Deliverable D1.1 User Requirements & Modelling

Work Package	WP1 - Clinical Requirements, Pre-Analytics, Sample Collection, Wet Lab Processing and Ethics
Task(s)	T1.1 Identification of Existing Body of Knowledge and Detailed User Requirements T1.2 Conceptual Modelling and Formal Requirement Specification
Author(s)	Massimo Mecella, Francesca De Luzi, Flavia Monti, Paolo Buono, Rosa Lanzilotti, Sabina Akram, Thomas Krause, Dominic Heutelbeck, Philippe Tamla, Haithem Afli, Bruno Andrade, Mike Leske, Andrea Leoni, Andrea Molinari
Creation date	01/02/2025
Last update	01/07/2025
Version	V1
Dissemination Level	PU

*All dates in dd/mm/yyyy format

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Table of Contents

Inhalt

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Executive summary

The GenDAI project is creating a medical diagnostics platform that leverages AI and metagenomic data for faster, more accurate clinical analysis. The platform automates data workflows to enable personalized medicine, initially focusing on tangible benefits for conditions like Irritable Bowel Syndrome (IBS).

This deliverable represents the foundational cornerstone of the GenDAI project, establishing the strategic and technical groundwork for all subsequent development. Following a rigorous methodology of state of the art research, expert interviews, and conceptual modeling, this document presents two primary sets of results:

(1) A comprehensive analysis of user needs, covering challenges in clinical workflows, data processing, security, AI, and visualization. (2) A complete conceptual and technical blueprint that translates these needs into a concrete system design using a full suite of models (Use Case, Information, Architecture, UI).

The primary value of this work is in creating a common, agreed-upon framework that aligns all partners, de-risks future development, and provides a traceable link from clinical needs to technical implementation. This ensures the final product will be fit-for-purpose.

The outputs of D1.1 are not static; they serve as the direct and essential inputs for the project's subsequent technical work packages. The models and specifications contained herein will guide the practical development and integration work across the project:

The architecture, component, and data processing models will directly inform the implementation of the GenDAI Diagnostics Workflow in WP2.

The data management and security requirements, including the specifications for Persistent Unique Identifiers (PUIs) and long-term archiving, will steer the development of the GenDAI Safe & Cloud Computing Platform in WP3.

The defined AI requirements for handling high-dimensional data, ensuring explainability, and enabling model training will guide the development of biomarker discovery models in WP4.

The UI and visualization techniques will be implemented to create the user-friendly GenDAI Interactive Reporting tools in WP5.

In summary, this deliverable provides the essential roadmap and technical specifications that will steer the hands-on development phases of the GenDAI project, ensuring the final platform successfully meets the validated needs of its intended clinical and research users.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

1. Introduction

1.1 Project Overview

The GenDAI project (Genomic applications for laboratory Diagnostics supported by Artificial Intelligence) aims to develop a novel medical diagnostics platform that leverages metagenomic data and Artificial Intelligence (AI) to improve the diagnosis and monitoring of inflammatory bowel disease (IBD). The platform will provide clinicians with a comprehensive, personalized view of a patient’s microbiome, enabling more targeted and effective treatment strategies.

1.2 Purpose of this Document

This document, D1.1 User Requirements & Modeling, serves as the foundational pillar for the GenDAI project. Its primary purpose is to present a comprehensive compilation of detailed user requirements and to translate these needs into a concrete conceptual system design. By meticulously defining the clinical and technical specifications, this document establishes a shared understanding that will guide all subsequent development, ensuring the final platform is aligned with validated user needs and is fit-for-purpose.

1.3 Document Organization

The document is structured to logically guide the reader from the problem space to the proposed solution.
Chapter 1 (Introduction): Outlines the document’s purpose, the scope of the GenDAI project, and the methodology employed for gathering and analyzing user requirements.
Chapter 2 (User Requirements and Challenges): Presents the core findings of our analysis. This section details the specific needs and challenges identified across key domains, including clinical diagnostics, data processing and automation, data security, AI, visualization, and regulatory considerations.
Chapter 3 (Conceptual Design and Modelling): Transitions from requirements to a concrete solution. This chapter presents a full suite of conceptual models, including Use Context, Use Case, Information, Component, Architecture, and UI Models, which form the technical blueprint for the GenDAI platform.
Chapter 4 (Conclusions and Next Steps): Summarizes the key conclusions and outlines how the results of this deliverable will be utilized as direct inputs for the project’s subsequent technical work packages.

1.4 Methodology

The overall methodology for the GenDAI project is guided by the Nunamaker research approach, which structures the project into four iterative phases: observation, theory building,

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

systems development, and experimentation[1]. This document, D1.1, represents the tangible output of the project's foundational Observation and Theory Building phases.

The identification and analysis of user requirements, which constitutes the core of the Observation phase, was based on a multi-faceted approach. We began with socio-economic, technical, and market modeling to profile the Supply and Demand sides of the Clinical Diagnostics ecosystem to understand the competitive landscape. This foundational modeling was then supported and informed by extensive desk research and direct fieldwork.

The process started with a background analysis, where we drafted preliminary statements on the issues under investigation based on the consortium's existing knowledge. This was systematically expanded through comprehensive desk research, which involved the collection and review of a wide variety of technological, medical, and market-related sources like scientific publications, market reports, and healthcare policy roadmaps. To enhance and validate these findings with real-world data, the team conducted extensive fieldwork. This primarily involved qualitative in-depth interviews with a range of experts, including managers, researchers, and practitioners. A cornerstone of this fieldwork was a comprehensive series of structured interviews with one of our clinical end-user, biovis Diagnostik MVZ GmbH, involving their key domain experts. These sessions, spanning a total of over four hours, were meticulously structured with participation from all partners of the GenDAI consortium, asking targeted questions to gather specific requirements for their respective areas. For these interviews, tailored discussion guidelines were developed to complement and illustrate our research. Furthermore, active participation in GenDAI-related events, such as conferences and workshops, provided invaluable opportunities to gather direct insights and engage with domain experts.

The rich data gathered during the observation phase served as the direct input for the subsequent Theory Building phase. This phase was executed using the principles of User-Centered System Design (UCSD), which places the end-user, their goals, and their tasks at the absolute center of the design process. This user-focused methodology guided the translation of abstract needs and challenges into the concrete, formal models presented in Chapter 3 of this document. To ensure clarity and standardization across the project, these models were specified using the Unified Modeling Language (UML), guaranteeing that every aspect of the conceptual design is directly traceable back to a validated user requirement.

Finally, ethical considerations were integrated throughout this entire process, from requirement gathering to system design, ensuring the GenDAI platform is developed responsibly from its very inception.

[1] Nunamer J-F, Chen M, Purdin T-D. System development in information system research. In: Twenty-Third Annual Hawaii International Conference on System Sciences; 1990. Kailua-Kona, HI, USA: IEEE; 1990. pp. 631–640. DOI: 10.1109/HICSS.1990.205401

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

2. User Requirements and Challenges

2.1 Overview

The European healthcare ecosystem faces increasing pressure from demographic shifts, such as an aging population, and the rising demand for more efficient and sustainable diagnostic solutions. To maintain high standards of care, there is a critical need for innovation that can deliver more accurate, faster, and cost-effective clinical analyses.

The GenDAI project addresses these challenges by developing a novel medical diagnostics platform that leverages Artificial Intelligence (AI) and metagenomic data. The platform aims to automate complex data processing workflows, improve diagnostic accuracy, and enable personalized medicine. The initial focus is on delivering tangible benefits for conditions like Irritable Bowel Syndrome (IBS), demonstrating the platform's potential to transform patient care.

To ensure the platform is fit-for-purpose, the requirements analysis focused on a clearly defined ecosystem of stakeholders. The primary users of the GenDAI platform are Clinical Analysis Laboratories, which can be independent service companies or integrated departments within hospitals and clinics. These laboratories serve a variety of purposes, including direct patient care, pharmaceutical drug development, and healthcare research.

To ground this analysis in a real-world context, the project's requirements are informed by the needs of a key end-user and project partner: Biovis Diagnostik MVZ GmbH. Biovis is a leading European medical diagnostic laboratory specializing in microbiome analysis. Processing approximately 1,000 stool samples per week, Biovis focuses on complex diagnostic cases and provides highly customized treatment recommendations, often for patients for whom conventional options have been exhausted. Their role as a key stakeholder ensures that the GenDAI platform is being developed to meet the practical challenges and advanced needs of an innovative, high-throughput clinical laboratory.
Other key stakeholder groups include:

- Clinicians (e.g., Gastroenterologists): As the "customers" of the laboratories, they order tests and use the diagnostic reports to make treatment decisions.
- Patients: The ultimate beneficiaries of the improved diagnostics and personalized treatment plans.
- Technology and Research Groups: Academic and commercial entities developing complementary or competing solutions.
- Regulatory Bodies: Institutions responsible for defining and enforcing the legal and quality standards that clinical laboratories must adhere to.

This framework of stakeholders and their respective needs establishes the context for the detailed user requirements presented in the subsequent sections of this chapter. The discussion systematically explores the challenges and needs across several key domains, encompassing the entire lifecycle of the diagnostic process. The analysis begins with the core

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

Clinical Diagnostics workflow, proceeds to examine the technical demands of **Data Processing and Automation** needed for handling high-throughput data, and then addresses critical requirements for **Data Management and Security**, including privacy and long-term archiving. Subsequent sections detail the specific needs for **Artificial Intelligence** in biomarker discovery and the user requirements for intuitive **Visualization and User Interaction** in reporting, concluding with the overarching **Regulatory and Ethical Considerations** that govern the entire system.

2.2 Clinical Diagnostics

2.2.1 State of the Art

Clinical diagnostics using metagenomics hinges on comprehensive and standardized methodologies for sample collection, processing, and reporting, essential for accurate and reproducible diagnostic results. Samples are collected from various body sites, including blood, saliva, and the gut, to assess the microbiome's complex interactions with human health conditions, as highlighted by numerous studies. Proper storage, processing, and quality control during these stages are paramount to maintain sample integrity and prevent contamination, with reporting guidelines emerging to support consistency across different studies and laboratories[2]. Sequencing technologies are the cornerstone of this diagnostic approach, particularly through 16S rRNA amplicon sequencing and shotgun sequencing. While 16S sequencing focuses on bacterial taxonomy by targeting specific genome regions, shotgun sequencing allows for more comprehensive profiling, analyzing genes across the entire microbial community[3]. Sequencing platforms like Illumina provide the necessary high-resolution data for characterizing complex microbial populations in clinical samples. These techniques have revolutionized the field by enabling large-scale genomic profiling at a fraction of the historical cost, thus making microbiome diagnostics more accessible.

Standardization in data formats, such as FASTQ for raw reads, enables consistent analysis and interoperability across platforms, databases, and tools. Key databases like GenBank, RefSeq, and KEGG offer extensive resources for annotating and validating diagnostic results, supporting both taxonomic identification and functional profiling by linking sample sequences to reference genomes and functional gene descriptions[4]. Ethical considerations are equally essential, as the integration of metagenomic data with patient information must adhere to GDPR standards for data protection and prioritize informed consent. Furthermore, compliance with the FAIR principles (Findable, Accessible, Interoperable, Reusable) is needed to ensure transparency and responsible data sharing across research networks, promoting trust and

[2] Mirzayi C, Renson A, Genomic Standards C, Massive A, Quality Control S, Zohra F, et al. Reporting guidelines for human microbiome research: the STORMS checklist. Nat Med. 2021;27:1885–92

[3] Ramazzotti, M.; Bacci, G. Chapter 5 - 16S rRNA-Based Taxonomy Profiling in the Metagenomics Era. In Metagenomics; Nagarajan, M., Ed.; Academic Press: London, UK, 2018; pp. 103–119. doi:10.1016/B978-0-08-102268-9.00005-7.

[4] Camacho, C., et al. "BLAST+: Architecture and Applications." BMC Bioinformatics 10 (2009), p. 421. ISSN: 1471-2105. doi:10.1186/1471-2105-10-421. ePrint: 20003500. URL: https://pubmed.ncbi.nlm.nih.gov/20003500/.

scientific reproducibility.

The interviews conducted within the project confirmed that while shotgun sequencing is of increasing interest, 16S rRNA amplicon sequencing (specifically the V3-V4 region for gut microbiome analysis, processed on platforms like the Illumina NextSeq 2000) remains the standard for routine, high-volume clinical diagnostics. This is largely due to practical constraints such as lower cost and faster turnaround times, which are critical in a clinical service environment. A key challenge highlighted was the significant impact of pre-analytical and analytical variability; the experts interviewed emphasized that the specific DNA extraction method and the chosen bioinformatics pipeline are often greater sources of inter-laboratory result deviation than sequencing errors themselves. This underscores the need for robust standardization beyond just the sequencing step. In their high-throughput setting, processing approximately 1,000 stool samples per week, strict quality control measures are essential. For instance, a minimum sequencing depth of 40,000 reads per sample is enforced to ensure data robustness. Furthermore, the analysis is contextualized with mandatory companion metadata such as patient age and sex, alongside other relevant clinical chemistry parameters measured from the stool (e.g., inflammation markers, pH), reinforcing the multi-modal nature of modern diagnostics. A significant barrier identified to the widespread adoption of metagenomics in diagnostics is the lack of a widely accepted consensus on the end-to-end process, from sample preparation to interpretation, a gap GenDAI aims to address.

2.2.2 Remaining Challenges / Requirements

The clinical adoption of metagenomic diagnostics faces several critical challenges. First, the sheer volume of metagenomic data generated by high-throughput sequencing platforms is difficult to process, interpret, and store, necessitating advanced computational resources and bioinformatics expertise. Efficient and interpretable machine learning (ML) models are still under development for clinical applications, and deep learning techniques, while promising, demand high levels of computational power and training data to reliably support diagnostics[5]. Second, the standardization of sample collection and processing methods across laboratories remains an unsolved issue. Variations in methodology can lead to discrepancies in diagnostic results, impacting the reproducibility of findings across different clinical settings. Although guidelines are emerging, comprehensive reporting standards that address each stage of the process are necessary to ensure that clinical metagenomic studies are both comparable and reproducible[6]. Another challenge arises from the ethical and privacy implications of handling vast amounts of sensitive data. Ensuring GDPR compliance, upholding informed consent, and maintaining adherence to the FAIR principles pose ongoing difficulties as the field scales up,

[5] Cacho, A.; Smirnova, E.; Huzurbazar, S.; Cui, X. A Comparison of Base-calling Algorithms for Illumina Sequencing Technology. Briefings Bioinform. 2016, 17, 786–795. doi:10.1093/bib/bbv088.
Teng, H.; Cao, M.D.; Hall, M.B.; Duarte, T.; Wang, S.; Coin, L.J.M. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. GigaScience 2018, 7. doi:10.1093/gigascience/giy037.

[6] Méndez-García, C.; Bargiela, R.; Martínez-Martínez, M.; Ferrer, M. Metagenomic Protocols and Strategies. In Metagenomics; Nagarajan, M., Ed.; Academic Press: London, UK, 2018; pp. 15–54. doi:10.1016/B978-0-08-102268-9.00002-1.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

especially in managing the balance between data accessibility and patient privacy[7]. Clinically, a significant hurdle is the current limited ability to link microbial genomic data to actionable health insights. For example, the functional analysis of genes within a microbiome is complex and requires deeper exploration to reliably associate specific microbial genes with health outcomes or therapeutic responses. In predictive modeling, the issue of high dimensionality (having more features than samples) presents a further obstacle, particularly when linking microbial genes to clinical phenotypes, where traditional ML techniques may struggle with sparse data[8]. Finally, there are interpretability issues with deep learning applications in microbiome classification, which can limit the ability of clinicians to understand or trust the results, thereby slowing down the integration of these technologies into routine diagnostics. Addressing these interconnected challenges through improved ML algorithms, better-defined methodological standards, and enhanced ethical frameworks will be crucial in unlocking the full clinical potential of metagenomics.

2.3 Data Processing and Automation

2.3.1 State of the Art

The process of metagenomic analysis has become a cornerstone in microbiome research and clinical diagnostics, enabling the identification and characterization of microbial communities within various environments. This analysis begins with raw sequence data obtained through amplicon or whole-genome sequencing. Depending on the nature of the study, researchers may utilize one of two primary clustering methods: Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs)[9]. OTUs, which group sequences based on a predetermined similarity threshold, are commonly used for studies focused on broader microbial taxonomy. ASVs, on the other hand, offer finer resolution by identifying unique sequence variants, often revealing subtle microbial population structures. Each method has advantages suited to particular study goals, with ASVs typically favored in high-precision analyses and OTUs useful for more generalized surveys.

The interviews conducted for this project confirmed that while both methods have been used, ASVs are now considered the state of the art for their higher precision. Consequently, leading diagnostic labs like our clinical partner Biovis have recently rebuilt their pipelines to adopt an ASV-based approach, often using algorithms like DADA2[10].

[7] Untergasser, A., et al. "RDML - RDML Compliant qPCR Instrument Software." Ed. by RDML Consortium, 2022. URL: https://rdml.org/instruments.html (visited on 07/03/2023).

[8] Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface 2018, 15. doi:10.1098/rsif.2017.0387.

[9] M. Chiarello, M. McCauley, S. Villéger, and C. R. Jackson, "Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold," PLoS ONE, vol. 17, no. 2, e0264443, 2022, doi: 10.1371/journal.pone.0264443.

[10] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson, and S. P. Holmes, "DADA2: High resolution sample inference from amplicon data," bioRxiv, 2015, doi: 10.1101/024034.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Following clustering, sequences undergo annotation and classification, relying on databases such as Greengenes, SILVA, and the NCBI RefSeq[11] to provide taxonomic identification. Our interviews revealed that in a clinical setting, labs use a combination of public databases but also heavily rely on their own curated internal databases, which are updated to reflect the latest taxonomic changes. For shotgun sequencing, functional annotation is also a crucial step, linking genetic information to potential functions and enabling insights into the microbial roles within the sampled environment. Databases tailored to gene functions, like KEGG and Pfam, are often incorporated here to align metagenomic sequences with known biological functions, which is particularly valuable in clinical diagnostics for understanding microbial impact on host health. Tools like QIIME2[12] and MGnify play significant roles in processing these annotations and classifications, offering extensive libraries of modules for metagenomic analysis. The data processed in these workflows are stored in standardized file formats like FASTA and FASTQ, which ensure compatibility across various platforms and ease of data sharing.

Beyond the initial analysis, data processing in metagenomics frequently includes steps like quality control, sequence alignment, and data visualization. Quality control, for instance, may involve trimming low-quality bases and filtering out sequences that fall below a certain threshold to maintain data integrity. The interviews highlighted that for a high-throughput lab, a key quality metric is sequencing depth, with a strict minimum threshold (e.g., 40,000 reads per sample) being enforced to ensure data robustness. Samples failing this check are reprocessed.

Alignment of sequences, or ‘binning,’ helps in assembling fragmented DNA by arranging overlapping sequences into a coherent contig, a set of sequences that aligns with a larger genome. These processing steps ensure high accuracy and relevance in downstream analyses, such as taxonomic profiling and functional analysis, which can reveal the abundance and relationships of various microbial taxa in the sample. Visualizations such as composition diagrams, phylogenetic trees, and correlation networks are often produced to summarize the microbial structure, improving data interpretability for researchers and clinicians alike. The final output can be a static report, typically in PDF format, or interactive visualizations that allow users to explore the microbial composition in more detail. Such interactive outputs are becoming increasingly popular for clinical applications, as they provide richer insights and are more adaptable to individual patient profiles. Our user study confirmed a desire for interactive visualizations that would allow clinicians to explore the data more dynamically.

Automation is a critical component in modern metagenomic workflows, especially as the volume of samples processed per batch has risen significantly in clinical and environmental studies. Platforms like QIIME2 and MG-RAST offer integrated metagenomic pipelines that guide users through predefined workflows, automating routine steps and allowing for some customization based on study-specific needs. However, these tools are generally limited in

[11] M. Balvočiūtė and D. H. Huson, "SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare?," BMC Genomics, vol. 18, Suppl 2, p. 114, 2017, doi: 10.1186/s12864-017-3501-4.

[12] E. Bolyen et al., "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," Nature biotechnology, vol. 37, no. 8, pp. 852–857, 2019, doi: 10.1038/s41587-019-0209-9.

their flexibility and scalability when handling very high sample volumes. Workflow Management Systems (WMS) such as Apache Airflow or Luigi have been adopted to address these limitations, enabling the orchestration of complex, multi-step workflows with high transparency and error handling. Airflow, in particular, is notable for its Directed Acyclic Graph (DAG) structure, which allows for the seamless integration of tasks within a single pipeline and provides robust logging and monitoring capabilities. This structure supports reproducibility by maintaining a documented sequence of events, essential for quality assurance in clinical diagnostics[13].

Our interviews showed that specialized labs often opt for custom-built solutions. A key requirement is a simple user interface that empowers trained laboratory staff, who are not bioinformatics specialists, to start, manage, and monitor analysis pipelines without deep command-line knowledge. This highlights a gap where existing tools may be too complex for routine clinical operation. Flexibility is another crucial need; labs require the ability to easily switch between different pipelines (e.g., for gut vs. vaginal microbiome analysis) and swap out specific algorithmic components (e.g., replacing a standard species classifier with a newer, more accurate one like the EMU algorithm) as the science evolves.

2.3.2. Remaining Challenges / Requirements

GenDAI's platform requirements for data processing and automation in metagenomic analysis reflect the need for a scalable, efficient, and reproducible pipeline capable of handling both high-throughput sequencing data and sensitive clinical applications. The GenDAI Diagnostics Workflow is designed to incorporate a reproducible microbiome annotation pipeline capable of identifying bacterial, viral, and fungal communities in patient samples. Leveraging machine learning models developed in the GenDAI Discovery workflow, this pipeline aims to automate the annotation process and enhance the accuracy of taxonomic and functional profiling. The workflow is also equipped to integrate with GenDAI's Interactive Reporting services, producing clinical reports that are not only visually rich but also dynamically generated to include microbiome compositions and reference profiles of healthy individuals. These interactive reports are intended to improve data interpretation for clinicians, providing detailed insights into microbial community structures and their potential impact on patient health.

The architecture of the GenDAI platform employs a cloud-based WMS to ensure reproducibility and transparency throughout the diagnostics pipeline. This system enables the automation of complex workflows by employing an event-based structure that logs each stage of data processing. This modular design directly addresses the requirement for flexibility identified in user interviews, allowing specific algorithmic components to be updated or swapped (e.g., integrating a new species classifier) without redesigning the entire workflow. This configuration supports traceability and accountability, essential for clinical diagnostics where reproducibility and data integrity are paramount. With high-throughput sequencing, the GenDAI pipeline can manage large volumes of data, enabling the analysis of multiple samples or batches per day. The event-based structure of the WMS not only enhances transparency

[13] T. Krause et al., "An Event-Driven Architecture for Genomics-Based Diagnostic Data Processing," Applied Biosciences, vol. 2, no. 2, pp. 292–307, 2023, doi: 10.3390/applbiosci2020020.

but also facilitates dynamic error handling, allowing specific steps to be re-run in cases of failure without disrupting the entire workflow. The cloud infrastructure, primarily hosted on Google Cloud, provides the scalability needed to handle varying levels of data load, with built-in support for containerized applications using Docker and Kubernetes, ensuring compatibility across different deployment environments.

A significant focus within the GenDAI workflow is the integration of long-term archival solutions developed through GenDAI Safe. This aspect of the workflow allows the preservation of all input data, intermediate results, and final outputs, ensuring that each analysis can be reproduced using the exact version of the software and ML models originally applied. The incorporation of persistent storage linked with the GenDAI Diagnostics Workflow and associated software components supports both data reproducibility and regulatory compliance, as it logs metadata such as model weights and software versions, meeting the stringent requirements of clinical and diagnostic data management. The workflow also integrates with the GenDAI Discovery Model for biomarker detection, allowing the seamless annotation of patient samples with newly identified biomarkers and thus enhancing the diagnostic precision.

The interactive reporting capability of the GenDAI Diagnostics Workflow provides a sophisticated, user-friendly platform for clinical reporting, wherein microbiome composition and patient health profiles are automatically generated based on processed data. The integration of visualization tools developed in GenDAI's visual services enhances the interpretability of these reports. With end results available in both static (e.g., PDF) and interactive formats, the pipeline is adaptable to the needs of different clinical environments. Finally, each stage of the pipeline includes rigorous archiving through the GenDAI Safe archiving services, which ensures that all inputs, model parameters, and processing outcomes are securely stored and linked to their respective pipeline versions. This approach reinforces the platform’s commitment to reproducibility, allowing future diagnostic reviews to accurately trace each data point back to its source, thus maintaining a high standard of transparency and reliability in diagnostic data processing.

2.4 Data Management and Security

2.4.1 State of the Art

The rapid growth in Genomics-Based Diagnostic Data (GBDD), driven by advancements in genomic sequencing technologies, presents complex challenges in data management and security. These technologies have enabled the quick, cost-effective generation of large-scale data, facilitating breakthroughs in diagnostics and personalized medicine. However, the vast volume, rapid velocity, and high variety of GBDD necessitate an infrastructure capable of handling this scale and complexity[14]. Effective GBDD management must not only offer extensive storage and processing capabilities but also accommodate various sequencing

[14] Krause, E. Jolkver, M. Kramer, P. Mc Kevitt, and M. Hemmje, "A Scalable Architecture for Smart Genomic Data Analysis in Medical Laboratories," in Angewandte Data Science: Projekte | Methoden | Prozesse, L. Blum, Ed.: Springer, 2023.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

approaches, such as amplicon and shotgun sequencing, each of which requires tailored storage and analysis techniques. Privacy concerns add another layer of complexity, as GBDD contains highly sensitive information. Adherence to regulations like GDPR demands strict data privacy protocols to safeguard patient confidentiality and ensure that only authorized users have access. To maintain scientific rigor and ensure reproducibility, PUIs are used to track and verify data workflows, reinforcing accountability and traceability. Technologies such as Apache Spark and Hadoop have proven essential for scalable data processing in genomics, while microservices and Event-Driven Architectures (EDA) provide modular and secure frameworks for handling GBDD, enhancing scalability, efficiency, and data security.

Given the sensitive nature of genomics data, access control plays a critical role with regard to compliance with privacy standards, such as HIPAA and GDPR. Currently, Attribute-based access control (ABAC)[15] provides the most flexible access control mechanism to implement various fine-grained access control models. ABAC allows for the specification of contextualised policies that apply to the discussed genomics use-cases. The most prominent implementations of this mechanism are XACML[16], NGAC[17], and recently OPC[18]. These implementations all follow a request-response communication pattern. Given the publish-subscribe nature of Event-Driven Architectures, this is a mismatch between communication paradigms. While an integration is possible, it can lead to bottlenecks by introducing polling of authorization endpoints and associated latencies. Therefore, in such systems, Attribute Stream-Based Access Control (ASBAC)[19] and an implementation based on the Streaming Attribute Policy Language (SAPL)[20] may be beneficial and provide a more natural pathway of integration.

Persistent Unique Identifiers (PUIs) in Medical Data are essential to enhance data integration, integrity and security, facilitating interoperability among disparate systems, and safeguarding patient privacy. Efficiently managing this data is critical for improving patient outcomes, advancing medical research, and optimizing healthcare delivery systems. PUIs are essential for the consistent and accurate identification of patients, healthcare providers, devices, and other entities within medical databases.

[15] Hu, V. C., Ferraiolo, D., Kuhn, R., Friedman, A. R., Lang, A. J., Cogdell, M. M., ... & Scarfone, K. (2013). Guide to attribute based access control (abac) definition and considerations (draft). NIST special publication, 800(162), 1-54.

[16] Standard, O. A. S. I. S. (2013). extensible access control markup language (xacml) version 3.0. A:(22 January 2013). URI: http://docs. oasis-open. org/xacml/3.0/xacml-3.0-core-spec-os-en. html.

[17] Information technology - Next Generation Access Control - Functional Architecture (NGAC-FA), INCITS 499--2013, American National Standard for Information Technology, American National Standards Institute, March 2013.

[18] Open Policy Agent. (n.d.). Open Policy Agent. Retrieved June 26, 2025, from https://www.openpolicyagent.org/

[19] Heutelbeck, D. Attribute Stream-Based Access Control (ASBAC) - Functional Architecture and Patterns. In Proceedings of the 2019 International Conference of Security and Management (SAM'19) (2019).

[20] Heutelbeck, D. The Structure and Agency Policy Language (SAPL) for attribute stream-based access control (ASBAC). In Proceedings of the ETAA 2019 : 2nd International Workshop on Emerging Technologies for Authorization and Authentication (2019).

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

PUIs are stable references assigned to entities (e.g., patients, specimens, devices) that remain consistent over time and across different systems, or system updates, ensuring that the entity can be reliably tracked and referenced throughout its lifecycle. For this, it is essential to create an Entity Lifecycle management system or service, able to guarantee the most possible primitive operations on entities: creation, deletion, attribute update, retrieval, but, mostly, more complex primitives like merging, intersecting, upgrading, etc.

Fundamental in entity management is also the preservation of the temporal dimension of PUIs: entities change over time, in nature and attribute, and their temporal evolution is an essential component of any AI-based solution to preserve knowledge graph consistency and, at the same time, guarantee coherent and meaningful inference processes. PUIs Persistent identifiers are foundational to the effectiveness of knowledge graphs and AI operations. They ensure that entities are uniquely and consistently identified, which is essential for data integration, quality, and interoperability. By facilitating unambiguous linking and retrieval of information, PIDs enhance the capabilities of AI systems, leading to more accurate models and insightful analyses.

In terms of user requirements for implementing PUI infrastructure in Gendai, its implementation requires a comprehensive set of user requirements to ensure effectiveness, security, and compliance. These requirements address the needs of various stakeholders, including patients, healthcare providers, researchers, and regulatory bodies. Here is a list of actions and requirements:

- Uniqueness: Each PUI identifier must be unique to prevent duplication and misidentification of entities such as patients, specimens, or devices.
- Persistence: PUI identifiers should remain constant over time and across different systems to enable long-term tracking and referencing.
- Adherence to Standards: Use standardized formats (e.g., UUIDs, HL7 standards) to ensure compatibility between different healthcare information systems.
- System Integration: The PUI identifier system must seamlessly integrate with existing and future healthcare applications and databases.
- Cross-System portability: Facilitate data exchange between disparate systems, both within and between organizations.
- Data Protection Measures: Implement encryption, access controls, and authentication protocols to safeguard PUIs identifiers from unauthorized access and breaches.
- Anonymization and De-identification: Ensure that patient privacy is maintained by properly de-identifying data when used for research or shared externally.
- Compliance with Privacy Laws: Adhere to regulations like HIPAA and GDPR regarding the handling and protection of personal health information.
- Legal Compliance: Systems must meet all legal requirements related to data storage, usage, and sharing.
- Audit Trails: Maintain detailed logs of data access and modifications to facilitate audits, AI explainability according to GDPR obligations, and demonstrate compliance.
- Reporting Capabilities: Provide tools for generating reports required by regulatory agencies.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

- Intuitive Design: The system should have an easy-to-use interface that integrates smoothly with healthcare providers' workflows. PUIs (like, for example, DOI for academic publications) are known for their complexity of memorization by humans, and at the same time, the need to have human-readable, mnemonic PUIsidentifiers. A good compromise among these different, contrasting needs is essential to ensure their usability
- Minimal Disruption: Implementation should not significantly disrupt existing processes or require extensive retraining.
- Support and Training: Offer comprehensive training programs and ongoing support for users about their interpretation, usage, creation and management.
- High Availability: Ensure that the PUI management system is reliable with minimal downtime to support critical healthcare operations.
- Efficiency: Optimize for quick data retrieval and processing to not hinder clinical decision-making.
- Scalability: The PUI management system should handle increasing amounts of data and users without performance degradation.
- Data Ownership Policies: Define clear policies regarding who owns the data associated with PUIs.
- Use Limitation: Ensure that data is used only for agreed-upon purposes, respecting patient autonomy and preferences
- Re-identification Prevention: Implement safeguards to prevent the re-identification of de-identified data.
- Budget Alignment: The implementation and maintenance costs should be justifiable and aligned with organizational budgets.
- Technology Agnostic Design: Design systems that can adapt to new technologies and standards.
- Regulatory Changes Adaptation: Ability to quickly comply with new laws or regulations as they arise.

Different types of PUIs for metagenomic analysis

In metagenomic analysis, several types of PUIs (persistent unique identifiers) are utilized to conduct and report research. Accession numbers are unique identifiers assigned to various types of data submitted to public databases like the European Nucleotide Archive (ENA) or the National Center for Biotechnology Information (NCBI). These include identifiers for raw sequencing data (SRA), annotated sequences (GenBank), biological samples (BioSample), and overarching projects (BioProject). Each type of accession number serves a distinct purpose, reflecting different levels of biological data organization. Another important common database for accession numbers is Gene Expression Omnibus (GEO), which manages genomics array- and sequence-based data submissions. International Classification of Diseases (ICD) provides codes to uniquely identify and classify health conditions (e.g. inflammatory bowel disease). Other standard PUIs are present in the Human Microbiome Project (HMP) to identify stool samples from individuals and in the Chemical Abstracts Service (CAS), referring to chemicals used during DNA extraction or analysis.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Taxonomies and Ontologies in Metagenomics

In order to classify and identify entities in the field of metagenomics, numerous taxonomies and ontologies are globally recognized and used. Some of the most relevant are: NCBI taxonomy (microorganisms classification); Greengenes, SILVA and RDP taxonomies (microbial 16S rRNA gene sequences classification); Gene Ontology (GO, annotate gene products in terms of their biological processes, molecular functions, and cellular components); Human Disease Ontology (DO, standardized definitions of human diseases) and Sequence Ontology (SO, biological sequences, such as exons, introns, promoters, and operons).

Other standards - generic for scientific products

In the field of metagenomic research, various standards play a crucial role in ensuring data integrity, reproducibility, and accessibility. Key identifiers such as DOI (Digital Object Identifier) are widely used to uniquely identify research articles, datasets, and experimental protocols. For instance, datasets shared in public repositories like the NCBI Sequence Read Archive (SRA) or European Nucleotide Archive (ENA) are often assigned DOIs. Similarly, ORCID (Open Researcher and Contributor ID) ensures clear attribution for researchers, avoiding confusion due to name similarities. DOIs for protocols, such as those shared via Protocols.io, further facilitate the dissemination and reuse of experimental methodologies.

Other standards - metagenomic reporting

Standards specific to metagenomic analysis include MIxS (Minimum Information about any (x) Sequence) and MIMARKS (Minimum Information about a MARKer gene Sequence), both developed by the Genomic Standards Consortium (GSC). MIxS defines a core set of metadata necessary for metagenomic samples, including sample type, sequencing methods, and environmental context, while MIMARKS targets marker gene sequences, like 16S rRNA, ensuring the quality and reproducibility of amplicon-based studies. The FAIR Principles (Findable, Accessible, Interoperable, Reusable) guide the management of metagenomic data to enhance its discoverability and usability. Similarly, the Global Alliance for Genomics and Health (GA4GH) develops standards like Phenopackets and RefGet for securely and interoperably sharing genomic and clinical data, which can be particularly relevant in health-related metagenomic studies.

Metadata standards such as BioSample and BioProject, used by repositories like NCBI and ENA, ensure consistent formats for describing projects and samples, providing critical context for metagenomic studies. Platforms like SEEK and the ISA Framework (Investigation, Study, Assay) offer structured metadata solutions for experimental workflows, with formats like ISA-Tab and ISA-JSON representing comprehensive descriptions of studies and assays. Tools from the Open Microscopy Environment (OME) and the Bio-Formats software support imaging metadata integration, which is particularly relevant in multi-omics approaches combining microscopy and metagenomics. Bioschemas, a community-driven initiative, enhances the discoverability of metagenomic datasets by adding structured metadata for web interoperability.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

For metagenomic studies involving clinical data, standards from the Clinical Data Interchange Standards Consortium (CDISC), such as CDASH and SDTM, ensure consistent clinical metadata, including patient health information. The Genomic Standards Consortium (GSC) also provides other frameworks, like Minimum Information about a Metagenome Sequence (MIMS), defining essential metadata for metagenomic samples.

Data preprocessing and quality control are supported by platforms like QIIME2 and Mothur, which establish reproducible workflows for sequence trimming, denoising, and taxonomic assignment. The Minimum Information about Bioinformatics Investigation (MIABi) standard further ensures transparency in bioinformatics workflows by outlining the required metadata, such as software used and parameter settings. Finally, the CAMI (Critical Assessment of Metagenome Interpretation) initiative benchmarks metagenomic analysis methods, providing guidelines for evaluating taxonomic binning, assembly, and functional annotation to ensure robustness in metagenomic research. Together, these standards create a robust framework for the reliable and reproducible analysis of metagenomic data.

Coverage of PUIs across biomedical entities

Persistent Unique Identifiers (PUIs) are not uniformly applied to all entities. While sequences, taxonomic units, and external datasets are commonly assigned persistent identifiers, often because of the need for publication, data sharing, or regulatory compliance, many internal entities such as users, projects, protocols, and local documents are either not assigned PUIs at all or are managed using ad hoc internal identifiers[21]. Thus, the allocation of identifiers tends to focus on objects essential for reproducibility or presentation of external data.

2.4.2 Remaining Challenges / Requirements

In GenDAI’s approach to data management and security, scalability, compliance, and secure data accessibility are key. The Diagnostics Pipeline emphasizes progressive enhancements in automated processing, interactive reporting, and secure data handling. The initial pipeline delivery includes foundational measures for secure sample processing with comprehensive documentation, while subsequent updates integrate AI-driven models and interactive reporting features to increase pipeline functionality. Over time, the pipeline incorporates long-term archiving and advanced security mechanisms, supporting data traceability and enhanced compliance. The GenDAI Safe and Cloud Computing Platform utilizes cloud infrastructure to provide GDPR-compliant, region-specific services while employing Docker and Kubernetes for scalable, adaptable data management.

Long-term archiving strategy

Policies and workflows for secure data access are developed alongside a Long-Term Archiving System based on ISO 14721 standards, which ensures that GBDD resources are stored and managed in ways that support reproducibility and transparency. Security measures also include fine-grained access control through dynamic models that safeguard sensitive

[21] Sansone, S. A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., ... & Hide, W. (2012). Toward interoperable bioscience data. Nature genetics, 44(2), 121-126.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

data, while ENS Integration supplies PIDs to enhance traceability across data workflows. This combination of scalable infrastructure, privacy-compliant storage, and robust traceability establishes a comprehensive foundation for managing and securing GBDD in line with both present and future diagnostic needs.

In addition, ensuring that the PUI infrastructure itself is inherently scalable to accommodate growing data volumes and evolving user bases, and adaptable to future technological shifts or emerging regulatory changes, presents a continuous design and maintenance challenge for long-term effectiveness. For instance, microservices architectures deployed via containerization (e.g., Docker, Kubernetes) on cloud platforms offer unparalleled flexibility and adaptability for big data processing. This modularity facilitates seamless integration of new analytical tools, rapid deployment of updates, and resilience against evolving technological and regulatory landscapes[22].

Coverage gaps and metadata standards

While sequence-level data often benefits from well-established identifiers such as GenBank accession numbers, many internal entities remain untracked by persistent identifiers. This limited coverage poses challenges to reproducibility, provenance, and interoperability across research infrastructures[23].

Practical implementation hurdles

In addition to conceptual challenges, the real-world implementation of PUI systems for medical data often encounters difficulties related to legacy systems, data migration, and stakeholder buy-in. In this direction, documenting and addressing these obstacles and learning from past experiences remains crucial.

Temporal evolution and entity tracking

Managing the temporal evolution of entities associated with PUIs is a complex problem in knowledge graphs and AI systems. In environments such as medical diagnostics and genomics, entities such as patient health status, microbial taxonomies, scientific knowledge, and bioinformatics methods are not static. Ensuring that the PUI infrastructure and associated knowledge graphs accurately reflect these evolving attributes and relationships is essential. Among the implications for Persistent Unique Identifiers (PUIs) is the need to ensure traceability of the evolving entities they refer to. For instance, if a PUI identifies a microbial sequence, it is essential to maintain the association between that PUI and the current or historical taxonomic classification, especially when the taxonomy of the organism changes over time. Another important issue concerns the identification of patients who are seen at different points in time: it is crucial to ensure the linkage between longitudinal microbiome data and the evolution of their clinical status, so that the information remains consistent and usable

[22] Akerele, J. I., Uzoka, A., Ojukwu, P. U., & Olamijuwon, O. J. (2024). Improving healthcare application scalability through microservices architecture in the cloud. *International Journal of Scientific Research Updates*, *8*(02), 100-109.

[23] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

over time.

Compliance with GDPR, HIPAA, and privacy standards

Clarifying how identification systems maintain compliance while preserving utility is a key issue, particularly in clinical settings. Therefore, the need to develop robust measures to ensure PUI compliance with privacy regulations such as GDPR and HIPAA while maintaining usability for healthcare providers and researchers remains a significant challenge. The challenge is therefore to specifically analyse the impact of GDPR on health data use and management, including consent requirements, data minimisation and technical/organisational measures, thus providing a crucial framework for understanding these complexities[24]. This includes robust anonymization/pseudonymization strategies and clear access control policies that go beyond simple identifier assignment.

Harnessing PUIs for enhanced AI capabilities

The potential for Persistent Unique Identifiers (PUIs) to enhance Artificial Intelligence (AI) systems within the organization remains a largely untapped challenge. While there is an explicit interest in leveraging AI for pattern detection, correlation with patient metadata, and suggesting biomarkers, current diagnostic workflows do not routinely employ AI or machine learning for result interpretation. PUIs facilitate crucial data integration, enhance data quality, and promote interoperability, but also link and retrieve information that are vital for training robust AI models and deriving accurate, insightful analyses, as widely discussed in the context of Knowledge Graphs in medical informatics and the necessity of FAIR data principles for AI readiness[25] [26].

The implementation of ICD codes (International Classification of Diseases) and DOIs (Digital Object Identifiers for data or research outputs) in automated result interpretation points towards a clear strategic direction. Integrating PUIs with standardized clinical terminologies like ICD codes would enable AI systems to link genomic and microbiome data directly to clinical diagnoses and outcomes, significantly enriching predictive models[27]. Similarly, the use of DOIs could facilitate the citation and traceability of specific data versions or analysis results within an AI pipeline, boosting reproducibility and auditability. The key drivers for such an implementation would be the ability to gain deeper, AI-driven insights, automate complex decision-making processes, and standardize reporting in a way that is interoperable with

[24] Budin-Ljøsne, I., Teare, H. J., Kaye, J., Beck, S., Bentzen, H. B., Caenazzo, L., ... & Mascalzoni, D. (2017). Dynamic consent: a potential solution to some of the challenges of modern biomedical research. BMC medical ethics, 18, 1-10.

[25] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9.

[26] Lu, Y., Goi, S. Y., Zhao, X., & Wang, J. (2025). Biomedical Knowledge Graph: A Survey of Domains, Tasks, and Real-World Applications. arXiv preprint arXiv:2501.11632.

[27] Chaturvedi, N., Yadav, M. K., & Sharma, M. (2024). Applications of artificial intelligence and machine learning in microbial diagnostics and identification. Methods in Microbiology, Academic Press, Volume 55, 2024, 213-230.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

broader healthcare systems.

2.5 Artificial Intelligence

The GenDAI project leverages advanced Artificial Intelligence (AI) to enhance genomic analysis, streamlining processes from pre-analytics through to ethical compliance. AI plays a pivotal role by enabling automated and high-throughput data processing in clinical diagnostics. By integrating machine learning (ML) and Natural Language Processing (NLP) techniques, AI facilitates the extraction of valuable insights from genomic and literature data, essential for refining sample collection protocols and pre-analytics. Specifically, NLP enables the mining of scientific literature to identify emerging trends and critical markers that can influence sample processing and experimental design. This approach supports a more nuanced and contextually rich understanding of genomic data, ultimately improving diagnostic accuracy and efficiency.

2.5.1 State of the Art

Since the introduction of the Transformer architecture[28], the NLP domain has seen unprecedented progress. By processing whole text sequences in parallel through the transformer blocks, a computational (self-)attention mechanism allows the model to learn complex word interactions and, as a consequence, to learn a complete "language model". Transformer-based (large) language models like BERT[29], GPT-2/3/4/o1[30][31][32], Gemini[33], Claude, or the Llama family[34][35] have benefited immensely from parallel GPU processing and the global accumulation of digital text data across the internet. Today, Large Language Models have been trained on trillions of words and are typically made of hundreds of billions of

[28] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[29] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.

[30] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[31] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[32] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.
Gemini Team, Google. (2023). Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805.

[33] Gemini Team, Google. (2023). Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805.

[34] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, P., ... & Lample, G. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[35] Meta. (2024). Llama 3: The Next Generation. arXiv preprint arXiv:2407.21783.

parameters. Next to text understanding, these models are also excellent in generating text that is indistinguishable from human-generated text.

To increase the factual correctness of text generated by Large Language Models, "Retrieval Augmented Generation" (RAG) has established itself as a successful concept to extend LLM capabilities for knowledge-intensive tasks. Here, large corpora of text are sliced into subunits, for example, sentences or paragraphs, and indexed using semantic embeddings. Semantic embeddings are vectorized representations of text chunks, where similar text input results in similar vector representations. Basic mathematical vector operations can subsequently calculate the similarity between two or more text fragments. Using RAG methodologies, AI-powered applications do not immediately send a user query or instruction straight to the LLM, but first convert the user query into a vector using the same semantic embeddings applied on the vectorized text corpus to retrieve the n most relevant text passages to answer a user's query. These text passages are then provided to an LLM together with the query. Key benefits of a RAG approach include that neither the LLM is required to answer a user query based on its acquired "world knowledge", nor does the user have to train or fine-tune an LLM with the domain knowledge.

Similar to natural languages, the genome is also a language in itself, with nucleotides as letters, genes as words, and functional pathways as phrases, featuring its own syntactic rules defining how genes interact with each other. As the architecture of transformer models improved to process sequences of gradually increasing length, a new field of Genome Language Models has emerged in which Transformers and their potential successor technologies are trained on massive genome datasets. The ability to process extremely long sequences compared to Natural Language Models is mandatory, as bacterial genomes as well as genes of mammals are easily multiple million base pairs long, and gene regulation often depends on very distant regions of the genome. The most prominent Genome Language Models are DNABERT-2[36], Nucleotide Transformer[37], and Evo[38]. Typical use cases for Genome Language Models include sequence classification, promoter detection, and transcription factor detection.

2.5.2 Remaining Challenges / Requirements

The integration of AI into biomarker identification and clinical diagnostics, particularly within the metagenomics domain, presents several significant challenges. Insights from user interviews highlight critical requirements for successful adoption and clinical utility.

[36] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri and Han Liu. (2024). DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genomes. arxivxiv, 2306.15006v2.

[37] Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J. et al. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. Nat Methods 22, 287–297 (2025). https://doi.org/10.1038/s41592-024-02523-z.

[38] Nguyen, E., Poli, M., Durrant, M. G., Kang, B., Katrekar, D., Li, D., … & Hie, B. (2024). Sequence modeling and design from molecular to genome scale with evo. Science, 386(6723). https://doi.org/10.1126/science.ado9336

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

A. Handling High-Dimensional and Complex Data: Metagenomic datasets are inherently high-dimensional, encompassing thousands to millions of microbial taxa, genes, and metabolic pathways, often alongside complex patient metadata and clinical outcomes.

- **Challenge:** The sheer volume and sparsity of this data make traditional statistical methods insufficient. AI models must effectively extract meaningful signals from noise, reduce dimensionality without losing critical information, and handle compositional data peculiarities.
- **Requirement:** Users require AI models capable of processing and integrating diverse data types (e.g., continuous clinical data, co-variables like ethnicity and sex) to refine feature selection, as envisioned in the improved BiGAMi algorithm (T4.1). This pre-processing is crucial for removing phenotypic variation not directly related to the microbiome.

B. Model Robustness and Generalisability: The clinical utility of AI models hinges on their ability to perform reliably across varied patient cohorts, different sequencing platforms, and diverse environmental conditions.

- **Challenge:** Models trained on specific datasets may not generalise well to unseen data or different populations, potentially leading to inaccurate diagnoses or irrelevant biomarker findings. Ensuring model stability and performance in real-world clinical settings is paramount.
- **Requirement:** Models must demonstrate high precision and recall, especially in sensitive diagnostic contexts. Rigorous validation, such as 100 rounds of 5-fold cross-validation (T4.2), is essential to build confidence in the model's reliability across diverse scenarios.

C. Explainability (XAI) and Interpretability: For AI-derived insights to be actionable in clinical practice, clinicians and researchers need to understand *why* a model made a particular prediction or identified a specific biomarker. This was a strong theme from user interviews.

- **Challenge:** Many advanced AI models, particularly deep learning architectures, operate as "black boxes," making their internal decision-making processes opaque. This lack of transparency can hinder trust, adoption, and regulatory approval in medical applications.
- **Requirement:** Users expressed a strong preference for AI systems that not only classify microorganisms but also **directly suggest biomarkers alongside genus/species classifications**, providing probabilistic assessments and confidence scores. They desire outputs presented as ranked lists or detailed reports to make AI-generated suggestions actionable. This necessitates models that offer clear insights into feature importance and the rationale behind biomarker identification.

D. Bias Mitigation and Fairness: AI models can inadvertently learn and perpetuate biases present in their training data, leading to unfair or inequitable outcomes, particularly across different demographic groups.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

- **Challenge:** Genomic and clinical datasets may not be representative of global populations, introducing biases related to ancestry, ethnicity, and socioeconomic factors. Ensuring that AI models provide unbiased and equitable outcomes for all patient demographics is crucial.
- **Requirement:** The project must prioritise the use of diverse datasets and implement strategies for bias detection and mitigation to ensure model fairness and ethical compliance (e.g., GDPR).

E. Handling Novel and Unclassified Species for Biomarker Discovery: Current classification methods often struggle with microorganisms that are uncharacterised or not yet included in reference databases, limiting the discovery of novel biomarkers. User interviews highlighted a significant interest in addressing this gap.

- **Challenge:** Many potentially critical microbial components might remain "unclassified" or "novel," yet hold significant diagnostic or therapeutic potential. Existing biomarker discovery pipelines may overlook these.
- **Requirement:** Users would greatly benefit from an AI system that can **flag potentially important novel species** and suggest experimental pathways for validating biomarkers derived from these species. This capability is seen as critical for addressing conditions with unmet clinical needs, such as rare diseases or treatment-resistant infections, and could reduce delays in personalised treatment plans.

F. Integration with Clinical Workflow and Domain Expertise: Successful adoption requires AI tools to seamlessly integrate into existing clinical and research processes, allowing domain experts to validate and refine findings.

- **Challenge:** Discrepancies between AI outputs and established clinical practice, or a lack of user-friendly interfaces for interaction, can impede adoption.
- **Requirement:** Users desire the ability to incorporate AI outputs into existing pipelines for personalised diagnostics (e.g., for IBD). A simple user interface (T4.4) for fine-tuning or retraining developed models for specific applications or updated sample collections is essential. This framework, potentially utilising split-fed learning (T4.2), allows users to tailor the model for their personal needs, increasing the potential for novel discovery and ensuring that AI-derived insights are clinically relevant and validated by experts. Collaborative efforts with industry experts (T4.3) will ensure practical insights into application, deployment, and commercialisation.

Computational and Data Requirements

The use of 16S rRNA sequences in taxonomic classification requires significant computational resources, including high-performance computing for managing and analysing large-scale genomic datasets. The computational infrastructure must be robust enough to support machine learning algorithms and NLP-based models for genomic data parsing. Additionally, to facilitate efficient biomarker identification, the project requires access to comprehensive and high-quality 16S rRNA datasets. These datasets should cover a broad spectrum of microbial

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

taxa to ensure the inclusivity and diversity needed for reliable biomarker discovery across different clinical conditions and demographics.

Machine Learning and Data Ethics in Genomic Data Processing

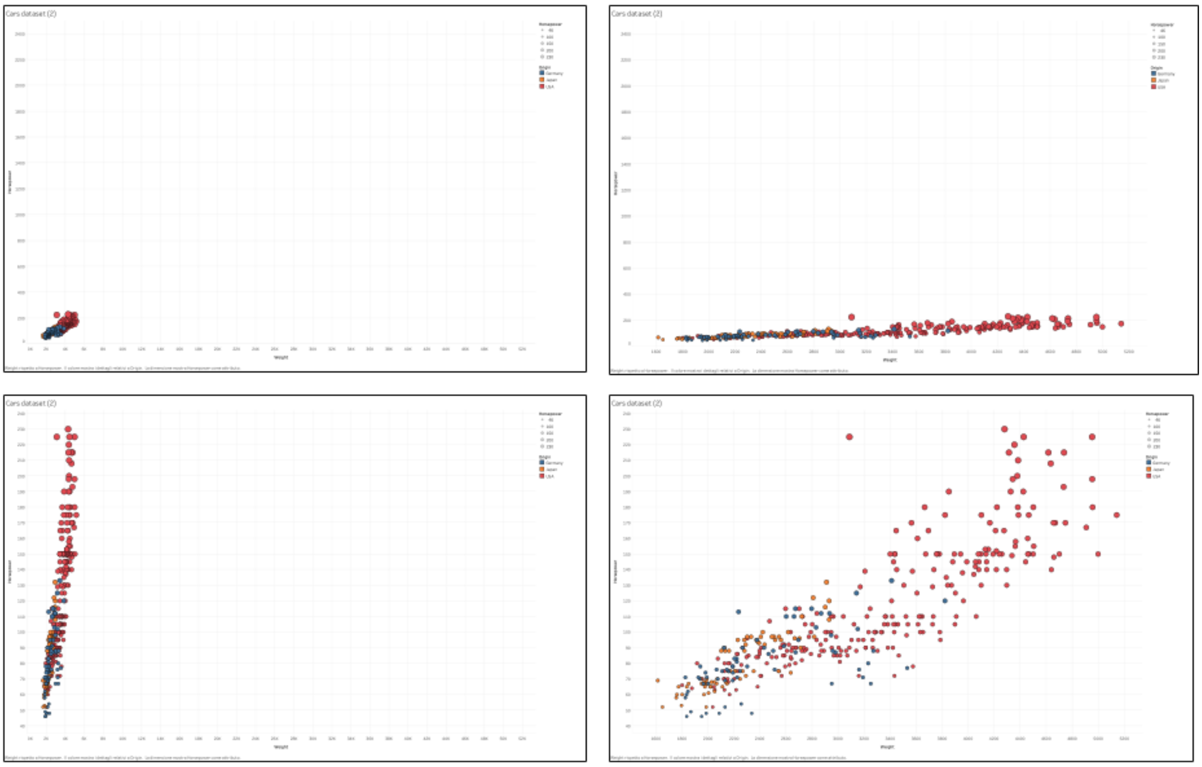
The application of AI in GenDAI also encompasses robust ethical considerations, ensuring alignment with GDPR and other regulatory standards for data privacy and compliance. The AI-driven approach involves anonymisation and secure management of genomic data, safeguarding sensitive patient information throughout the analysis. Furthermore, model fairness and transparency are prioritised by incorporating diverse datasets and validating models across various population groups. This ensures that AI models provide unbiased and equitable outcomes for all patient demographics, addressing potential ethical concerns in AI-driven genomic research.

2.6 Visualization and User Interaction

Visualization is the act of communicating information through graphical representations. The visual channel has been used as a communication means since before the formalization of written language. One single picture is language-independent and can contain a lot of information that can be processed in parallel by the human perception system. In contrast, text is slower because of two main reasons: 1) the human processes text sequentially; 2) in order to create a mental model of the data, the human must transform text into images, which in the visualization are already available. Visualizing data is a process that transforms numbers and text into a visual form in order to improve the speed of accessing the information and creating opportunities to see patterns that are not trivial to identify using computational methods.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

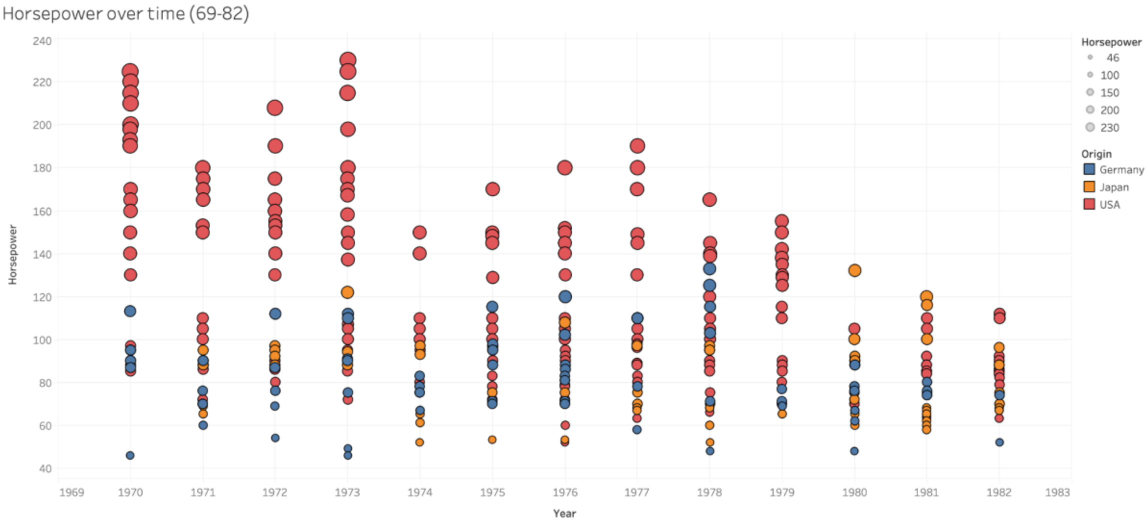
Funded by the European Union



Visualization is important for many reasons, but making good visualizations is more important; otherwise, the power of visualizations is not well exploited. The figure above reports a sample dataset shown in four scatterplots with different scales on the x and y axes. The scatterplot is useful and well-known to reveal correlations. The scatterplot at the top left is not clear, and all items are overlapped and concentrated in a very tiny area. The top-right and bottom-left scatterplots show the same dataset where only the x or y axis is normalized, respectively. In the scatterplot at the bottom right both axes are normalized and clearly show the data distribution. The dataset is about cars, and the axes compare the weight with the horsepower. The bottom right scatterplot clearly shows that heavier cars have more horsepower, but this relationship is less strong as the weight increases. Everyone can immediately see the outlier at the top of the chart, which is referred to as a Buick Estate wagon (1970), characterized by a very high power (225HP) with an average weight. Other cars with such power are two Pontiacs and a Buick Electra, which have a weight that is almost double compared to the Estate. The scatterplot also reveals another dimension, which is the nationality of the car brands, shown using the colors. Another visible information is that bigger and heavier cars are from the USA (red dots on the top-right quadrant of the bottom right scatterplot). By changing the compared variable, is it possible to grasp further information.

The following scatterplot shows the same dataset as the previous figure, where horsepower and construction year are compared. It is clear that the trend of horsepower has been towards power reduction. USA keeps the most powerful cars, but it is also visible that in 1971 there was a first attempt to reduce power, and again in 1974, and since 1977 there has been a progressive reduction of power. Interestingly, no aggregation or statistical computation has been performed to show the data; the data have been just plotted in the scatterplot, with a proper encoding, and the user can derive information, provided that the encoding is clear.

In this second example, the analyst has changed the visualization, replacing the variable of interest. The aspect of interaction is very important in visualization, because the observer may notice something that is interesting for the analysis and the possibility of modifying the view can allow the analyst to see things from a different perspective, in order to get a better mental model of the data.



Such information, if provided with data only, would be more difficult to grasp, because the text is read sequentially, takes typically more time to parse, and more space than visual items. Als,o the trend is difficult to get in the textual form. Consider that the example is composed of 9 attributes and 392 rows, but even if the visualization remains almost the same, if we add data, in textual for,m the readability and understandability would decrease very quickly.

Interaction with visualization is thus necessary to allow freedom to the user. Shneiderman proposed a strategy (a "mantra") for visual information seeking that included several activities and features that a visualization tool must have[39]. Specifically, Shneiderman's mantra is "Overview first, zoom and filter, details on demand". The overview provides a summary of the database content. With view manipulation tools, such as filter (choosing the variables and the inclusion/exclusion criteria) and zoom (deciding the level of detail), the user can customize the perspective and the subset of data to analyze. The last step is the detail, which is on-demand and is given when a reasonable number of items are available for the analysis.

Addressing the user's mental model reduces the time taken to get the required information. But when the data is too big, the overview may not be feasible. Also ,when the number of variables is too big, the human might not be able to manage them. Daniel Keim revised the Shneiderman's mantra introducing the Visual Analytics Mantra, as "Analyse first, Show the important, Zoom, filter and analyze further, Details on demand"[40]. The difference in the new

[39] Ben Shneiderman. 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In Proceedings of the 1996 IEEE Symposium on Visual Languages (VL '96). IEEE Computer Society, USA, 336.

[40] Keim, D., Andrienko, G., Fekete, JD., Görg, C., Kohlhammer, J., Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges. In: Kerren, A., Stasko, J.T., Fekete, JD., North, C. (eds)

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

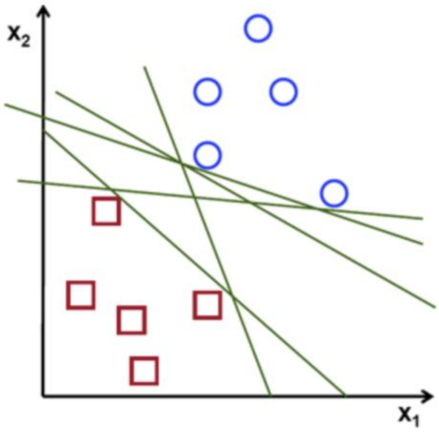
mantra is that the user doesn't choose the most interesting area, but it is an intelligent system that decides what to see, according to given heuristics, tasks, or AI prompts. Here, the concept of 'importance' is relevant because it is necessary to provide the system with this notion in order to train the system to select important data. Another difference is that here it is more explicit the iteration, which is performed in collaboration between the user and the system.

Since the analysis is driven by needs, the analyst should decide what to check and what to see, so the user must be involved in the analytical process. In order to better exploit the user perception, the visualization can use eight visual variables (marks and channels): position, mark, size, brightness, color, orientation, texture, and motion. Each one is more suitable for specific tasks. For instance, the position is well-exploited when the user wants to perceive precision values in the data space.

2.6.1 State of the Art

GenDAI is characterized by AI, which includes Machine learning. In the literature, there are specific visualization techniques that have been introduced to support Machine learning analysis. Examples are Support Vector Machines (SVM), Decision trees, Random Forest (RF), and Clustering.

Given an n-dimensional data space, SVMs have the goal of finding the best hyperplane that splits the data space in the n dimensions. In general, the number of hyperplanes can be infinite. The following figure shows an example of infinite lines dividing the two sets of marks (circles and squares) on a 2D plane.

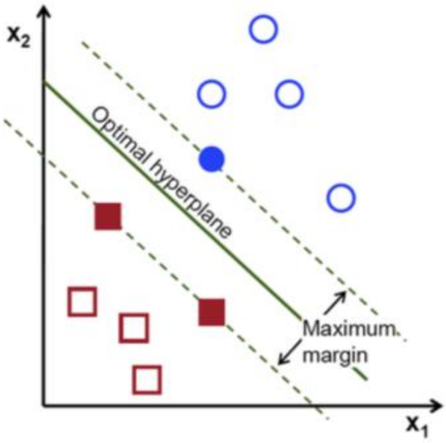


SVM exploits a set of mathematical functions called a kernel, which has the task of finding the optimal hyperplane. An example of an optimal hyperplane is indicated in the next figure, which shows sucha hyperplane surrounded by a max margin, which denotes the distance between the two sets. More specifically, the colored marks are those that compose the support vector, which is the vector of items that contribute to the identification of the optimal hyperplane.

Information Visualization. Lecture Notes in Computer Science, vol 4950. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70956-5_7

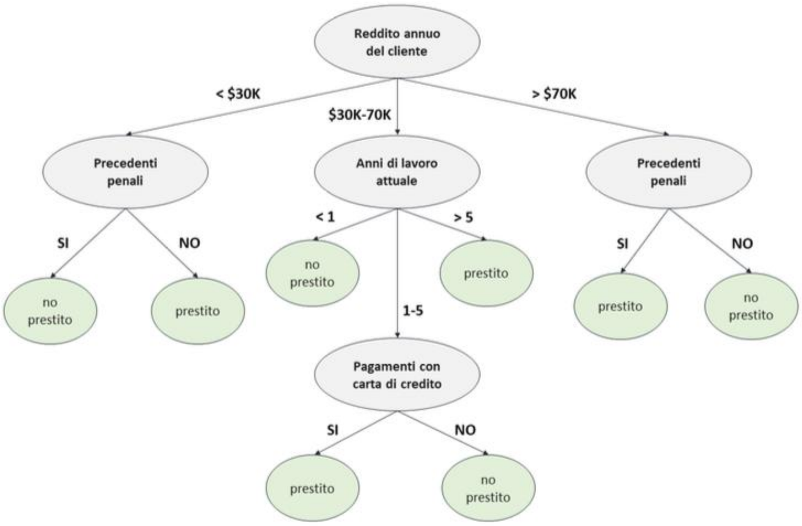
This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union



SVMs are well-suited for Natural Language processing, image recognition, and computer vision. The main problem of this approach is that in real-world examples is difficult to have a full separation of items, but they are efficient to train and provide good results on many datasets.

Decision Trees (DT) are ta ree data structure whose node splits data set into two subsets according to a given condition (called a predicate). The split is repeated for each node of the tree. This approach is very efficient and deterministic, but the tree must be balanced to optimize the time to process the information. The example in the following figure presents a simplified version of a decision tree where the nodes are weighted, but the decision tree also works without the weights.



While the decision tree can go through all nodes to find the optimal path to get to the solution, Random Forest, as the name suggests, creates a random forest of decision trees ot improve the decision capacity. A random set of features is selected according to their separation capability of item classes. The selected features are also called feature bagging, while the

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801
Funded by the European Union

excluded features are called out-of-bag. Even if RF is typically more accurate than DT, it is also computationally more expensive, and is also difficult to explain.

When there is a need to classify items, it is almost obvious to think about clustering them according to some given criteria and using known algorithms. Clustering algorithms create groups of similar data so that it is more evident their similarity. This concept can be easy for a very small number of variables, but becomes a quickly complex topic when dealing with more variables. To make things worse, the similarity is not transitive, that means that: if A and B are similar, and B and C are similar, it is not obvious that A and C are similar, and in general they are not.

Other visualizations focus on metrics for ML assessment. The most used and classic metrics are Precision, Recall, F1 score, and G-score.

In microbiome data analysis, techniques can be related to the data structure, such as hierarchical and network data, asl well as the temporal data. These three basic data types can also be combined, making it challenging to visualize them. Indeed, there are a few techniques and tools in the literature that are capable of visualizing dynamic hypergraphs, which are techniques that show complex relationships that change over time in a single visualization.

The scatterplot can be improved by composing scatterplot matrices and thus adding one or more variables to the analysis.

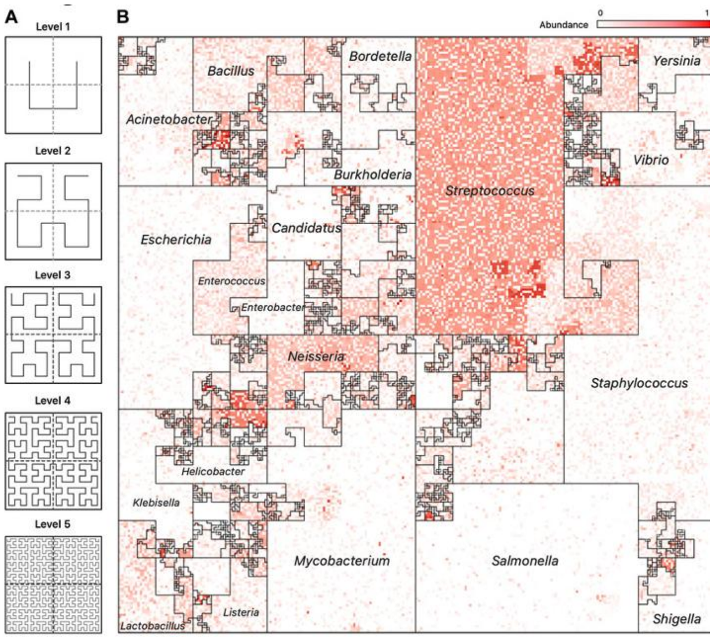
Almost all techniques can then be represented in a Cartesian plane with linear representations. Specifically in biology, there is the option to represent data in a radial space. Radial representations can better reveal direct random connections among data and cyclic behaviors.

Visualization techniques and tools of the microbiome have been proposed in the literature. Maps can be useful to get an overview of the data. Valdes et al propose a visualization that exploits the Hilbert Curve[41]. In the following figure, A (left) shows how the map is created, while B (right) shows the map of a mWGS Buccal Mucosa sample, where the color intensity of each position in the image represents the abundance of one microbial genome, the bordered regions represent groups of related microbes and their size corresponds to the number of genomes in the reference collection.

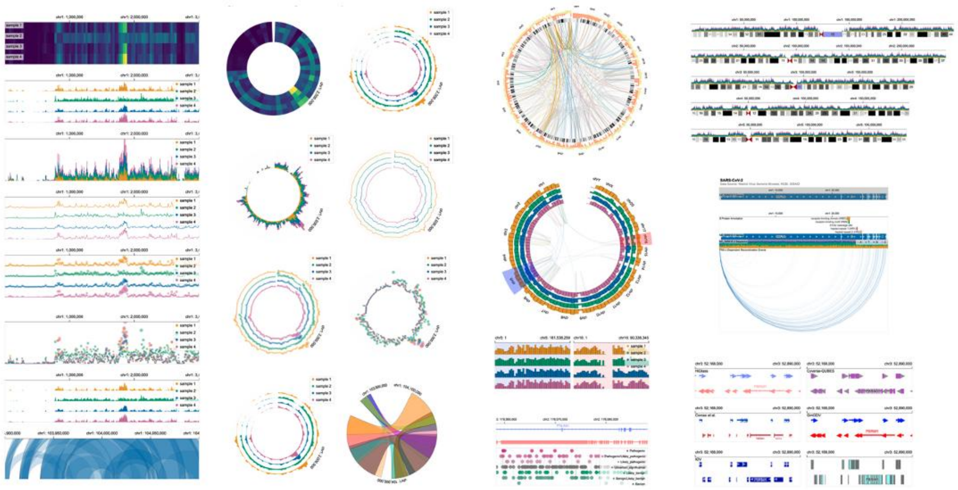
[41] Valdes, C., Stebliankin, V., Park, J. I., Lee, H., & Narasimhan, G. (2023). Microbiome maps: Hilbert curve visualizations of metagenomic profiles. Frontiers in Bioinformatics, 3, 1154588. https://doi.org/10.3389/fbinf.2023.1154588

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union



Other proposals explore different techniques to perform different tasks. For instance, GOS[42] is a declarative genomics visualization library designed for genomics visualization. According to several factors, such as the data or the task, the visualization can be linear, circular, or based on glyphs. In the next figure, there are a few examples of different visualization techniques that are used in biology.



When a user asks for a diagnosis, he or she delivers a sample to analyze (a doctor could do this on user's behalf). Lab reports are typically composed of a list of parameters that are

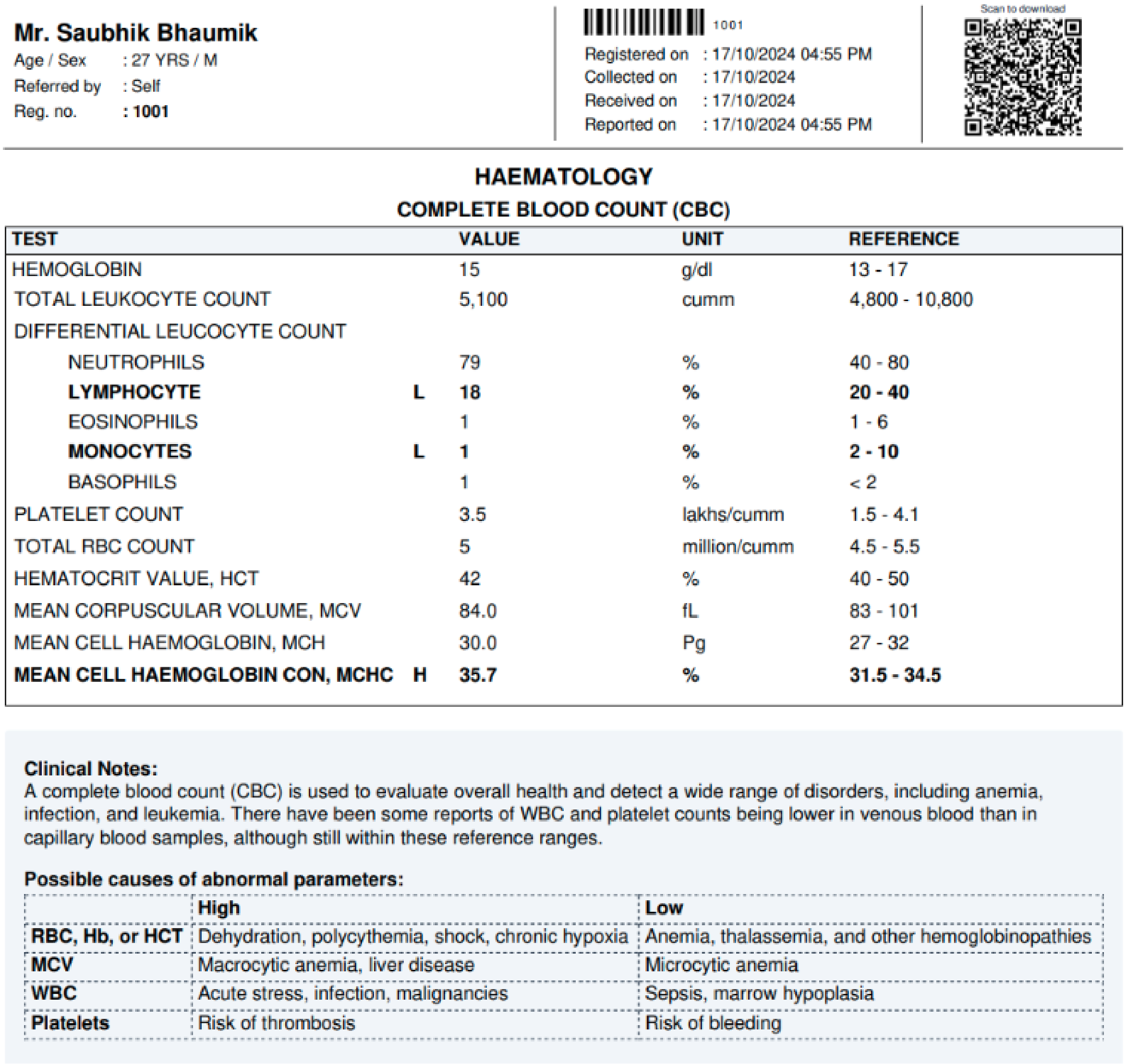
[42] Manz T, L'Yi S, Gehlenborg N. Gos: a declarative library for interactive genomics visualization in Python. Bioinformatics. 2023 Jan 1;39(1):btad050. doi: 10.1093/bioinformatics/btad050. PMID: 36688709; PMCID: PMC9891240. https://github.com/gosling-lang/gos

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

analyzed, their values detected, and the range of validity, with the indication of normal range and levels of values that are outside the typical range or the safe range.

One of the most know test reports is the blood test, which contains information related to the health status of the patient. The problem is tha,t still today, a blood test looks more similar to an after WWII secret file than a human-understandable report. The following figure[43] reports a blood test of a fictional person with the list of parameters tested, the unit, and the reference range. The user can read immediately that three parameters are outside the recommended range, two have low values (lymphocyte and monocytes) and one has high values (mean cell haemoglobin con, mchc) because, close to the value column there is an L or an H where the value is Lower or Higher than the recommended range.



Mr. Saubhik Bhaumik

Age / Sex	: 27 YRS / M
Referred by	: Self
Reg. no.	**: 1001**

1001

Registered on	: 17/10/2024 04:55 PM
Collected on	: 17/10/2024
Received on	: 17/10/2024
Reported on	: 17/10/2024 04:55 PM

Scan to download

HAEMATOLOGY

COMPLETE BLOOD COUNT (CBC)

TEST		VALUE	UNIT	REFERENCE
HEMOGLOBIN		15	g/dl	13 - 17
TOTAL LEUKOCYTE COUNT		5,100	cumm	4,800 - 10,800
DIFFERENTIAL LEUCOCYTE COUNT				
NEUTROPHILS		79	%	40 - 80
LYMPHOCYTE	**L**	**18**	**%**	**20 - 40**
EOSINOPHILS		1	%	1 - 6
MONOCYTES	**L**	**1**	**%**	**2 - 10**
BASOPHILS		1	%	< 2
PLATELET COUNT		3.5	lakhs/cumm	1.5 - 4.1
TOTAL RBC COUNT		5	million/cumm	4.5 - 5.5
HEMATOCRIT VALUE, HCT		42	%	40 - 50
MEAN CORPUSCULAR VOLUME, MCV		84.0	fL	83 - 101
MEAN CELL HAEMOGLOBIN, MCH		30.0	Pg	27 - 32
MEAN CELL HAEMOGLOBIN CON, MCHC	**H**	**35.7**	**%**	**31.5 - 34.5**

Clinical Notes:
A complete blood count (CBC) is used to evaluate overall health and detect a wide range of disorders, including anemia, infection, and leukemia. There have been some reports of WBC and platelet counts being lower in venous blood than in capillary blood samples, although still within these reference ranges.

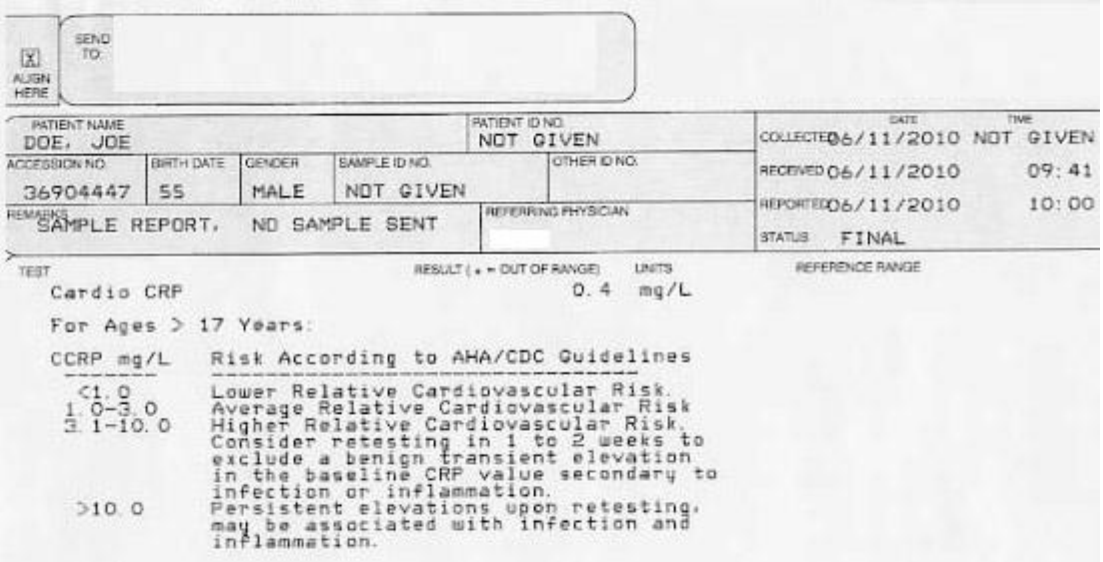
Possible causes of abnormal parameters:

	High	**Low**
RBC, Hb, or HCT	Dehydration, polycythemia, shock, chronic hypoxia	Anemia, thalassemia, and other hemoglobinopathies
MCV	Macrocytic anemia, liver disease	Microcytic anemia
WBC	Acute stress, infection, malignancies	Sepsis, marrow hypoplasia
Platelets	Risk of thrombosis	Risk of bleeding

Information is Beautiful website reports the result of a competition in proposing a visualization of blood tests that overcomes the problem of being capable of reading, and possibly

[43] Blood count report example: https://www.labsmartlis.com/cbc-report-format, last visit 25.06.25

understanding the blood test. The challenge was to make the cholesterol level clear. The provided report is the one in the following figure, completely text-based.



☒ ALIGN HERE

SEND TO:

PATIENT NAME DOE, JOE			PATIENT ID NO. NOT GIVEN		
ACCESSION NO. 36904447	BIRTH DATE 55	GENDER MALE	SAMPLE ID NO. NOT GIVEN	OTHER ID NO.	
REMARKS SAMPLE REPORT. NO SAMPLE SENT				REFERRING PHYSICIAN	

	DATE	TIME
COLLECTED	06/11/2010	NOT GIVEN
RECEIVED	06/11/2010	09:41
REPORTED	06/11/2010	10:00
STATUS	FINAL	

TEST	RESULT (* = OUT OF RANGE)	UNITS	REFERENCE RANGE
Cardio CRP	0.4	mg/L	

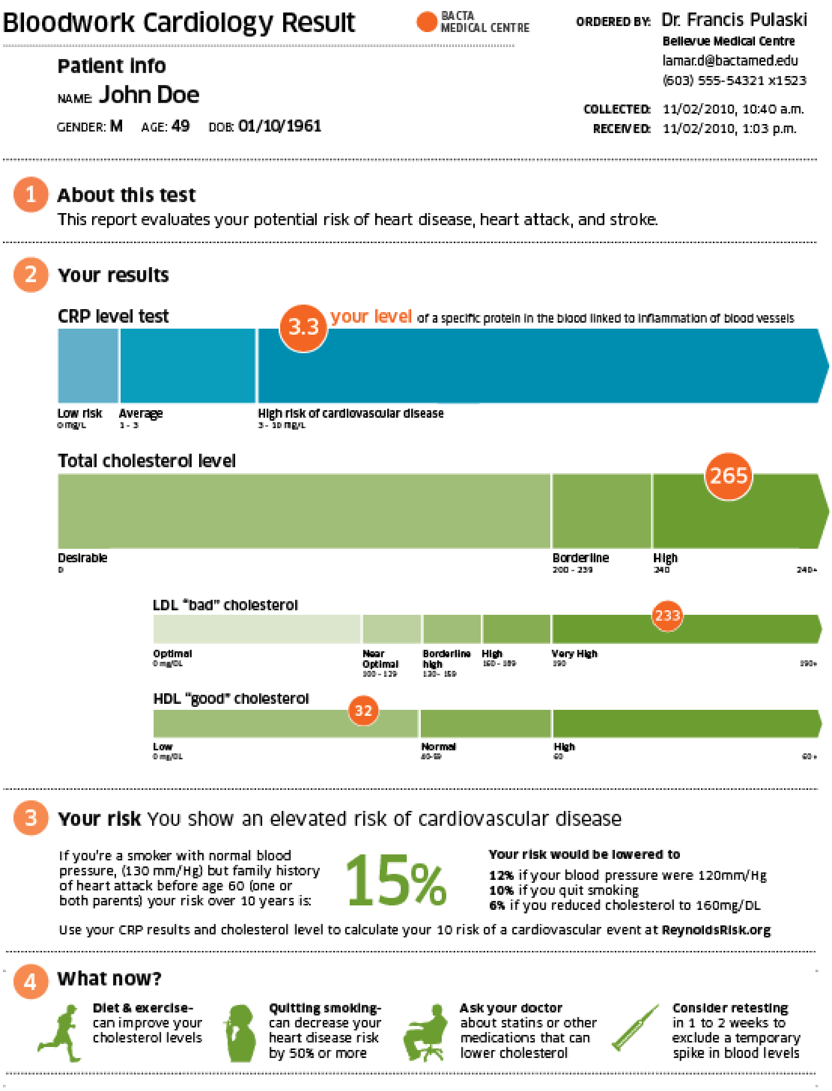
```
For Ages > 17 Years:

CCRP mg/L   Risk According to AHA/CDC Guidelines
---------   ------------------------------------
  <1.0      Lower Relative Cardiovascular Risk.
 1.0-3.0    Average Relative Cardiovascular Risk
 3.1-10.0   Higher Relative Cardiovascular Risk.
            Consider retesting in 1 to 2 weeks to
            exclude a benign transient elevation
            in the baseline CRP value secondary to
            infection or inflammation.
  >10.0     Persistent elevations upon retesting,
            may be associated with infection and
            inflammation.
```

The authors redesigned the report into the following one

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union



Bloodwork Cardiology Result

BACTA MEDICAL CENTRE

ORDERED BY: Dr. Francis Pulaski
Bellevue Medical Centre
lamar.d@bactamed.edu
(603) 555-54321 x1523

COLLECTED: 11/02/2010, 10:40 a.m.
RECEIVED: 11/02/2010, 1:03 p.m.

Patient Info

NAME: John Doe

GENDER: M AGE: 49 DOB: 01/10/1961

1 About this test

This report evaluates your potential risk of heart disease, heart attack, and stroke.

2 Your results

CRP level test

3.3 your level of a specific protein in the blood linked to inflammation of blood vessels

Total cholesterol level

LDL "bad" cholesterol

HDL "good" cholesterol

3 **Your risk** You show an elevated risk of cardiovascular disease

If you're a smoker with normal blood pressure, (130 mm/Hg) but family history of heart attack before age 60 (one or both parents) your risk over 10 years is: 15%

Your risk would be lowered to

12% if your blood pressure were 120mm/Hg
10% if you quit smoking
6% if you reduced cholesterol to 160mg/DL

Use your CRP results and cholesterol level to calculate your 10 risk of a cardiovascular event at **ReynoldsRisk.org**

4 What now?

Diet & exercise- can improve your cholesterol levels

Quitting smoking- can decrease your heart disease risk by 50% or more

Ask your doctor about statins or other medications that can lower cholesterol

Consider retesting in 1 to 2 weeks to exclude a temporary spike in blood levels

David McCandless & Stefanie Posavec for Wired Magazine // informationisbeautiful.net

The new version, which can be further improved, contains at the top the same personal information of the patient, the reason for reading the test, and then the results in graphical form. The bars and the trend line are evident, so the value of the patient. Notice that the orange text "your level" recalls the color of the orange circle that is positioned on the bar and contains the value.

The bar is split into different parts, each representing a different interval with its meaning. For instance, the patient has the value of 265 mg/DL, which is in the high-risk area. The report also contains further indications such as LDL "bad" cholesterol, because people don't easily remember the difference between LDL and HDL.

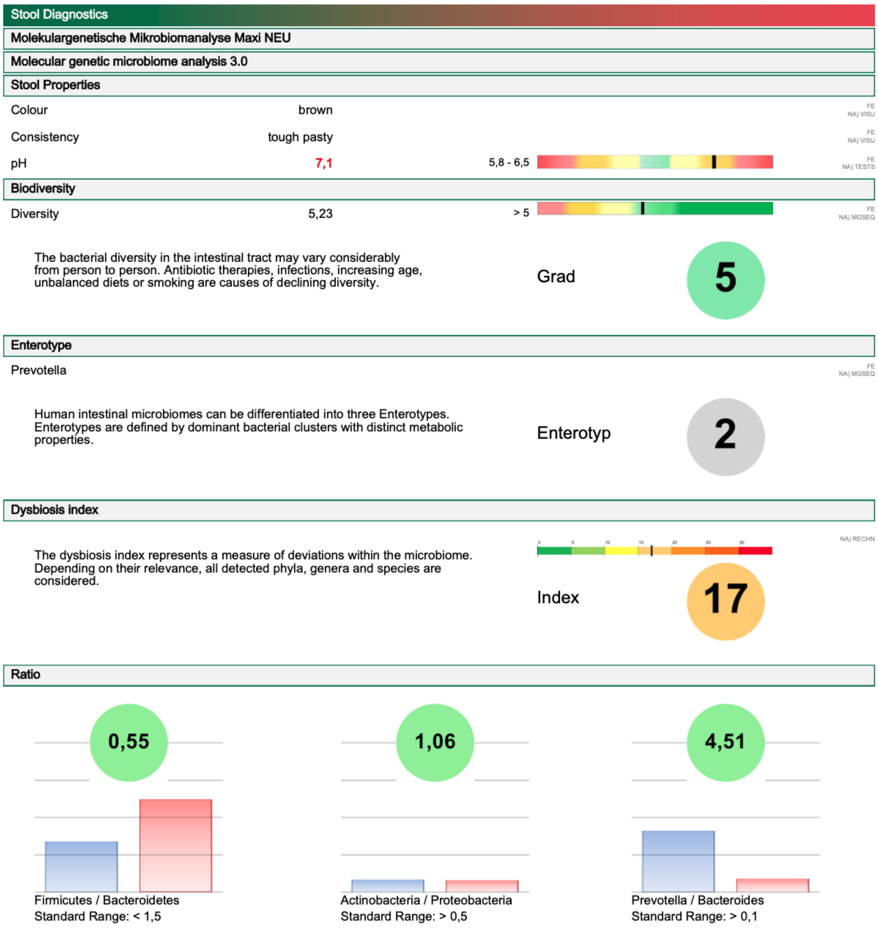
At the bottom of the page, there is the diagnosis: elevated risk of cardiovascular disease, and what to do.

In the case of biological visualization, the situation is similar, and a bit more complex due to the many variables the lab should take into account. A reference starting point is the report

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

produced by BioVis diagnostik, which contains, as it happens for the blood test, the personal data of the patient at the topo, then the test performed and, for each the results, the unit, the standard range and possible previous results, for an easier comparison. This is already a quite well-presented report that can still be improved in various ways.



Stool Diagnostics

Molekulargenetische Mikrobiomanalyse Maxi NEU

Molecular genetic microbiome analysis 3.0

Stool Properties

Colour	brown		FE NA) VISU
Consistency	tough pasty		FE NA) VISU
pH	7,1	5,8 - 6,5	FE NA) TESTS

Biodiversity

Diversity	5,23	> 5	FE NA) MGSEQ

The bacterial diversity in the intestinal tract may vary considerably from person to person. Antibiotic therapies, infections, increasing age, unbalanced diets or smoking are causes of declining diversity.

Grad 5

Enterotype

Prevotella — FE NA) MGSEQ

Human intestinal microbiomes can be differentiated into three Enterotypes. Enterotypes are defined by dominant bacterial clusters with distinct metabolic properties.

Enterotyp 2

Dysbiosis index

The dysbiosis index represents a measure of deviations within the microbiome. Depending on their relevance, all detected phyla, genera and species are considered.

NA) RECHN

Index 17

Ratio

0,55 — Firmicutes / Bacteroidetes — Standard Range: < 1,5

1,06 — Actinobacteria / Proteobacteria — Standard Range: > 0,5

4,51 — Prevotella / Bacteroides — Standard Range: > 0,1

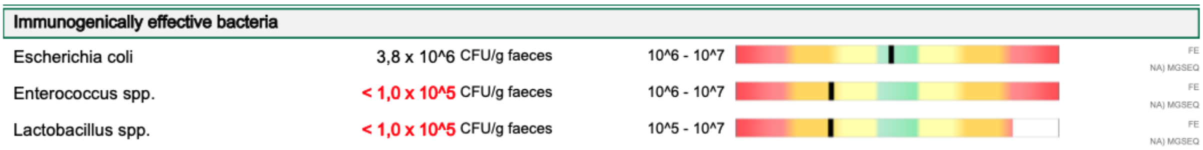
The previous image shows a portion of a sample report of stool diagnostics. In the report, there are several analyses whose results are reported similarly to the just-mentioned *Information is Beautiful* proposal. It is visible that the pH of the stool is too high because the value is 7,1. In the report, too high or too low is assumed to be the same red color, so to know if it is too high, it is necessary to look at the bar on the right. The bar shows the ideal range in green, then the traffic light colors reveal the other ranges that are progressively farther from such a range, and a black vertical rectangle reveals the position of the value.

The biodiversity of bacteria is indicated as a number and with a scale similar to the pH, with the difference that while the pH has a diverging scale (acid vs basic), the biodiversity is an index of variability of bacteria; the more diversity, the better, and the minimum value is 5. However, in this patient, the value is 5.23, which is bigger than 5, so the user should read the number in all cases. Here, the well-visible indicators (the big numbers, for instance) are a quick reference for the user. Indeed, after a few initial pages that contain the short information in a graphical form, there are many pages that explain in detail the numbers.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

In the following image, three items are reported (Ascherichia coli, Enterococcus spp., and Lactobacillus spp.); one is in the normal range, while the other two have a low value.



Immunogenically effective bacteria			
Escherichia coli	3,8 x 10^6 CFU/g faeces	10^6 - 10^7	FE NA) MGSEQ
Enterococcus spp.	< 1,0 x 10^5 CFU/g faeces	10^6 - 10^7	FE NA) MGSEQ
Lactobacillus spp.	< 1,0 x 10^5 CFU/g faeces	10^5 - 10^7	FE NA) MGSEQ

The explanation is provided far below in the document, and it is also not focused on the three values belonging to immunogenically effective bacteria. All three of them are described under the Proteobacteria section, which is a different section than the chart above.

2.6.2 Remaining Challenges / Requirements

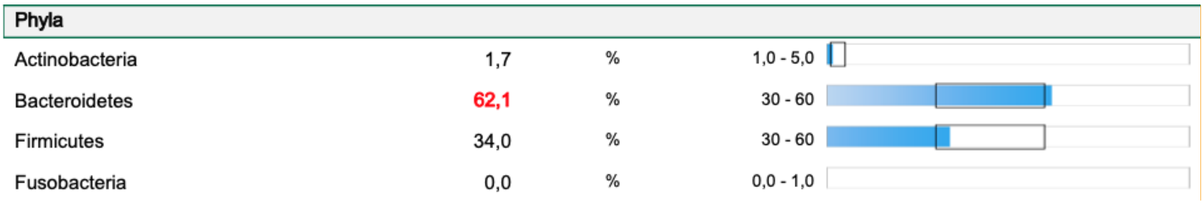
The state of the art reports the attention that the most innovative proposals have towards the understandability of the user. The current implementation, however, can be improved in different ways to make the report completely understandable by the user and to let the user read the results more quickly.

There are different improvements that can be done both to the report structure and to the specific visualizations. The report currently contains two main sections: at the beginning, a synthesis of the diagnosis that reports the raw data and the basic indication of the ranges. A second, and bigger section reports the more detailed report that explains the results and provides a more comprehensive description and diagnosis.

Instead of having just two sections, since the report is digital, is it possible to create an improved report that allows browsing the content in order to get to the most interesting information. The interest of the user is on the parameters that are not in the expected ranges, and the explanation of why this happens and what the implications are. The user reads in the first part, which is composed of a few (e.g., 4) page,s the parameters, and then jumps into the longer part to seek the explanation. Instead, the report can provide a section of the parameters out of range, in groups composed of semantically close parameters, and an explanation close to those groups.

In this way, the report is bigger, but easier to read. Another improvement is in the provided visualizations. Currently ,the charts are good but can be better in giving consistent representation and the right context in which the data is obtained.

Not only the representation but also the reported scale should be considered. Indeed, in the following example, four bacteria are reported, and on the right are the values for the patient. Here, the recommended ranges are very different: [1-5], [30-60], [0-1]. If we consider cholesterol, the difference in scale could be even worse [0-600].



Phyla			
Actinobacteria	1,7	%	1,0 - 5,0
Bacteroidetes	62,1	%	30 - 60
Firmicutes	34,0	%	30 - 60
Fusobacteria	0,0	%	0,0 - 1,0

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Different representations must be investigated to improve the readability of values that are intrinsically small or big, as well as the representation of suggested ranges.

2.7 Regulatory and Ethical Considerations

A comprehensive framework for all regulatory and ethical considerations within the GenDAI project is established in the dedicated deliverable D7.1: Data Management Plan, Knowledge Management Resources, and Quality & Risk Management Plan.

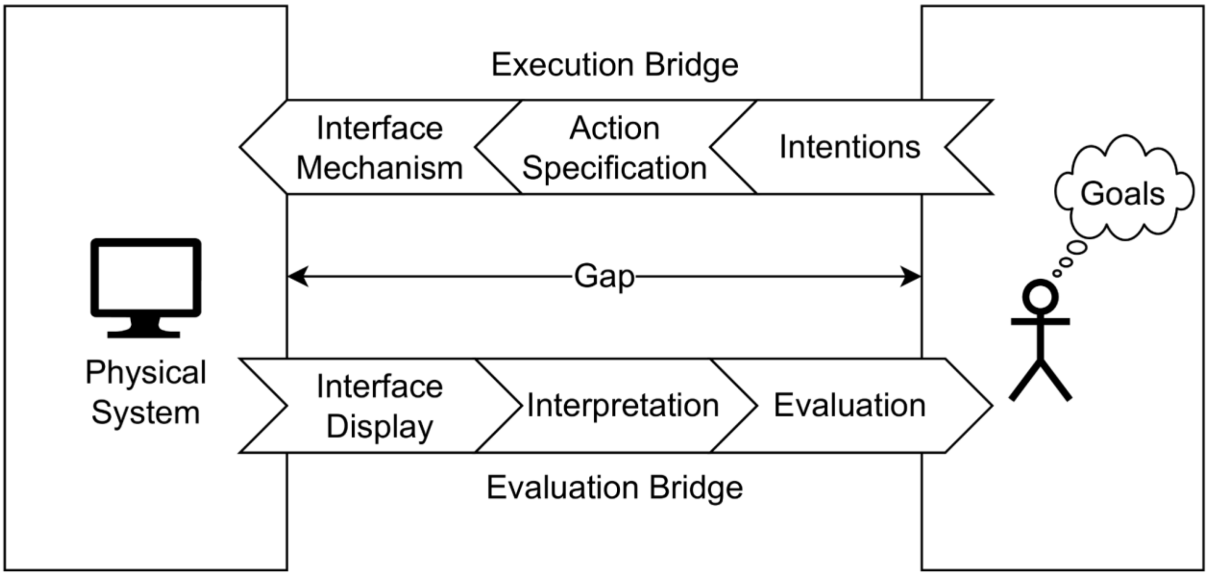
This plan details the project's commitment to responsible data handling, including strict adherence to the FAIR principles (Findable, Accessible, Interoperable, Reusable) and full compliance with legal frameworks such as the General Data Protection Regulation (GDPR). It covers critical procedures for obtaining informed consent, implementing data anonymization and pseudonymization, ensuring robust data security, and managing project-wide quality and risks.

As these topics are extensively covered in D7.1, they will not be detailed further in this document. All requirements and models presented herein are designed to be fully compliant with the principles outlined in the Data Management Plan.

3. Conceptual Design and Modelling

The methodology employed in this project is grounded in the principles of User-Centered System Design (UCSD), as described by Norman and Draper. UCSD emphasizes placing users and their tasks at the core of the design process, ensuring that the system is aligned with user goals and workflows. This approach views the interaction between users and systems as a process of bridging a gap: users must transform their psychological goals into physical actions that can be performed using the system's interface, while the system must effectively translate these actions into feedback that helps users evaluate whether their goals have been achieved.

This interaction can be conceptualized as two bridges. The first, the execution bridge, represents how users convert their intentions into system actions, starting with goal formulation, followed by action specification, and culminating in execution. The second, the evaluation bridge, illustrates how users interpret the system's feedback and determine whether it meets their initial goals. Narrowing these gaps is critical to creating intuitive, user-friendly systems.



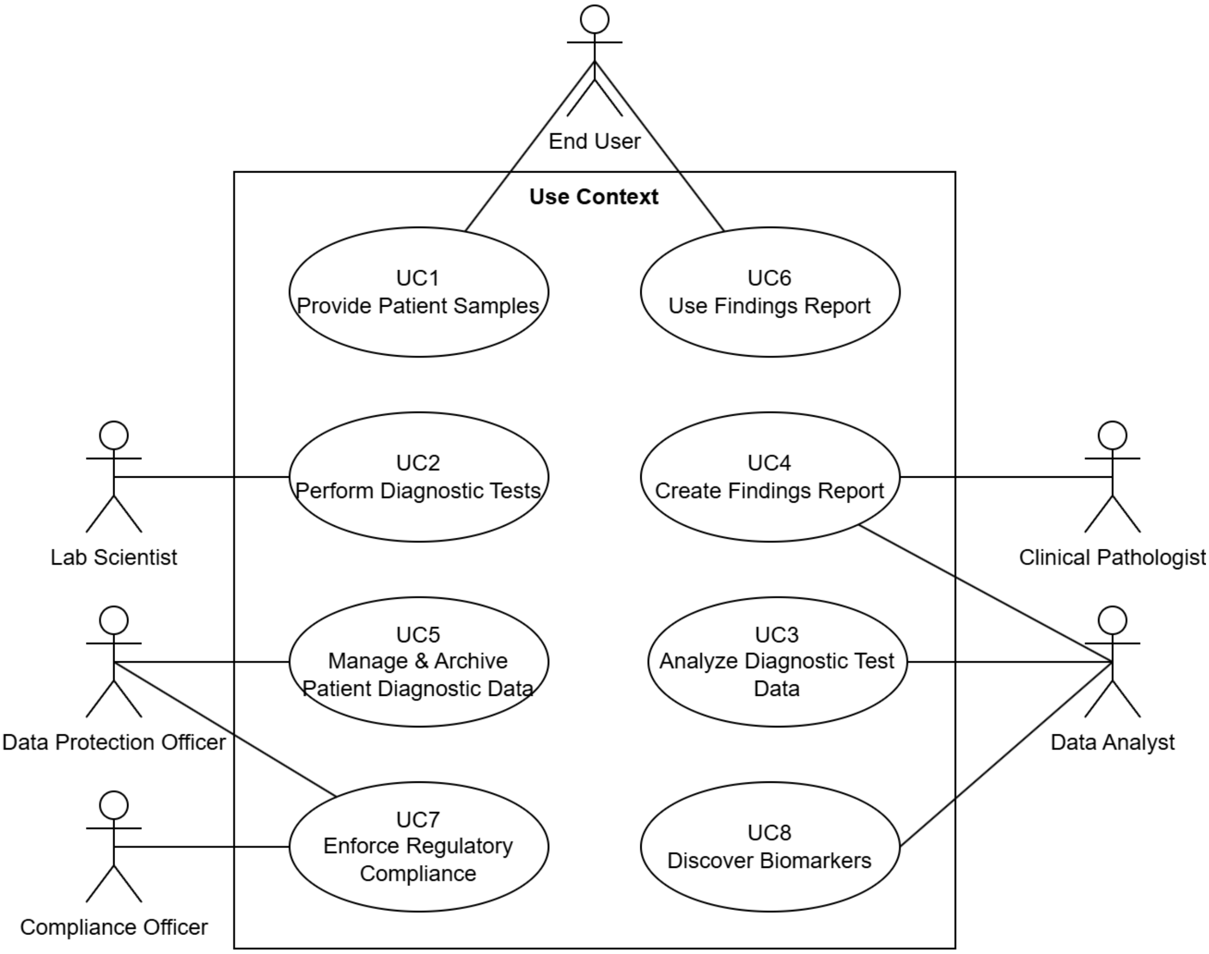
To address this, UCSD advocates for a common conceptual model that represents both the users' mental models and the system's interface mechanisms. Developing this model involves studying the use context through methods such as interviews and observations to understand user goals, tasks, and workflows. In this project, these insights informed the design of use cases and workflows, which were detailed through modeling techniques such as UML diagrams. This process ensured a shared understanding of the system's requirements.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.1 Overview and Use Context

3.1.1 Use Context



3.1.2 UC1 Provide Patient Samples

Actor: End User (e.g., patient or physician on patient's behalf)

- The End User submits one or more patient samples (blood, tissue, swab, etc.) to the laboratory.
- Basic patient or sample metadata is registered so the lab can track and process the samples.
- Ensures sample integrity (e.g., correct labeling, proper storage conditions).

3.1.3 UC2 Perform Diagnostic Tests

Actor: Lab Scientist

- The Lab Scientist follows the approved protocol to run tests

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

- They document the steps taken, verify sample quality, and collect any measurement data produced by instruments.
- Results are then handed off to the next stage (analysis, archiving) or flagged for re-testing if there are quality issues.

3.1.4 UC3 Analyze Diagnostic Test Data

Actor: Data Analyst

- The Data Analyst receives raw diagnostic data, performs data cleaning, and applies various analysis tasks (e.g., statistical analysis, clustering, machine learning).
- They interpret patterns or anomalies, perform quality checks, and may consult with domain experts as needed.
- Relevant findings are used to inform the final diagnostic report

3.1.5 UC4 Create Findings Report

Actors: Data Analyst, Clinical Pathologist

- After the analysis is complete, the Data Analyst generates a preliminary report summarizing the key results, interpretations, and any recommended actions.
- The Clinical Pathologist reviews the report, checks for consistency and correctness, and finalizes it before releasing it to the ordering physician or patient records.

This step may involve interactive data visualization or review tools to ensure accuracy and clarity of the final output.

3.1.6 UC5 Manage & Archive Patient Diagnostic Data

Actor: Data Protection Officer (and possibly others, e.g., Lab Scientist or IT staff)

- This use case focuses on safely storing, archiving, and retrieving patient-related diagnostic data.
- The Data Protection Officer ensures that data management complies with privacy regulations and internal security policies.
- Activities include setting up secure storage, access rights, audit trails, defining proper PID for various entities in the system, and data-retention procedures.

3.1.7 UC6 Use Findings Report

Actor: End User (e.g., physician, or even the patient in some contexts)

- Once the diagnostic results are available, the End User reviews or discusses them.
- The insights in the findings report may inform treatment decisions or further follow-up testing.
- The End User may provide feedback or clarification requests if the results are ambiguous.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.1.8 UC7 Enforce Regulatory Compliance

Actors: Compliance Officer, Data Protection Officer

- The Compliance Officer oversees the entire diagnostic process, making sure every stage complies with relevant regulations (quality assurance, privacy laws, medical device regulations, etc.).
- If any breaches or deviations are found, the Compliance Officer initiates corrective actions
- The Data Protection Officer ensures that privacy and data-handling policies are enforced, collaborating with the Compliance Officer on regulatory audits and reporting, and implementing communication obligations about the data breach to relevant authorities.

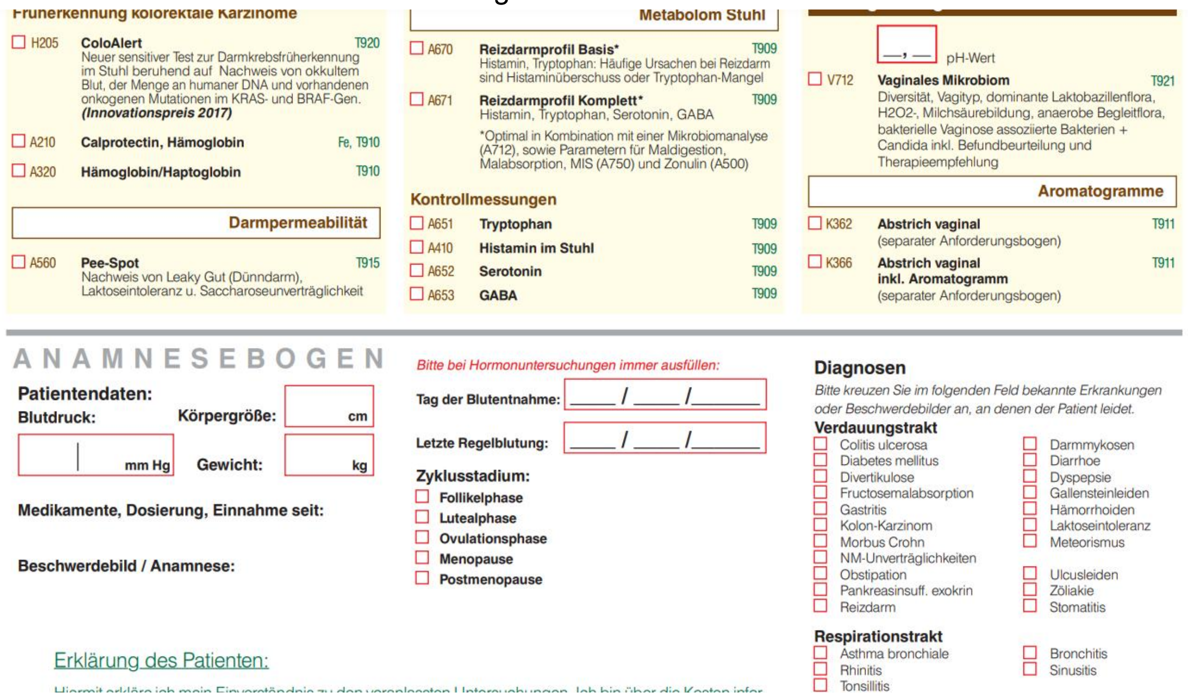
3.1.9 UC8 Discover Biomarkers

Actor: Data Analyst

- The Data Analyst runs specialized algorithms or machine learning tools on aggregated, large-scale datasets to identify novel biomarkers.
- Potential biomarkers may help refine or develop new diagnostic tests and improve future clinical decision-making.
- Once potential biomarkers are found, they are documented and passed to clinical or lab staff to develop new test protocols.

3.2 Provide Patient Samples (UC1)

To start the process of IBS diagnostics, a stool sample from the patient needs to be collected and an order form such as the following needs to be filled out.



Früherkennung kolorektale Karzinome

□ H205 **ColoAlert** T920
Neuer sensitiver Test zur Darmkrebsfrüherkennung im Stuhl beruhend auf Nachweis von okkultem Blut, der Menge an humaner DNA und vorhandenen onkogenen Mutationen im KRAS- und BRAF-Gen. ***(Innovationspreis 2017)***

□ A210 **Calprotectin, Hämoglobin** Fe, T910

□ A320 **Hämoglobin/Haptoglobin** T910

Darmpermeabilität

□ A560 **Pee-Spot** T915
Nachweis von Leaky Gut (Dünndarm), Laktoseintoleranz u. Saccharoseunverträglichkeit

Metabolom Stuhl

□ A670 **Reizdarmprofil Basis*** T909
Histamin, Tryptophan: Häufige Ursachen bei Reizdarm sind Histaminüberschuss oder Tryptophan-Mangel

□ A671 **Reizdarmprofil Komplett*** T909
Histamin, Tryptophan, Serotonin, GABA

*Optimal in Kombination mit einer Mikrobiomanalyse (A712), sowie Parametern für Maldigestion, Malabsorption, MIS (A750) und Zonulin (A500)

Kontrollmessungen

□ A651 **Tryptophan** T909
□ A410 **Histamin im Stuhl** T909
□ A652 **Serotonin** T909
□ A653 **GABA** T909

__,__ pH-Wert

□ V712 **Vaginales Mikrobiom** T921
Diversität, Vagityp, dominante Laktobazillenflora, H2O2-, Milchsäurebildung, anaerobe Begleitflora, bakterielle Vaginose assoziierte Bakterien + Candida inkl. Befundbeurteilung und Therapieempfehlung

Aromatogramme

□ K362 **Abstrich vaginal** T911
(separater Anforderungsbogen)

□ K366 **Abstrich vaginal inkl. Aromatogramm** T911
(separater Anforderungsbogen)

ANAMNESEBOGEN

Patientendaten:

Blutdruck: | **mm Hg**

Körpergröße: **cm**

Gewicht: **kg**

Medikamente, Dosierung, Einnahme seit:

Beschwerdebild / Anamnese:

Bitte bei Hormonuntersuchungen immer ausfüllen:

Tag der Blutentnahme: ____ / ____ /______

Letzte Regelblutung: ____ / ____ /______

Zyklusstadium:
□ **Follikelphase**
□ **Lutealphase**
□ **Ovulationsphase**
□ **Menopause**
□ **Postmenopause**

Erklärung des Patienten:

Hiermit erkläre ich mein Einverständnis zu den veranlassten Untersuchungen. Ich bin über die Kosten infor

Diagnosen

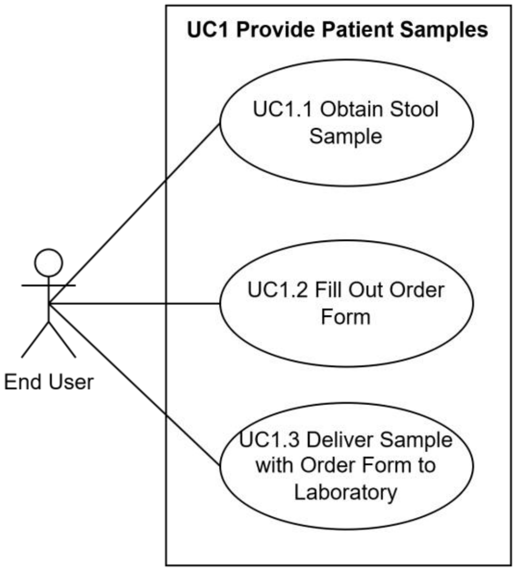
Bitte kreuzen Sie im folgenden Feld bekannte Erkrankungen oder Beschwerdebilder an, an denen der Patient leidet.

Verdauungstrakt
□ Colitis ulcerosa
□ Diabetes mellitus
□ Divertikulose
□ Fructosemalabsorption
□ Gastritis
□ Kolon-Karzinom
□ Morbus Crohn
□ NM-Unverträglichkeiten
□ Obstipation
□ Pankreasinsuff. exokrin
□ Reizdarm
□ Darmmykosen
□ Diarrhoe
□ Dyspepsie
□ Gallensteinleiden
□ Hämorrhoiden
□ Laktoseintoleranz
□ Meteorismus
□ Ulcusleiden
□ Zöliakie
□ Stomatitis

Respirationstrakt
□ Asthma bronchiale
□ Rhinitis
□ Tonsillitis
□ Bronchitis
□ Sinusitis

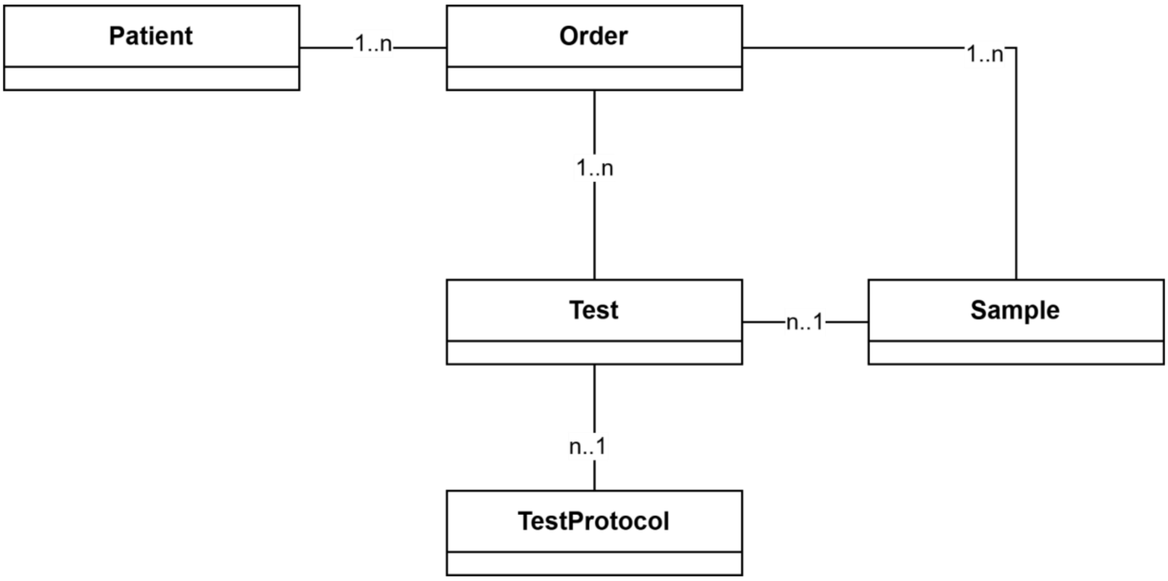
3.2.1 Sub Use Cases

From the above description, the following sub use cases can be derived:


UC1 Provide Patient Samples
UC1.1 Obtain Stool Sample
UC1.2 Fill Out Order Form
UC1.3 Deliver Sample with Order Form to Laboratory
End User

3.2.2 Partial Information Model

Based on the order form, we can also derive an initial, partial information model.



The information model describes the relationships between key entities in the stool sample diagnostic workflow. The order form, in its abstracted form, represents an **Order** for a single patient. An order thus represents a laboratory request to analyze stool samples for specific diagnostic purposes. The same order form allows ordering multiple **Tests** for a patient, as multiple analyses may be required to fulfill the diagnostic goals. Each Test is performed on exactly one **Sample**, ensuring that every analysis is tied to a specific stool specimen provided as part of the Order.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

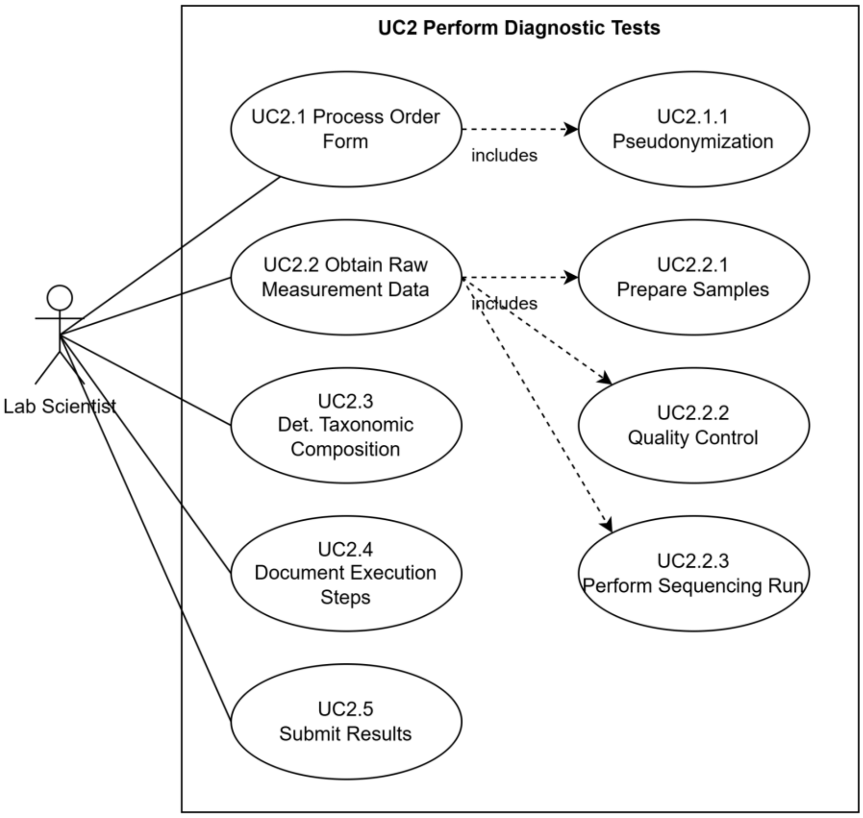
A **Sample** represents a physical stool specimen collected from the patient. While each Sample belongs to one specific Order, it may be analyzed in multiple Tests to extract different types of information or to apply complementary protocols. This allows flexibility in analyzing stool specimens for various diagnostic or research purposes without needing multiple physical samples.

The **Test** entity describes a specific diagnostic analysis performed on the stool Sample. Each Test follows a clearly defined **Test Protocol**, which outlines the standardized laboratory procedures, such as stool preparation, DNA extraction, or sequencing methods, to ensure reproducibility and accuracy. A single Test is always associated with exactly one TestProtocol, but the same TestProtocol can be applied across multiple Tests, promoting consistency across different analyses.

3.3 Perform Diagnostic Tests (UC2)

3.3.1 Sub Use Cases

The sub use cases for UC2 can be defined as follows.


UC2 Perform Diagnostic Tests
UC2.1 Process Order Form
includes
UC2.1.1 Pseudonymization
UC2.2 Obtain Raw Measurement Data
includes
UC2.2.1 Prepare Samples
Lab Scientist
UC2.3 Det. Taxonomic Composition
UC2.2.2 Quality Control
UC2.4 Document Execution Steps
UC2.2.3 Perform Sequencing Run
UC2.5 Submit Results

UC2.1 Process Order Form

The lab scientist needs to process the submitted order form to determine the requested tests and subsequent steps. As part of this process, the lab scientist also pseudonymizes the sample and all data associated with it by assigning a unique ID to the sample. This ID acts as

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

a PID for the sample. Pseudonymization techniques are applied while maintaining data usability for research and diagnostics. These privacy-enhancing measures ensure compliance with GDPR, HIPAA, ISO/IEC 27001, and IVDR 2017/746 regulations.

UC2.2 Obtain Raw Measurement Data

In this sub use case, the Lab Scientist generates raw sequencing data (e.g., FASTQ files) from the samples. Activities such as preparing the samples, ensuring they meet quality standards, and operating the sequencing instrument all fall under this umbrella. Any issues that arise—like contaminations or equipment malfunctions—are resolved here to produce high-quality raw data.

UC2.2.1 Prepare Samples

Here, the Lab Scientist verifies sample integrity (labeling, storage conditions) and performs the required preparation steps for microbiome analysis—typically extracting DNA or RNA, then preparing libraries for next-generation sequencing. Proper reagent handling, accurate pipetting, and correct instrument setup are crucial so that subsequent steps yield meaningful results.

UC2.2.2 Quality Control

After extraction and library prep, the Lab Scientist checks that samples meet defined quality parameters, such as DNA purity or library concentration. If samples fail these checks, they may be re-extracted or re-librarized until they pass. By detecting problems early, the Lab Scientist helps ensure that the eventual sequencing data will be robust and interpretable.

UC2.2.3 Perform Sequencing Run

Once sample preparation and QC are complete, the Lab Scientist loads the samples onto the sequencing instrument (e.g., Illumina or similar). They confirm correct run parameters, such as cycle count, read length, or reagent lot numbers, and initiate the run. At the end, raw reads (for example, 16S rRNA or metagenomic sequences) are exported for downstream analysis.

UC2.3 Determine Taxonomic Composition

In this sub use case, the Lab Scientist (using an automated pipeline) classifies the sequencing reads to identify which organisms are present in the microbiome sample. By mapping each read to reference databases, the pipeline yields a taxonomic profile (e.g., genus or species abundances). If the classification results appear incomplete or contradictory, the scientist may adjust parameters—like minimum read quality or taxonomic confidence thresholds—to refine the analysis.

The automated pipelines process raw diagnostic data, applying transformations such as metadata extraction, normalization, and classification. These steps ensure compatibility across clinical and research systems, improving interoperability and enabling scalable

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

analysis. Automated logs track all processing stages to ensure reproducibility and compliance with regulatory requirements.

UC2.4 Document Execution Steps

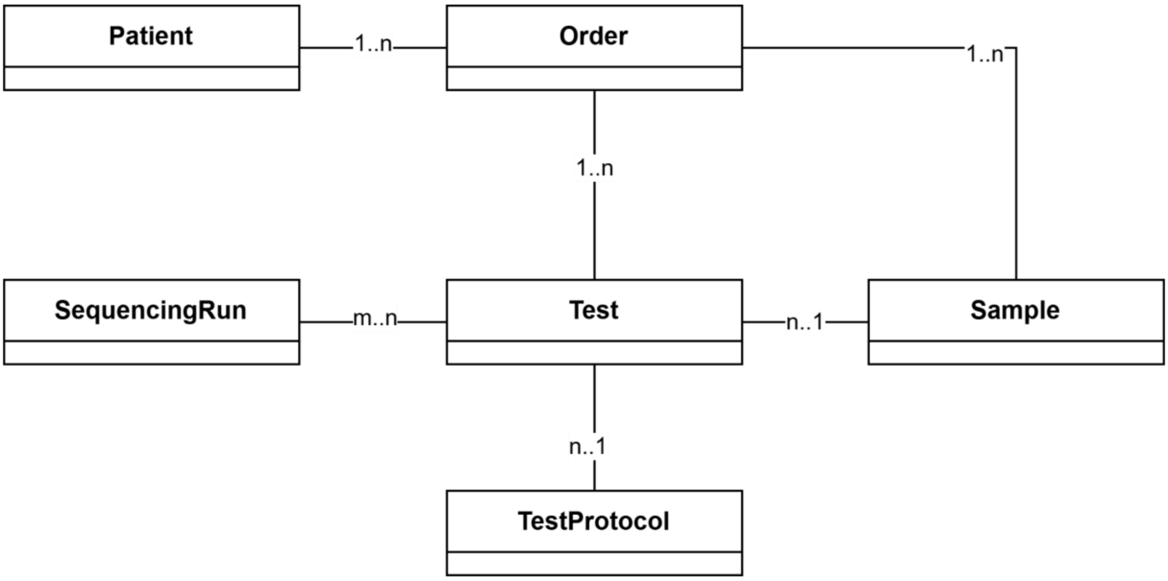
Throughout the entire diagnostic process, the Lab Scientist records essential details: timestamps, reagent lot numbers, instrument identifiers, and any deviations from the standard protocol. Clear and consistent documentation ensures each run is fully traceable and compliant with laboratory and regulatory standards. It also supports potential troubleshooting if the sequencing run or taxonomic analysis requires review later.

UC2.5 Submit Results

After the sequencing run and initial microbiome analysis, the Lab Scientist compiles the raw data and metadata into a standardized format. This package, which typically includes sequence files (FASTQ), QC metrics, and reference database details, is then transferred to the appropriate data system—often a Laboratory Information Management System (LIMS) or directly to the Data Analyst. Submitting results in a well-organized form facilitates further interpretation, archiving, or regulatory review.

3.3.2 Initial Information Model

Based on the fact that the microbiome analysis of IBD patients requires a sequencing run to determine the taxonomic composition of samples, we can add the **sequencing run** as an additional entity to the information model.



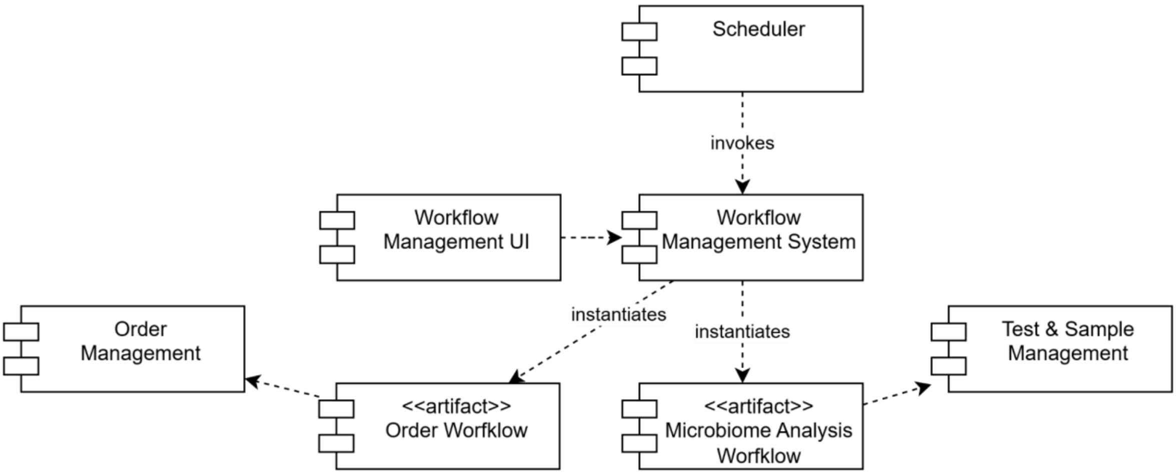
A SequencingRun represents a single batch or machine run performed on a sequencing instrument, such as an Illumina platform. Multiple Tests can be processed together in a single

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

SequencingRun, leveraging multiplexing capabilities to optimize sequencing resources. Similarly, a single Test can be part of multiple SequencingRuns if reruns or complementary sequencing methods are needed to meet quality or coverage requirements.

3.3.2 Initial Component Model



Based on the provided use cases and requirements, the following components and dependencies can be derived.

To automate the processing of orders and samples in a predictable, yet customizable way, a workflow management system (WMS) is needed. Within this WMS, at least two workflows are needed. One workflow needs to govern the activities for an individual order, which can contain multiple tests. Another workflow needs to be concerned with processing and analyzing sequencing runs, which usually contain samples from many patients. These respective workflows need to communicate with the respective subsystems for order management and for test management.

To track the progress of workflows or otherwise manage them, a workflow management UI is required. Finally, a scheduler allows workflows to be triggered at the appropriate time.

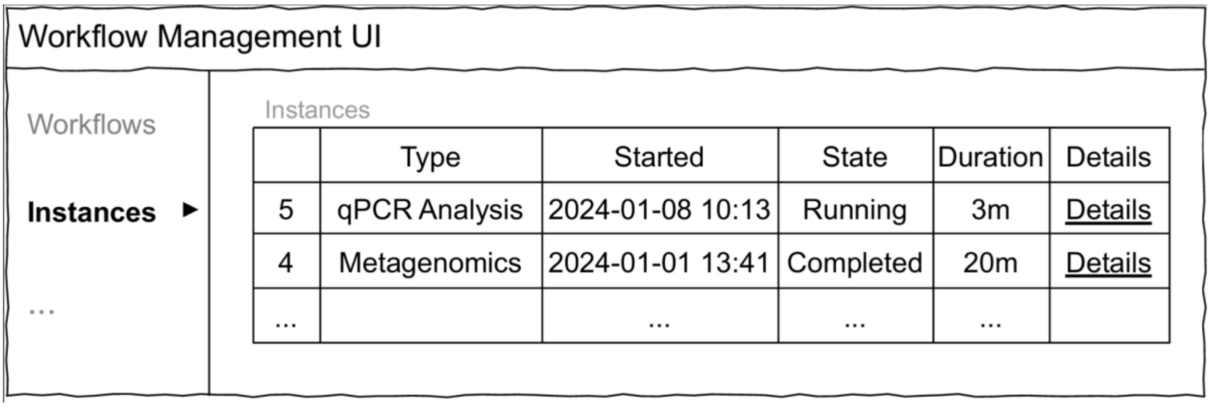
3.3.3 Visualization & UI Models

As the workflows are supposed to work automatically, no extensive user interface is required for test execution. There is however, a need for a UI to see the state and details of running workflows for introspection.

An instance view should provide an overview of recent workflows with the current state and basic information. It should also provide the option to request further details.

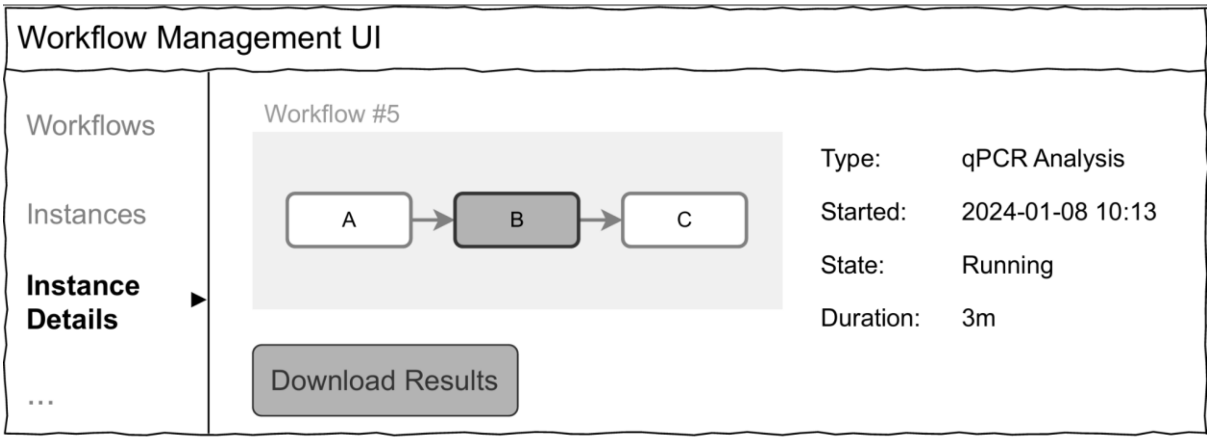
This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union



	Type	Started	State	Duration	Details
5	qPCR Analysis	2024-01-08 10:13	Running	3m	Details
4	Metagenomics	2024-01-01 13:41	Completed	20m	Details
…		…	…	…	

The Instance details view should provide more detailed information, such as the current workflow step that is executed.

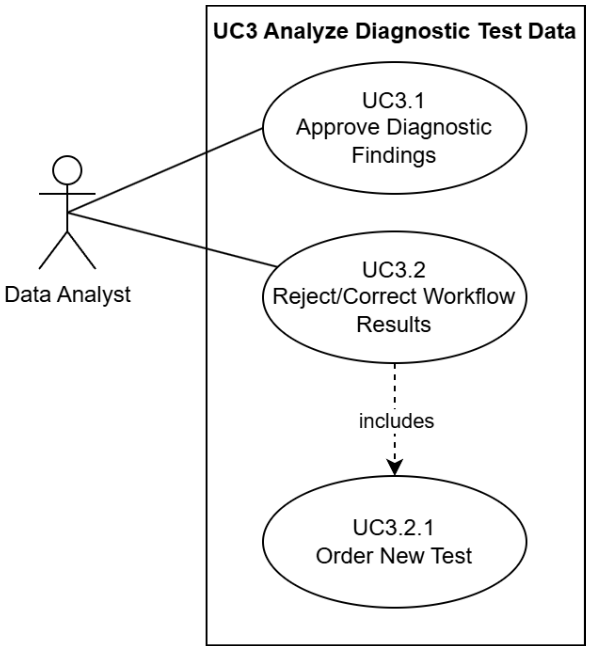


Type: qPCR Analysis
Started: 2024-01-08 10:13
State: Running
Duration: 3m

3.4 Analyze Diagnostic Test Data (UC3)

This use case focuses on the validation and analysis of diagnostic data generated from patient samples, ensuring that results meet quality, accuracy, and reproducibility standards. The process integrates AI-powered inference models to assist clinicians and analysts in interpreting microbiome and genomic findings, reducing manual workload while improving diagnostic precision.

3.4.1 Sub Use-Cases



UC3: Analyze Diagnostic Test Data

A Data Analyst examines the results produced by the diagnostic workflow—such as microbiome analyses, gene expression profiles, or other relevant data—to decide whether the findings are accurate, complete, and ready for further reporting. This overarching use case is subdivided into two principal activities:

UC3.1 Approve Diagnostic Findings

If the analysis appears valid and consistent with the required quality standards, the Data Analyst confirms and approves the diagnostic output. This approval indicates that the results are trustworthy enough to be included in a final report or shared with other stakeholders (e.g., clinical staff).

UC3.2 Reject/Correct Workflow Results

If the Data Analyst detects issues—such as inconsistent sequencing data, insufficient coverage, or results that contradict expected patterns—they reject or correct the workflow output. This process might involve reanalyzing partial data, updating parameters, or adjusting bioinformatics pipelines. In some cases, the Data Analyst may decide that completely new data is needed.

UC3.2.1 Order New Test

When the existing data cannot be salvaged or sufficiently improved, the Data Analyst proceeds to order a new test (for instance, by requesting a fresh stool sample or re-running the sequencing experiment under revised conditions). This sub use case is included under UC3.2,

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

since it represents one of the possible corrective actions taken when the analysis outcome is deemed unsuitable.

3.4.2 Role of AI in Diagnostic Data Analysis

The diagnostic workflow leverages pre-trained AI models that have been developed and validated as part of UC8 (Discover Biomarkers). These models do not require further training at this stage but are instead deployed to infer diagnostic insights.

- Inference Engine for Diagnostic Validation:
  The system applies trained Genome Foundation Models (GFMs) to patient microbiome data, providing confidence scores and anomaly detection. AI-generated outputs support analysts in assessing whether results are consistent, clinically relevant, and within expected thresholds.
- Decision Support & Explainability:
  To enhance trust in AI-driven decisions, explainability techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) are integrated. These tools highlight which microbial features contribute most to a classification, allowing human experts to verify the validity of AI-generated insights.

3.4.3 Workflow Integration

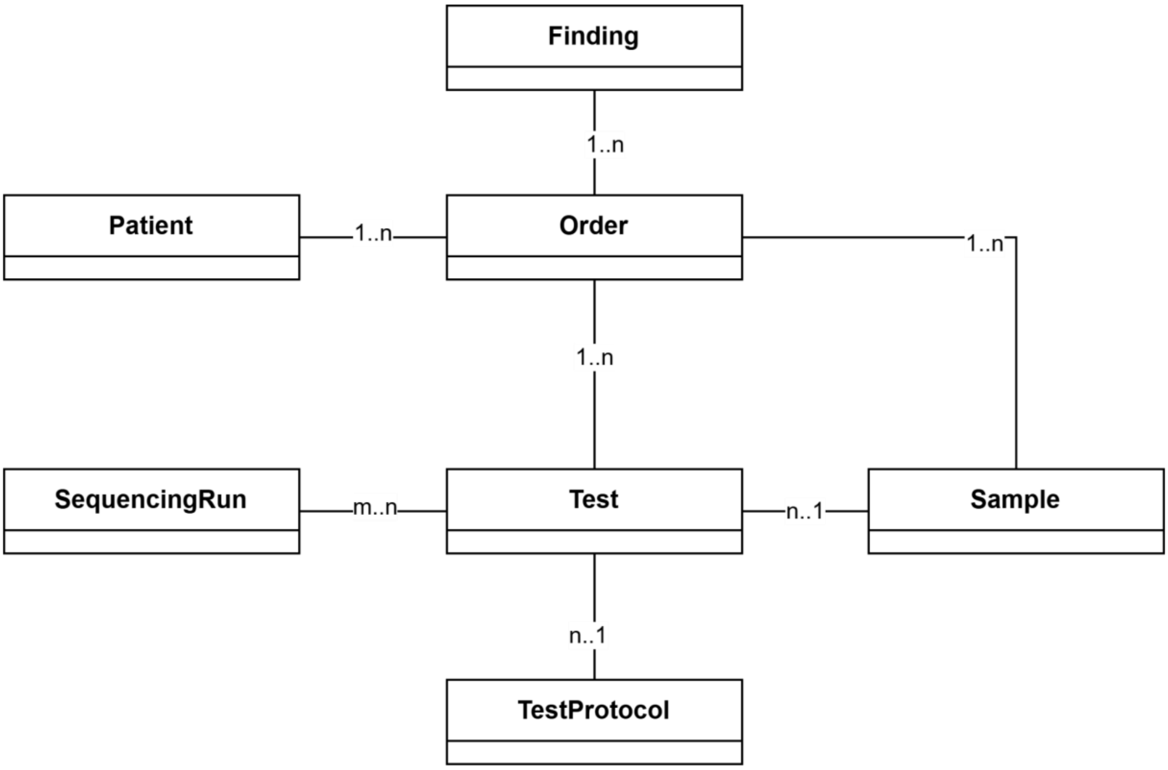
1. **Pre-processing of diagnostic results**: Raw sequencing data undergoes quality control and standardisation before analysis.
2. **AI-driven validation**: The trained models assess diagnostic patterns, identifying potential biomarkers, outliers, or inconsistencies.
3. **Human-in-the-loop verification**: AI confidence scores and explainable outputs are reviewed by clinical analysts for final validation.
4. **Final diagnostic reporting**: Verified results are formatted into clinical reports, integrating **GenDAI's interactive visualisation tools** for improved interpretability.

3.4.4 Augmented Information Model

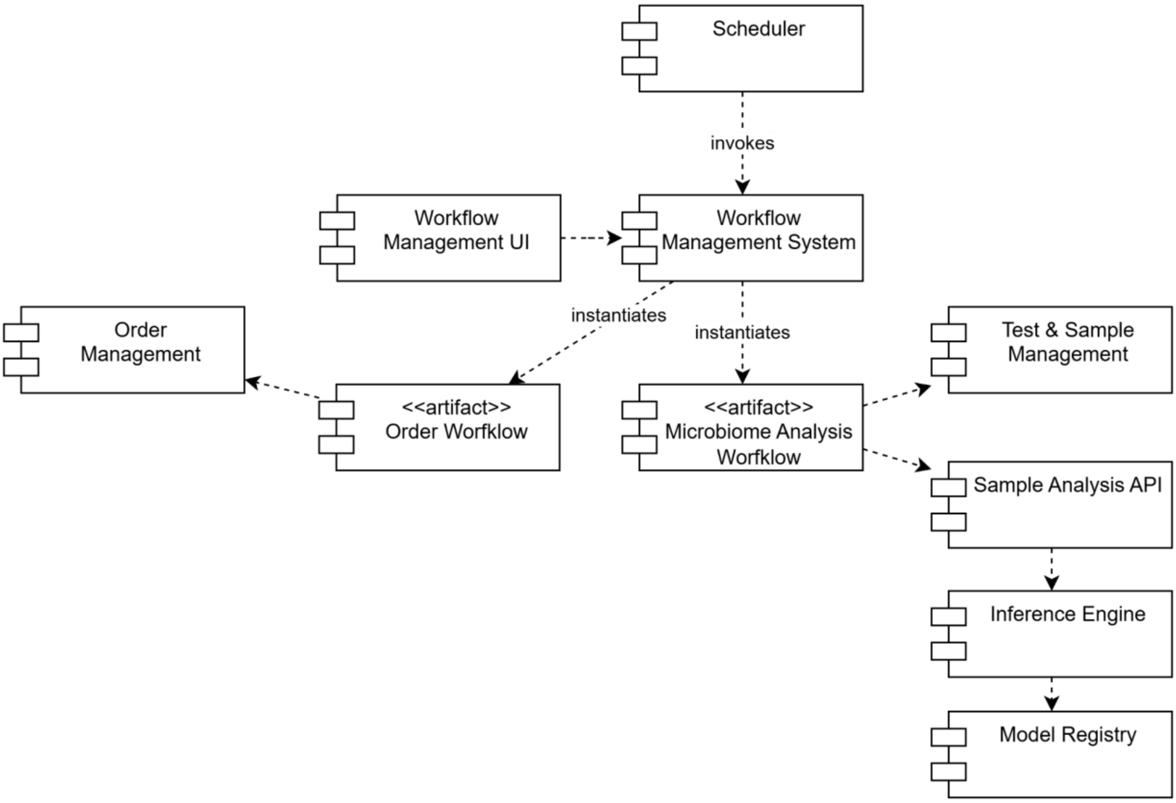
Based on the analyzed use case, we can extend our previous partial information model with the addition of a **Finding** entity that represents any finding produced as part of the analysis. As a finding might depend on the combined results of multiple tests, the entity is connected directly to an order instead of an individual test.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union



3.4.5 Augmented Component Model



This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Based on the component model so far, the following additional components have been identified for the described use cases.

- **Sample Analysis API:** An abstraction that encapsulates the logic to analyze microbiome sample data to derive findings using AI
- **Inference Engine**: A component that applies the trained models to new data for prediction and analysis.
- **Model Registry**: A repository that stores trained models and their metadata, allowing for version control and easy deployment.

3.4.6 Preliminary Partial Architecture Model

Based on the analyzed use cases so far, a preliminary, partial architecture model can be constructed.

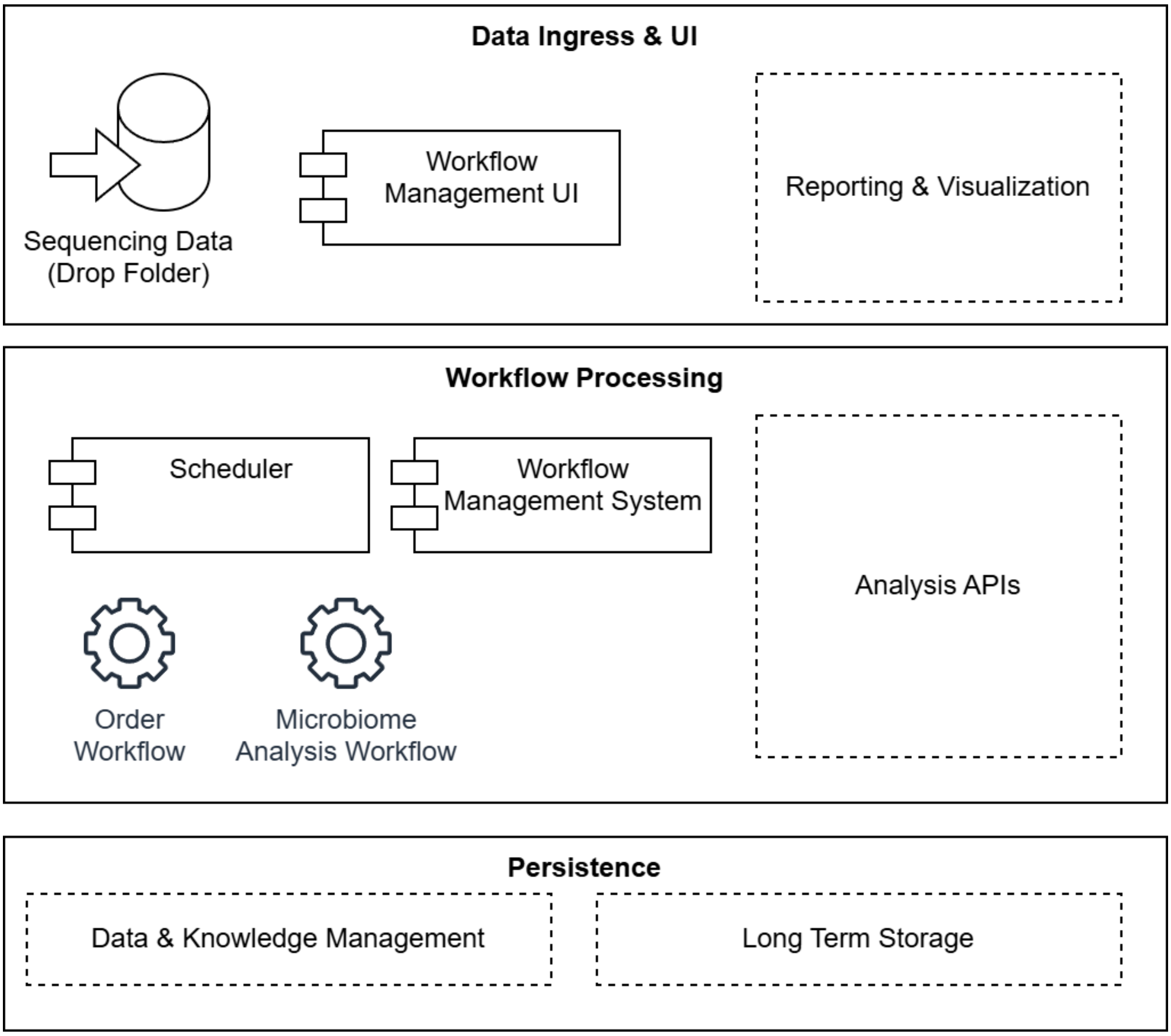
The architecture separates the system into three main layers, each handling distinct responsibilities while interacting with the other layers through clearly defined interfaces. At the top, the **Data Ingress & UI** layer provides entry points for new sequencing data and user interactions. A drop folder acts as a landing zone for raw files produced by the sequencing instrument, and the Workflow Management UI allows lab personnel and analysts to configure and monitor workflows. Reporting and visualization functions are also located in this layer, although the details of these components appear in another section of this document.

Beneath the UI, the **Workflow Processing** layer orchestrates the actual diagnostic and analysis procedures. A Scheduler drives the start and timing of processes by calling into the Workflow Management System, which instantiates and supervises the Order Workflow and the Microbiome Analysis Workflow. These workflows communicate with the Analysis APIs in order to process data, run QC checks, and generate meaningful results. The Workflow Management System coordinates the flow of information between the top-level user interactions and the underlying analysis logic, ensuring that each task is triggered, tracked, and completed in a transparent, auditable manner.

At the foundation of the architecture, the **Persistence** layer ensures long-term data retention and knowledge sharing. The system relies on a dedicated store for large raw and processed data (labeled “Long Term Storage”) and a separate Data & Knowledge Management component for curated information, references, and contextual metadata. Both of these dashed components are explained further in other sections of this document. They are integral to maintaining traceability, enabling reanalysis, and preserving findings for compliance, research, or future diagnostics.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union


Data Ingress & UI
Sequencing Data (Drop Folder)
Workflow Management UI
Reporting & Visualization
Workflow Processing
Scheduler
Workflow Management System
Analysis APIs
Order Workflow
Microbiome Analysis Workflow
Persistence
Data & Knowledge Management
Long Term Storage

3.4.7 Visualization & UI Models

The data analyst decides about the findings of the diagnostic workflow. To make a decision, the data analyst needs to see the relevant data in the easiest and fastest form, as high-quality as possible. The findings must be accurate and complete; thus, the visualization must show all data (completeness) and be precise. As reported in Section 2.6 Visualization and User Interaction, the best visual variable for precision is the position, on a 1D or 2D chart. Textual labels help reveal the content but must be used sparingly to avoid visual cluttering and confusion.

The approval of diagnostic findings is subject to the required quality standards, which means that the visualization will also indicate the standard reference values and how the results relate to such standards.

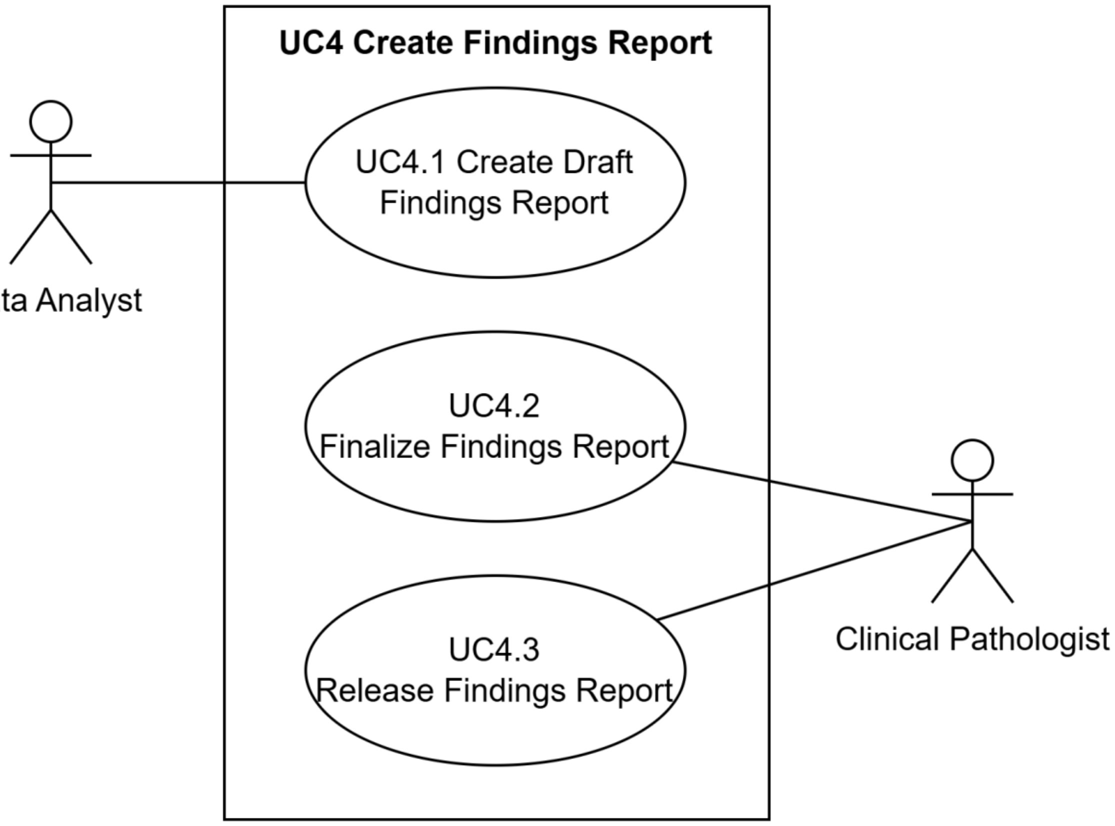
This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

Data coverage must be reported, as expected patterns, if they are defined beforehand, to perform a match with actual results. The visualization should also show the differences between expected patterns and the actual values.

3.5 Create Findings Report (UC4)

3.5.1 Sub Use Cases


UC4 Create Findings Report
UC4.1 Create Draft Findings Report
Data Analyst
UC4.2 Finalize Findings Report
UC4.3 Release Findings Report
Clinical Pathologist

UC4.1 Create Draft Findings Report
The data analyst creates an initial draft findings report, summarizing the analysis results and providing lifestyle recommendations.

UC4.2 Finalize Findings Report
The clinical pathologist, who is ultimately responsible for the findings report correctness, integrates the report with further information if necessary.

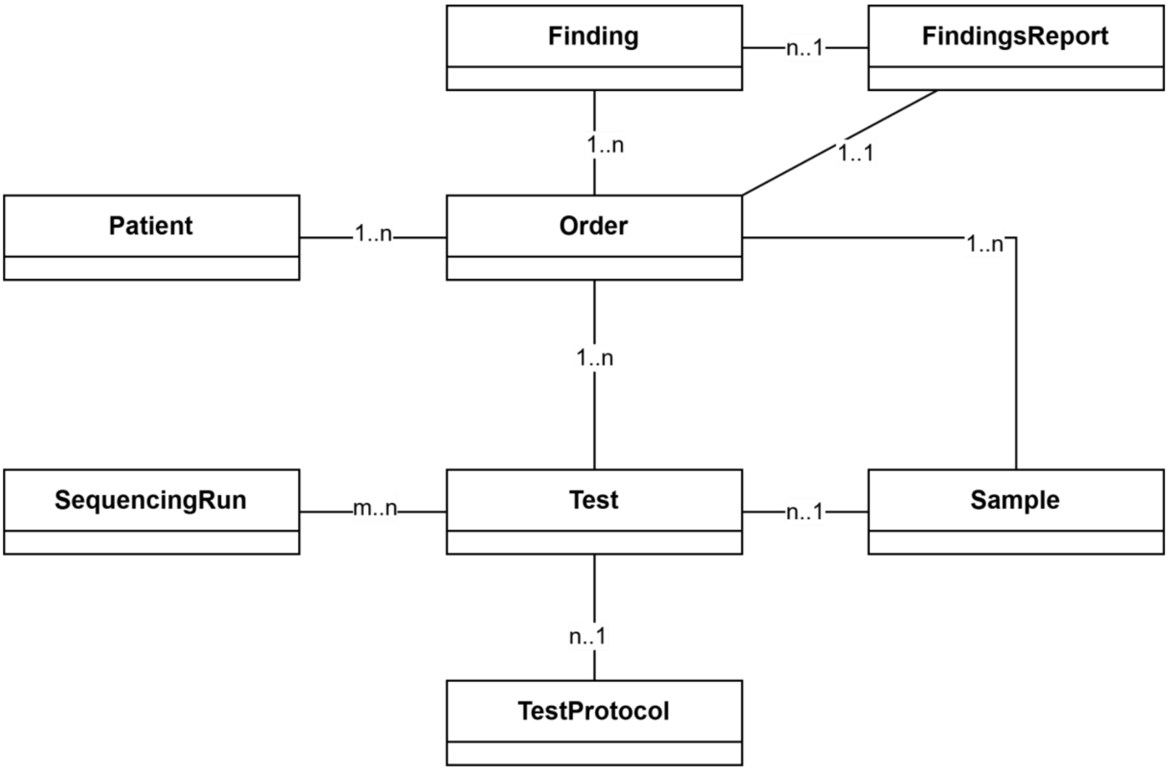
UC4.3 Release Findings Report
The final findings report is approved by the clinical pathologist before it is delivered to the end user.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.5.2 Augmented Information Model

The **findings report** can be added to our existing information model by attaching it to the order. It is expected that each order has exactly one findings report, which summarizes one or more findings that are already associated with the order.



3.5.3 Visualization & UI Models

The report format is determined by the company workflow and the tools in use. The traditional reports should be complemented by visualizations to reduce time to process information and improve understandability. The data analyst summarizes the key results. For data that can be expressed in numbers or mapped to visual structures, the report can integrate visualizations that will be integrated with the text containing explanation and recommended actions. The Clinical Pathologist is supported by the visual findings in the finalization of the report and its delivery.

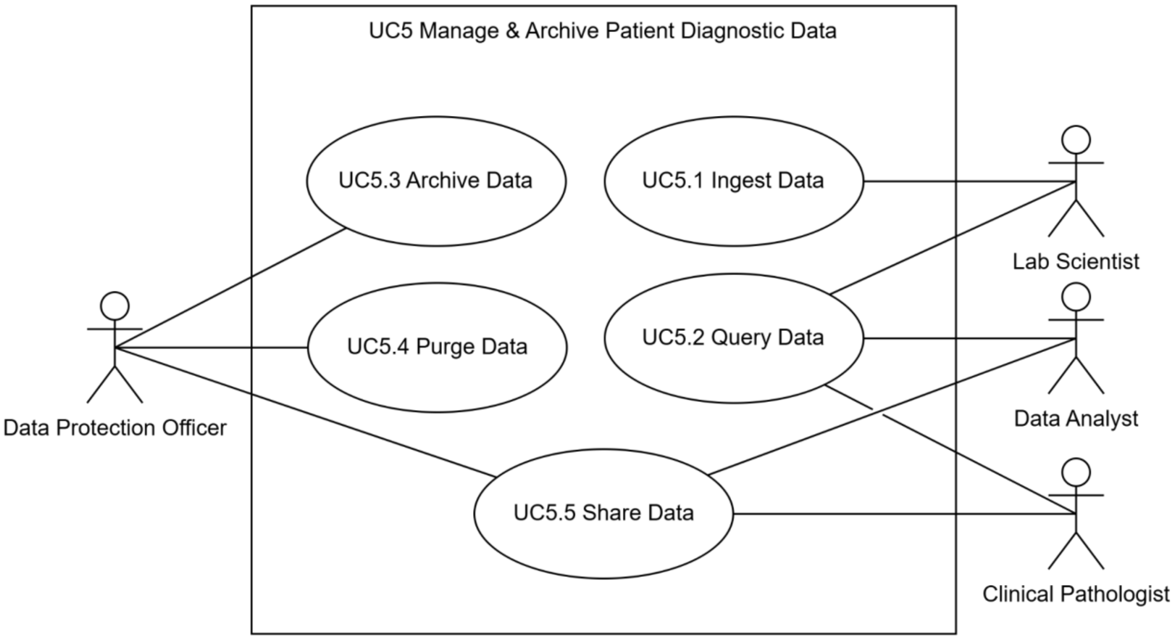
Interactive data visualization is useful for those situations where data exceeds the screen or the size of a page. As it happens for other diagnostic analysis, such as ultrasound, MRI and CT scans, where the operator chose which image most represents the situation, also in interactive visualizations, the Clinical Pathologist, but also the Data Analyst, must have the possibility to take screenshots and chose the best images that represent the data and the message they want to pass to the other steps of the process. This represents an additional review tool that can improve accuracy and clarity of the final output.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.6 Manage & Archive Patient Diagnostic Data (UC5)

3.6.1 Sub Use Cases



UC5.1 Ingest Data

Once patient samples are registered, associated biomedical data (from sequencing machines, imaging devices, and research databases) is securely ingested and stored. Each dataset is assigned a Persistent Identifier (PID) to ensure traceability, prevent duplication, and maintain data integrity. Storage follows Open Archival Information System (OAIS) standards (ISO 14721) for long-term accessibility and reproducibility. Encryption mechanisms and controlled access protect sensitive data from unauthorized modification or breaches.

UC5.2: Query Data

Authorized users, such as lab scientists, data analysts, and clinicians, can securely search, query, and retrieve patient datasets. The system normalizes and indexes stored data for efficient retrieval while ensuring all access requests are logged for audit purposes. Users may annotate and update datasets while adhering to strict access controls that prevent unauthorized modifications or data leaks.

UC5.3 Archive Data

Biomedical data is archived in accordance with OAIS (ISO 14721) standards and GDPR obligations to ensure long-term preservation, identification, accessibility, and integrity. The system enforces structured data retention policies to define how long diagnostic data remains available before archiving or deletion, ensuring alignment with institutional and regulatory

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

requirements. To maintain security and compliance, the archiving process integrates privacy controls, access restrictions, compliance audits, and governance policies that safeguard sensitive patient information while allowing authorized retrieval when necessary.

Strict privacy controls define how patient data is stored, accessed, and shared. De-identification techniques protect sensitive information, dynamically adjusting access permissions based on regulatory updates and evolving security policies.

The system supports periodic compliance audits by tracking all data handling activities. Regulatory assessments verify adherence to GDPR, HIPAA, IVDR 2017/746, and institutional privacy policies. Any discrepancies are flagged for corrective actions.

Data governance policies define structured access control mechanisms, security logging, and retention strategies. Data protection officers monitor these policies to align with legal and ethical standards, ensuring secure data management.

UC5.4 Purge Data

Strict policies also define if and when data is allowed or required to be purged. Some regulations might require data to be stored securely for a specific amount of time, while other regulations might require data to be deleted, i.e., based on user request. This UC will take care of any individual request to erase any of their personal data, thus exercising the “right to be forgotten” or “right to erasure”, determined by art. 17 GDPR.

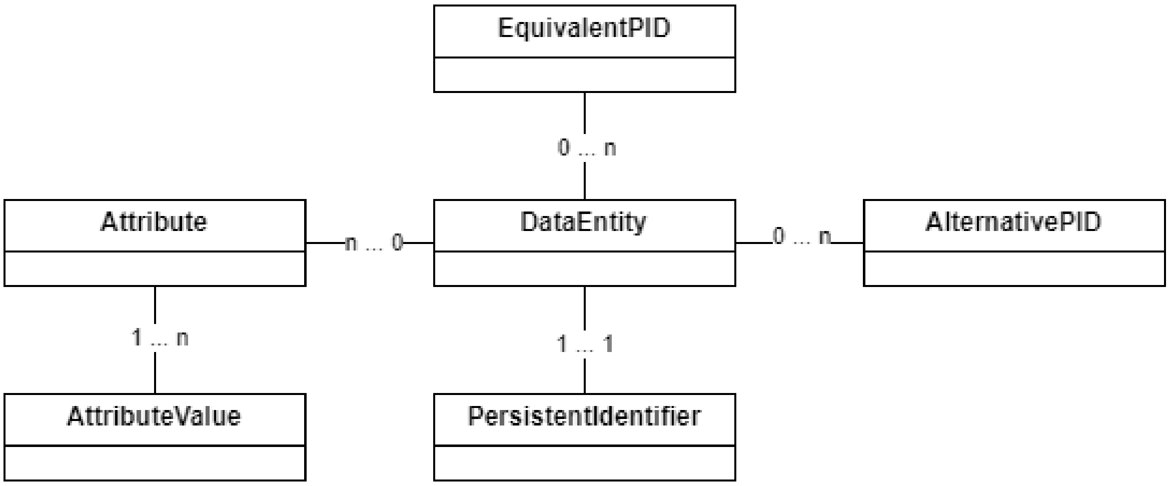
UC5.5 Share Data

Data-sharing processes enforce authentication and authorization policies, enabling controlled exchange of information between laboratories, hospitals, and research institutions. Standardized data formats ensure interoperability with Laboratory Information Management Systems (LIMS) and external databases, while legal agreements dictate the terms of access and usage.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.6.2 Partial Information Model



The Information Model establishes the fundamental entities and relationships in the Data Management and Security Architecture system, ensuring secure, traceable, and compliant biomedical data processing. It provides a structured approach to data ingestion, storage, processing, and sharing, supporting regulatory adherence, security enforcement, and interoperability across research and clinical environments.

Entities in the system are derived from different processes: patient specimens, laboratory tests, or biomedical imaging. These entities can be semantically different from each other. In fact, an entity semantically categorized as “Patient” would have different attributes (date of birth, gender, etc.) from an entity of type “Sample” (patientId, date of sampling, etc.). What is critical is that each entity is ingested into the system through a controlled process that verifies data integrity and assigns a **Persistent Identifier** to ensure uniqueness, traceability and prevent duplication. This identifier enables seamless data linkage across multiple workflows, enhances interoperability with external laboratory information systems and will be managed by a system that will ensure its complete lifecycle (creation, deletion, duplication, merge etc.).

An **Attribute** is a generic structure that can store metadata or other information about an **Entity.** It can be keyed or key-less, and multiple **AttributeValue**(s) can be associated with. This pattern permits an improved research approach, which is not guided only by the **PersistentIdentifier**. When entities are merged, their attributes are joined, with the aim of preserving information through the entire lifecycle.

Other relevant information for the **Entity** lifecycle is stored as **AlternativePID**(s) and **EquivalentPID**(s). An **AlternativePID** is a PersistentIdentitifier that represents the entity in another system that has its own convention to uniquely identify an entity. Differently, an **EquivalentPID** is a PersistentIdentifier belonging to another entity, which is added during the process of merging and splitting. In the first case, the PersistentIdentifier of the merging entity is added to the EquivalentPID(s) of the destination entity. When splitting an entity, resultant entities will contain the PersistentIdentifier of the original entity in their EquivalentPID(s).

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

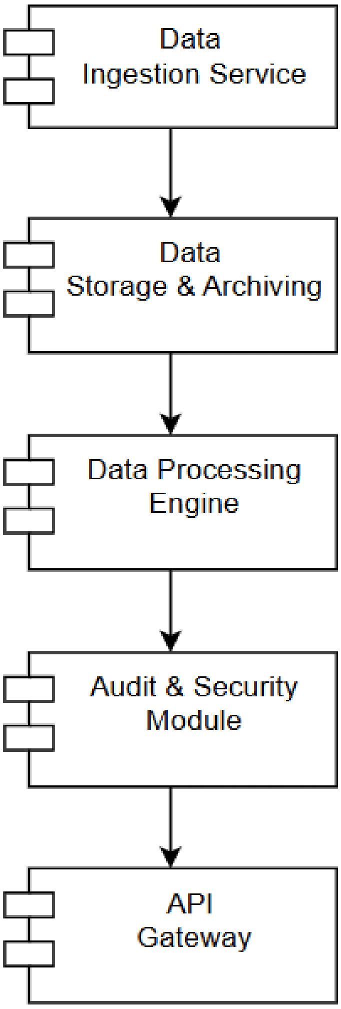
Entities in the system can be searched and retrieved using AlternativePID(s) and EquivalentPID(s) to guarantee the persistent availability of the information present in the system.

To maintain security and compliance, all interactions with data are recorded in **Audit Logs**, creating an immutable record of modifications, access requests, and processing activities. These logs support real-time security monitoring, regulatory audits, and data validation, ensuring compliance with *GDPR, HIPAA,* and *ISO/IEC 27001*. The system actively detects unauthorized access attempts and applies automated security measures to enforce access control policies and mitigate potential threats.

At the highest level, the **Data Entity** serves as the core structure that integrates samples, attributes, persistent identifiers, and security logs into a unified framework. This data governance model ensures a well-organized, secure, and scalable data management system that supports controlled data retrieval, automated processing pipelines, and compliant data sharing through authenticated APIs.

By implementing this structured and policy-driven approach, the system ensures that biomedical data remains secure, traceable, and interoperable, supporting high-quality research, diagnostics, and regulatory oversight. This model facilitates seamless data exchange, enhances reproducibility, and provides a scalable foundation for evolving clinical and research environments.

3.6.3 Partial Component Model



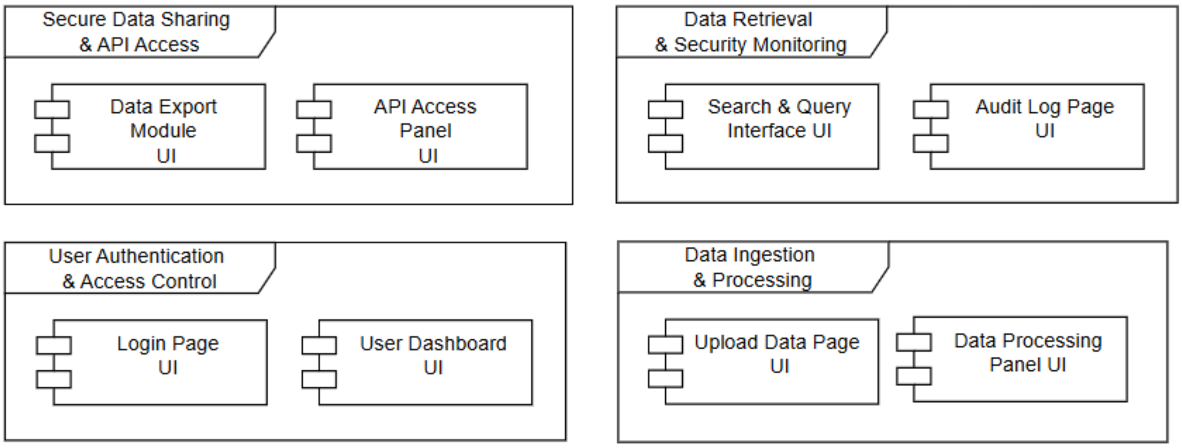
The Component Model for Data Management and Security Architecture defines the core building blocks of the GenDAI Data Management System, ensuring secure data ingestion, storage, processing, access control, and compliance enforcement. It provides a structured, scalable, and interoperable framework that supports biomedical research, clinical diagnostics, and regulatory adherence across diverse computing environments.

At the foundation of the architecture, the **Data Ingestion Service** collects and validates biomedical data from laboratory instruments, imaging devices, and research databases. Each incoming dataset is assigned a Persistent Identifier to ensure traceability, prevent duplication, and enable interoperability across different workflows. Verified data is stored in the **Data Storage & Archiving** component, which supports both structured and unstructured formats while ensuring long-term accessibility and compliance with standards such as OAIS (ISO 14721). This guarantees that stored data remains reproducible, transparent, and available for research and regulatory audits.

To facilitate efficient retrieval and processing, the **Data Processing Engine** extracts metadata, normalizes datasets, and applies automated classification and indexing. These operations enhance the organization, accessibility, and usability of biomedical data, enabling efficient search and analysis.

By implementing this modular, scalable, and policy-driven framework, the Component Model ensures that biomedical data remains secure, traceable, and interoperable. The seamless integration of persistent identifiers, audit logging, encryption, and access control mechanisms establishes a future-proof architecture that supports high-quality research, diagnostics, and regulatory compliance in evolving biomedical and clinical environments.

3.6.4 UI Models


Secure Data Sharing & API Access
Data Export Module UI
API Access Panel UI
Data Retrieval & Security Monitoring
Search & Query Interface UI
Audit Log Page UI
User Authentication & Access Control
Login Page UI
User Dashboard UI
Data Ingestion & Processing
Upload Data Page UI
Data Processing Panel UI

The **GenDAI Data Management System UI Model** is structured into four key components, each representing a critical aspect of data security, management, and interoperability. These components ensure that users can securely authenticate, ingest data, retrieve information, monitor security, and share data with external systems while adhering to compliance and regulatory requirements.

The **User Authentication & Access Control** component is responsible for managing user identities and ensuring secure access to the system. It includes the **Login Page UI**, where users authenticate using secure credentials, often with multi-factor authentication (MFA) for additional security. Once authenticated, users interact with the **User Dashboard UI**, which provides role-based access to system functionalities. Depending on their permissions, users can view, manage, or modify datasets, ensuring that only authorized personnel can interact with sensitive biomedical data.

The **Data Ingestion & Processing** component enables users to upload, validate, and process biomedical datasets securely. The **Upload Data Page UI** facilitates data submission while ensuring integrity through validation checks and the assignment of persistent identifiers. This process prevents duplication and maintains traceability. Once ingested, data is processed through the **Data Processing Panel UI**, where metadata normalization, indexing, and classification are applied. This structured approach ensures that datasets are ready for retrieval, analysis, and compliance with data governance standards.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

The **Data Retrieval & Security Monitoring** component provides users with the ability to search, query, and audit data interactions within the system. The **Search & Query Interface UI** allows users to efficiently retrieve indexed datasets while ensuring data integrity and controlled access. Complementing this is the **Audit Log Page UI**, which continuously tracks user interactions, modifications, and access requests. These logs support compliance with regulations such as GDPR and HIPAA, ensuring that all actions within the system are transparent and auditable for security purposes.

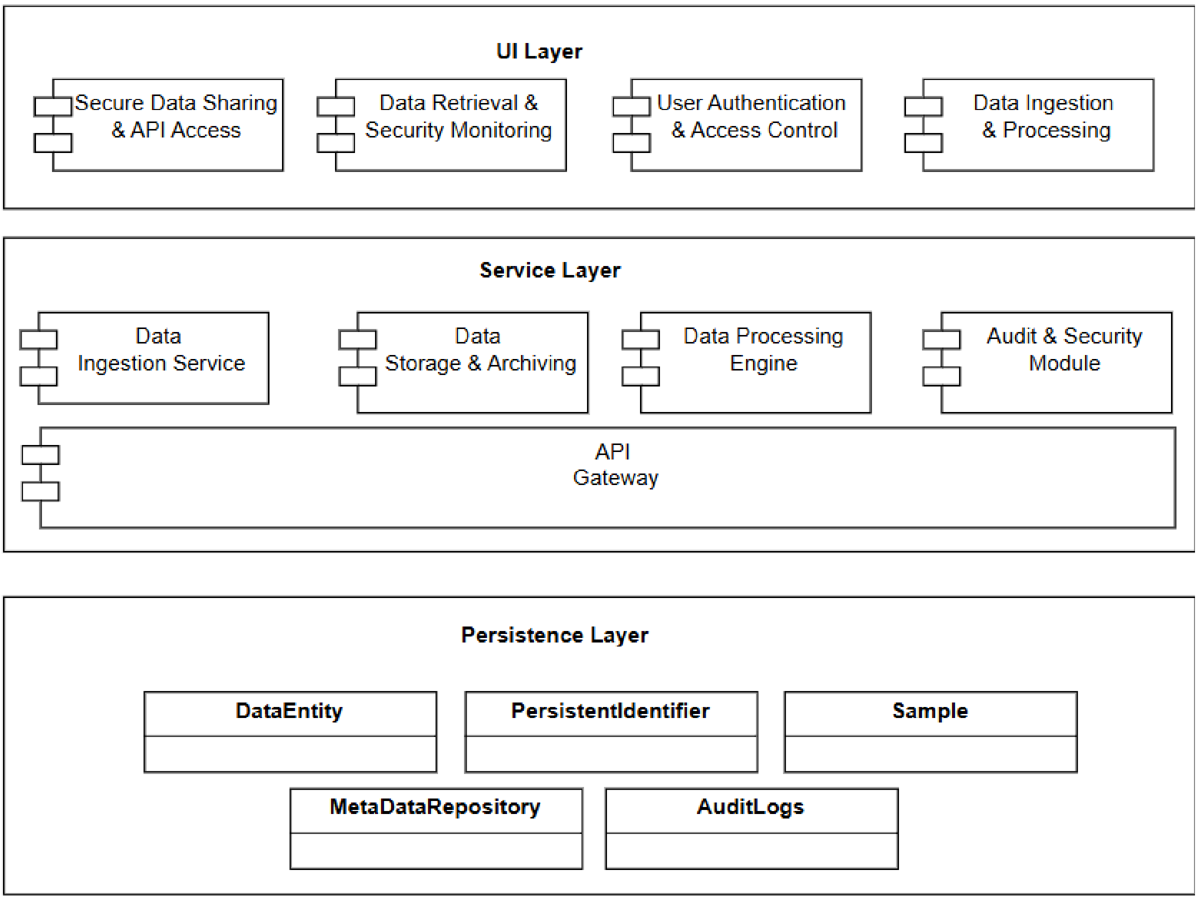
The **Secure Data Sharing & API Access** component facilitates interoperability between the GenDAI Data Management System and external platforms, ensuring that data is exchanged securely and in compliance with predefined policies. The **Data Export Module UI** allows users to generate secure data packages for external sharing, enforcing encryption and anonymization where necessary. The **API Access Panel UI** manages external data requests, ensuring authentication and authorization before serving data. This approach ensures that research institutions, clinical laboratories, and regulatory bodies can securely interact with the system while maintaining strict data governance.

Together, these four components create a structured and secure UI framework for the GenDAI Data Management System. The architecture ensures seamless data management, regulatory compliance, and controlled access while providing users with an intuitive interface to perform their tasks efficiently. This design supports high-quality biomedical research and diagnostics by integrating authentication, data processing, security monitoring, and controlled sharing into a unified and scalable system.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

3.6.5 Architecture Models



The GenDAI Data Management System follows a structured *Model-View-Controller (MVC)* architecture, integrating the UI Model, Component Model, and Information Model to ensure secure, scalable, and compliant biomedical data management.

The **UI Layer (View)** provides a structured interface for user interactions, including *User Authentication & Access Control, Data Ingestion & Processing, Data Retrieval & Security Monitoring,* and *Secure Data Sharing & API Access*. It ensures secure access, intuitive data submission, retrieval, auditing, and controlled external data exchange. Role-based access control (RBAC) and Multi-Factor Authentication (MFA) safeguard user interactions, while the audit log UI enables compliance tracking.

The **Service Layer (Controller)** acts as the system's processing core, enforcing security policies, managing data workflows, and executing business logic. It consists of the Data Ingestion Service, Data Storage & Archiving, Data Processing Engine, Audit & Security Module, and API Gateway. These services validate, store, process, and regulate data access while applying encryption and anonymization techniques to ensure GDPR, HIPAA, and ISO/IEC 27001 compliance.

The **Persistence Layer (Model)** maintains structured biomedical data, integrating DataEntity, PersistentIdentifier, Sample, MetadataRepository, and AuditLogs. It ensures data integrity, traceability, and long-term accessibility. Samples are assigned unique persistent identifiers,

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

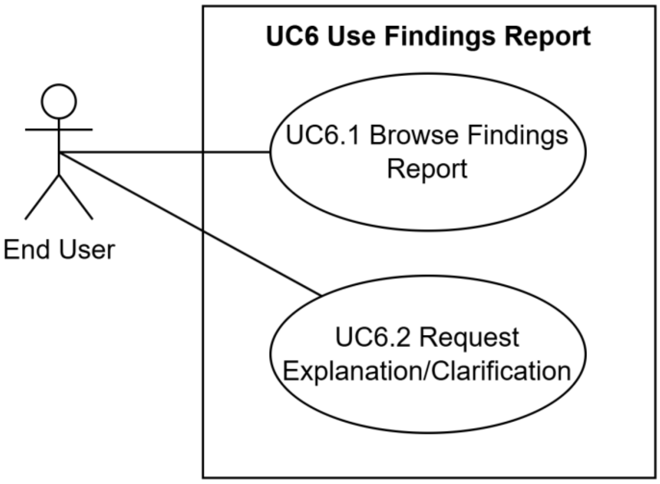
Funded by the European Union

while metadata repositories capture essential contextual details. Audit logs maintain a complete record of data interactions, supporting security and regulatory requirements.

By combining these three layers, the MVC architecture ensures modularity, security, and efficiency in managing biomedical data. The UI Layer facilitates user interactions, the Service Layer enforces policies and business logic, and the Persistence Layer provides structured data storage and integrity. This integrated approach enables seamless data management, advanced research capabilities, and controlled interoperability with external platforms.

3.7 Use Findings Report (UC6)

3.7.1 Sub Use Cases


UC6 Use Findings Report
UC6.1 Browse Findings Report
UC6.2 Request Explanation/Clarification
End User

UC6.1 Browse Findings Report

Upon receiving a findings report, the end user can start browsing its content. The findings report can be in printed form, a static electronic report, or an interactive report.

The first two are equivalent, the interactive report requires a computer and provides the user several features to make the reading of the report. easier Specifically, using filters, the user can focus only on relevant items of the report; using sort features, the user can rank the items according to some importance criteria, in order to have at hand the most important information; using different visualization techniques, the user can see different perspectives of the data. The visualization must be complemented with explanations or descriptions of the items, also containing the indication of ranges and, if possible and available, an indication of the average values of similar people. An example of this last concept is in the leaflet that is provided in the drug box, which explains the frequencies of side effects, typically providing the number of side effects observed out of the size of the sample taken into account (e.g. this side effect has been observed on 10 out of 1000 people).

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

UC6.2 Request Explanation/Clarification

While browsing the report, the end user might require explanations or clarifications on one or more topics in the report.

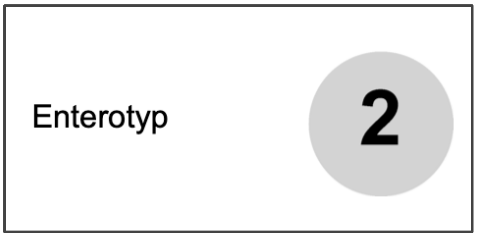
The reports are typically provided as a printed form or the digital version of a static report. The addition of interactivity cannot change this, because of legal, formal, and procedural reasons. Nevertheless, interactivity, which is feasible only on electronic reports, can help the readability of reports and augment them with useful information.

When an end user reads a medical report, there are two main reaction to the results: the first is to ask a doctor for explanation, who has to tell the patient the meaning of the data; the second is to search herself for possible explanations, often to have an idea about the situation and be prepared to ask more useful questions when she meets the doctor.

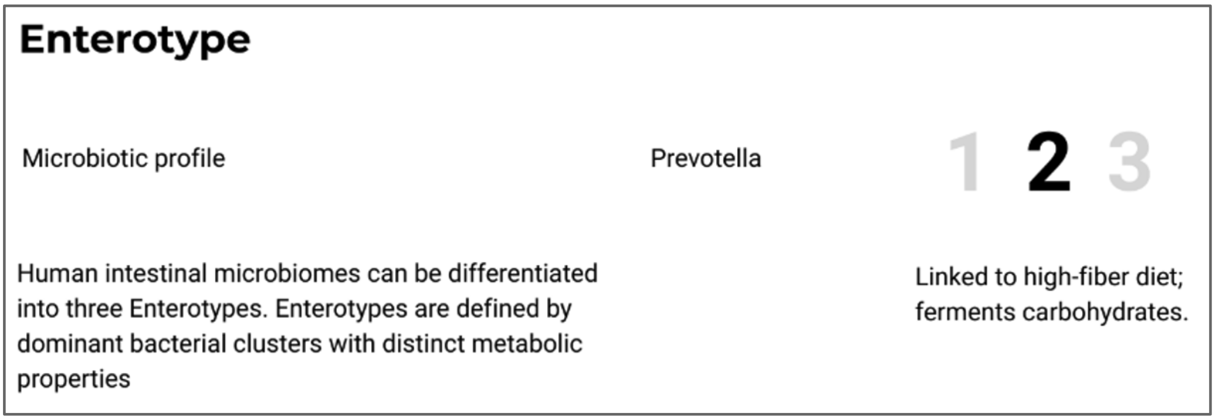
Interaction can be performed by providing additional information on demand, for instance, after a click on specific items shown on the report, or by filtering the report according to specific criteria, in order to reduce the visible items and be able to compare only relevant ones.

?
MEDICAL REPORT
INFO
MEDICAL REPORT

The report can also be advanced and simple at the same time. For instance, the detection of enterotypes, which are the dominant bacterial clusters, can be very evident, such as the following figure.


Enterotyp 2

However, a person might not know how many enterotypes are present, so a better user interface could be one that reveals the enterotype among the existing ones, like the following figure, which also contains the description of the profile, in addition to the enterotype number (see next figure).



Enterotype

Microbiotic profile

Prevotella

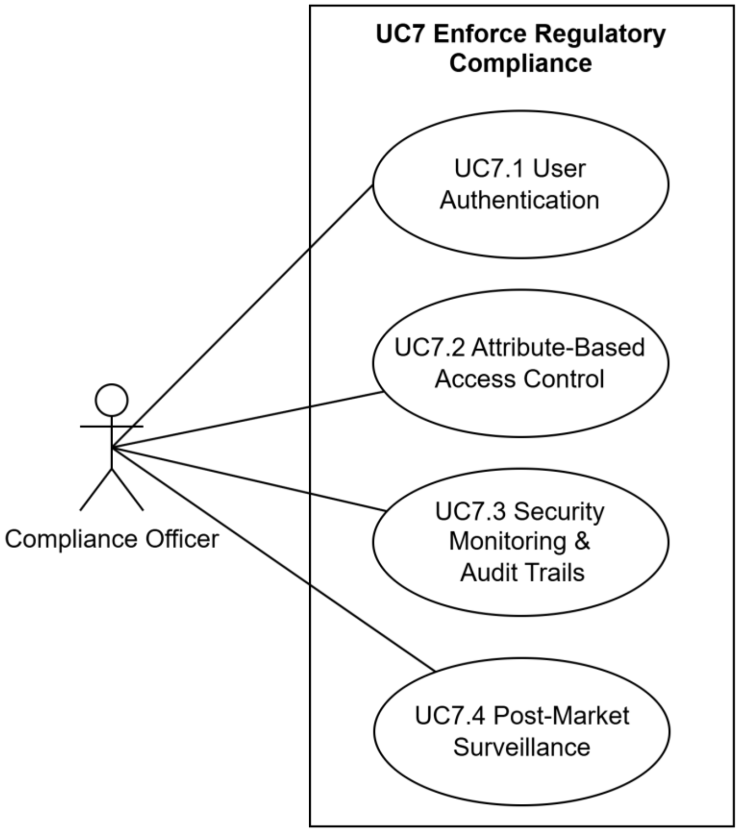
1 **2** 3

Human intestinal microbiomes can be differentiated into three Enterotypes. Enterotypes are defined by dominant bacterial clusters with distinct metabolic properties

Linked to high-fiber diet; ferments carbohydrates.

3.8 Enforce Regulatory Compliance (UC7)

3.8.1 Sub Use Cases


UC7 Enforce Regulatory Compliance
UC7.1 User Authentication
UC7.2 Attribute-Based Access Control
UC7.3 Security Monitoring & Audit Trails
UC7.4 Post-Market Surveillance
Compliance Officer

UC7.1: User Authentication

Before accessing sensitive data or operations, users must authenticate using Multi-Factor Authentication (MFA).

UC7.2 Attribute-Based Access Control

The system enforces Attribute-Based Access Control (ABAC) policies, granting permissions based on user attributes. Unauthorized access attempts trigger security alerts, ensuring continuous monitoring and compliance. Security policies dynamically regulate data access, ensuring that only authorized personnel can retrieve, modify, or process sensitive diagnostic information. ABAC frameworks prevent unauthorized changes, and any compliance violations trigger alerts for immediate review and corrective actions.

UC7.3 Security Monitoring & Audit Trails

All interactions with biomedical data, including ingestion, retrieval, processing, and modifications, are monitored and logged for security and compliance. Every interaction with a patient and diagnostic data is recorded in an audit trail to maintain accountability. These logs allow compliance officers to verify adherence to data governance policies and detect anomalies or unauthorized access attempts.

The system continuously analyzes security events, flagging potential breaches such as unauthorized access, policy violations, or suspicious user behavior. Automated responses, such as access revocation or additional authentication challenges, mitigate risks in real time, ensuring the integrity of diagnostic data.

3.8.2 Augmented Component Model

The **Audit & Security Module** continuously logs all interactions, modifications, and access requests, creating a timestamped audit trail that supports security monitoring, forensic investigations, and compliance verification in accordance with GDPR, HIPAA, and ISO/IEC 27001.

Data access is governed by the Policy Enforcement & Compliance Engine, which dynamically applies Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) to ensure fine-grained permission management. Security policies regulate data retrieval and processing activities, preventing unauthorized modifications or breaches. The Encryption & Privacy Layer ensures that biomedical data remains protected both at rest and in transit, applying encryption, anonymization, and pseudonymization techniques to safeguard sensitive information.

User authentication and authorization are handled by the Identity & Access Management (IAM) system, which integrates Multi-Factor Authentication (MFA) and Federated Identity Management to ensure secure user verification. The **API Gateway** provides controlled access to external applications and research platforms, enforcing strict security protocols to enable secure and compliant data exchange between GenDAI and external systems, including LIMS and clinical databases.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.9 Discover Biomarkers (UC8)

This section outlines the AI models employed for biomarker identification and diagnosis, focusing on their conceptual design, training processes, and feature selection. Furthermore, it addresses crucial aspects of model explainability, robustness, and bias mitigation, ensuring the reliability, transparency, and fairness of the AI-driven diagnostic system. The research draws significant inspiration from the latest advancements in Genome Foundation Models (GFMs), which offer a transformative approach to bioinformatics by leveraging large-scale AI models trained on extensive genomic data, as detailed in Section 2.5. These models are essential for overcoming challenges related to high-dimensional data, generalisability, and the discovery of novel microbial insights.

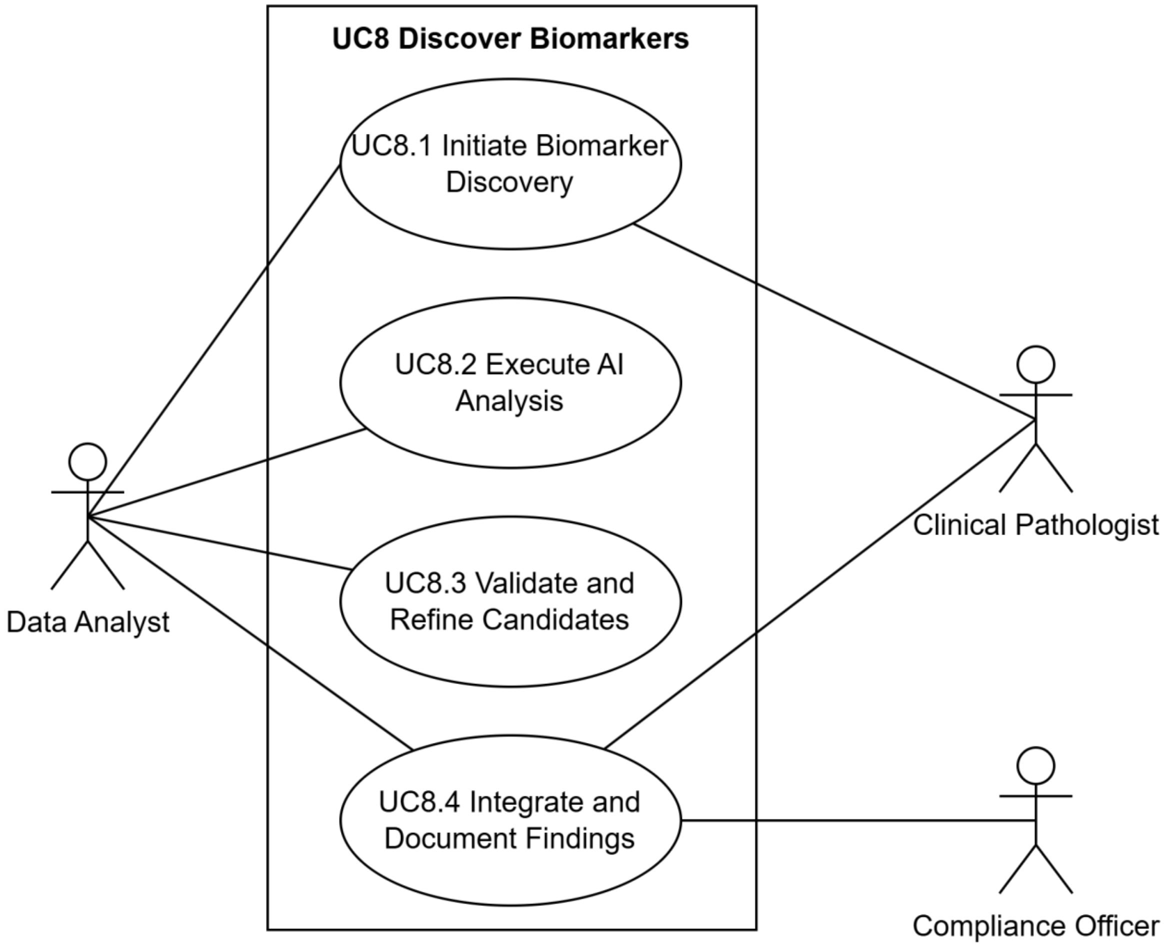
This use case focuses on identifying and validating novel biomarkers using advanced AI models trained on metagenomic data. It encompasses both the development of biomarker discovery models (training) and their application for clinical inference, directly addressing the user requirements identified in Section 2.5.

To ensure clarity and align with best practices in AI-driven diagnostics, this section is divided into model training (UC8.1) and model inference (UC8.2).

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

3.9.1 Sub Use Cases


UC8 Discover Biomarkers
UC8.1 Initiate Biomarker Discovery
UC8.2 Execute AI Analysis
UC8.3 Validate and Refine Candidates
UC8.4 Integrate and Document Findings
Data Analyst
Clinical Pathologist
Compliance Officer

UC8.1 Initiate Biomarker Discovery

The biomarker discovery process is initiated and guided by the clinical pathologist in conjunction with the Data Analyst. AI can assist in the early stages by helping to identify relevant genomic datasets or by supporting the refinement of initial hypotheses based on existing scientific literature, leveraging NLP for literature mining (as described in Section 2.5). The Data Analyst selects and loads relevant genomic datasets, preparing for the automated biomarker discovery workflow.

UC8.2 Execute AI Analysis

Raw genomic data undergoes rigorous pre-processing, including cleaning, tokenisation (e.g., k-mer creation), and normalisation to remove noise, inconsistencies, and handle compositional data peculiarities. Following this, advanced Genome Language Models (e.g., DNABERT-2, Nucleotide Transformer, Evo) are employed to analyse the sequences. These models leverage transformer architectures, similar to those used in natural language processing (Section 2.5.1), but adapted for genomic "languages". They are designed to identify candidate biomarkers by capturing complex patterns and relationships within the vast genomic data. This identification can be further enhanced using sophisticated visualisation techniques, aiding the user in discerning complex data patterns and relationships.

UC8.3 Validate and Refine Candidates

The Data Analyst reviews the AI-generated candidate biomarkers using statistical metrics and advanced visualisation tools. Crucially, this stage heavily incorporates explainability tools (XAI), such as SHAP values, LIME, and attention weights from Transformer models, as highlighted in Section 2.5.2. These tools provide insights into *why* a model made a particular prediction, thereby addressing user requirements for transparency and interpretability. This allows for rigorous validation and, if needed, refinement of the results, ensuring clinical actionability.

UC8.4 Integrate and Document Findings

Validated biomarkers are integrated into the diagnostic workflow, ensuring seamless transition from discovery to clinical application. All steps and outcomes are meticulously documented and archived for compliance with regulatory standards (e.g., GDPR), traceability, and future reference, promoting transparency and reproducibility. This step also accounts for feedback loops from clinicians, allowing for validation or overriding of predictions based on domain expertise, as identified in user interviews (Section 2.5.2).

Taxonomic Classification of DNA Sequences

To advance biomarker identification in the GenDAI project, WP4 focuses on the taxonomic classification of bacterial DNA sequences, particularly through the use of 16S rRNA data. This bacterial DNA sequence data is essential for distinguishing microbial taxa in clinical samples, providing a foundation for identifying health-related biomarkers. The 16S rRNA sequences are highly conserved within bacteria, allowing for reliable taxonomic classification across diverse microbial communities. This data is collected and processed in accordance with strict pre-analytics standards to ensure accuracy and reproducibility in biomarker discovery. The principles of Genome Language Models (Section 2.5.1) are central to this classification.

NLP and Semantic Embeddings for Enhanced Metagenomic Classification

AI, and specifically NLP techniques adapted for genomic data, address the limitations of traditional alignment-based genomic methods by implementing alignment-free classification techniques. Using semantic embeddings, which are vectorised representations of DNA sequences, AI models can classify metagenomic data with greater precision and reduced computational demand. Semantic embeddings are particularly useful for parsing metagenomic sequences, facilitating the identification of unique microbial markers relevant to clinical diagnostics. This innovative approach not only increases the speed of processing vast genomic datasets but also strengthens the robustness of taxonomy classification, thereby meeting the rigorous clinical requirements outlined in WP1 and addressing the challenge of handling high-dimensional data (Section 2.5.2).

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by
the European Union

Towards Personalised Genomic Analysis Through AI-Driven Modelling

A significant feature of AI in GenDAI's clinical framework is the capability for personalised genomic analysis, a key user requirement highlighted in Section 2.5.2. The split-fed learning framework (from T4.2) enables the customisation and training of specific layers of the GenDAI model using local, pre-processed data. This allows for more tailored model outcomes that can be adapted to unique clinical environments and individual patient data, significantly enhancing the potential for discovering novel biomarkers. By refining the classification and predictive capabilities of metagenomic samples, this personalised approach supports the development of bespoke clinical solutions that address the varying needs of patient populations and improves the model's generalisability.

AI Models for Biomarker Identification and Diagnosis

The AI models used in this project are primarily based on **Genome Foundation Models (GFMs)**, a novel class of large-scale AI models trained on extensive genomic datasets. These models are designed to handle tasks such as sequence analysis, gene annotation, and gene expression prediction, offering significant improvements in accuracy, scalability, and efficiency over traditional bioinformatics methods like BLAST.

Conceptual Design of Genome Foundation Models (GFMs)

GFMs are employed for sequence analysis and classification. Their primary purpose is to analyse DNA sequences, classify them into taxonomic groups, and identify potential biomarkers. These models are capable of processing large-scale genomic datasets, such as 16S rRNA sequences, to classify microbial communities and detect anomalies, directly addressing the challenge of high-dimensional data as detailed in Section 2.5.2. Algorithms such as DNABERT-2 and Nucleotide Transformer (Section 2.5.1) are specifically utilised for these sequence classification tasks. These models leverage transformer architectures to process long DNA sequences, enabling them to capture complex patterns and relationships within the data. For feature selection, GFMs automatically learn relevant features from raw genomic data, thereby reducing the need for manual feature engineering. This is particularly useful for tasks like taxonomic classification, where traditional methods often rely on curated reference databases. The improved BiGAMi algorithm (T4.1) will be integrated to further enhance feature selection, accepting continuous data and leveraging co-variables to remove confounding phenotypic variation before feeding data to a Random Forest (RF) model.

Furthermore, GFMs are used for gene annotation and expression prediction. Their purpose is to predict gene functions and expression levels based on genomic data, a crucial step for understanding the role of specific genes in disease pathways and identifying potential therapeutic targets. Algorithms like Evo (Section 2.5.1) and DNABERT-S are employed for these tasks, having been trained on large datasets of annotated genomes, which allows them to predict gene functions and regulatory elements with high accuracy. In terms of feature selection, GFMs use embeddings to represent DNA sequences in a high-dimensional space, where similar sequences are mapped to nearby points, enabling efficient comparison and classification of sequences even when they are not identical.

Lastly, GFMs are applied to microbiome analysis. Their purpose in this context is to analyse microbiome data, identifying microbial communities and their interactions with the host, which is particularly important for understanding the microbiome's role in health and disease. Crucially, this application directly addresses the user requirement for handling novel and unclassified species, as highlighted in Section 2.5.2.E. Models like DNABERT-S are used to generate DNA sequence embeddings that can be compared to reference sequences for taxonomic classification. These embeddings facilitate the identification of microbial species even when they are not present in reference databases, including instances of "microbial dark matter". For feature selection in microbiome analysis, GFMs use contrastive learning to ensure that similar DNA sequences produce similar embeddings. This is especially useful for taxonomic classification, where the objective is to group sequences based on their similarity. By training the model to minimise the distance between similar sequences and maximise the distance between dissimilar sequences, contrastive learning significantly improves the model's ability to classify sequences accurately, thereby supporting the discovery of novel biomarkers.

Training Processes

The training of GFMs involves several key steps to ensure that the models are accurate, generalisable, and robust, directly addressing the requirements from Section 2.5.2.B. Raw genomic data, such as 16S rRNA sequences, is cleaned and normalised to remove noise and inconsistencies, which includes filtering out low-quality reads and correcting sequencing errors. Missing data is handled using imputation techniques, and sequences are tokenised into smaller units (e.g., k-mers) for processing by the model. The dataset is then split into training, validation, and test sets. The training set is used to train the model, while the validation set is used to tune hyperparameters and prevent overfitting. Rigorous cross-validation techniques, such as k-fold cross-validation and specifically 100 rounds of 5-fold cross-validation (T4.2), are employed to ensure that the model performs well on unseen data and to enhance its generalisability. The model's performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). These metrics are used to assess the model's ability to classify sequences and predict gene functions. The test set is used to assess the model's generalisation ability, ensuring that it performs well on new, unseen data. Finally, model optimisation involves hyperparameter tuning performed using techniques like grid search or Bayesian optimisation to find the optimal model configuration. Regularisation techniques, such as L1/L2 regularisation, are applied to prevent overfitting and improve model robustness.

Feature Selection

Feature selection is a critical step in building effective GFMs, as it helps to reduce the dimensionality of the data and improve model performance, especially important for high-dimensional metagenomic data (Section 2.5.2.A). Several techniques are employed for this purpose. GFMs use DNA sequence embeddings to represent sequences in a high-dimensional space. These embeddings capture the semantic meaning of the sequences, allowing for efficient comparison and classification. Similar sequences are mapped to nearby points in the embedding space, enabling the model to identify patterns and relationships within the data. Additionally, contrastive learning is used to ensure that similar DNA sequences produce similar embeddings. This is particularly useful for tasks like taxonomic classification, where the goal is to group sequences based on their similarity. By training the model to

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

minimise the distance between similar sequences and maximise the distance between dissimilar sequences, contrastive learning significantly improves the model's ability to classify sequences accurately, supporting the discovery of novel species (Section 2.5.2.E). An improved version of the BiGAMi algorithm (T4.1) is also employed for robust feature selection. Originally designed for discrete variables, the enhanced BiGAMi will now also accept continuous data and integrate co-variables. This allows for the removal of phenotypic variation directly caused by other sources (e.g., ethnicity, sex) before data is fed into a Random Forest (RF) model, thereby refining the relevance of selected microbiome features.

Model Explainability, Robustness, and Bias Mitigation

These aspects are paramount for the clinical utility and trustworthiness of AI-driven diagnostics, directly addressing key user requirements and challenges outlined in Section 2.5.2. To ensure that GFMs are interpretable and their insights actionable, techniques like SHAP values and LIME (Local Interpretable Model-agnostic Explanations) are used to explain the model's predictions. These techniques provide clear insights into which features are driving the model's decisions, making it easier for clinicians and data analysts to trust and understand the results. Tokenisation, the process of slicing DNA sequences into smaller units, is critical for explainability, allowing researchers to gain insights into how the model processes and classifies sequences. Users require outputs presented as ranked lists, confidence scores, and detailed reports. For robustness, GFMs are rigorously tested by evaluating their performance on diverse datasets, including data from different populations or collected under varying conditions. This ensures that the models are not overfitted to specific datasets and can generalise well to new data, meeting the stringent requirements for clinical application. Adversarial training techniques are also employed to ensure that the models are resistant to small perturbations in the input data, such as sequencing errors or noise. Regarding bias mitigation, bias in the training data is identified and mitigated using techniques like re-sampling, re-weighting, or adversarial debiasing. This ensures that the models do not exhibit biased behaviour, particularly when applied to underrepresented groups. The models are regularly audited to ensure fairness and equitable outcomes for all patient demographics, aligning with GDPR and other ethical compliance standards (Section 2.5.4).

Information Models

The information model for AI training and execution includes:

- **Data Set for Training/Execution**: The dataset used for training and testing the AI models includes genomic, metagenomic, and clinical data.
- **Model Metadata**: Information about the trained models, including hyperparameters, performance metrics, and feature importance.

3.9.2 AI Inference for Biomarker Identification

Once trained, the models are deployed for real-time biomarker discovery and validation in new patient datasets. This stage does not involve retraining but applies pre-trained models for automated classification and interpretation, delivering actionable insights to clinicians.

Inference Pipeline

1. **Data Pre-processing:** Patient microbiome samples undergo taxonomic classification using pre-defined reference databases (SILVA, Greengenes, KEGG), integrating with the advanced classification methods described in Section 2.5.
2. **AI-Driven Biomarker Prediction:** The system assigns biomarker probabilities to microbial taxa, flagging potential disease-associated patterns. Explainability techniques (SHAP, Attention Weights in Transformer Models) are actively used during inference to indicate which microbial features contribute most to the prediction, providing the transparency required by users (Section 2.5.2.C). This directly supports the preference for systems that suggest biomarkers alongside classifications with confidence scores.
3. **Validation Against Clinical Metadata:** Biomarker candidates are cross-checked with comprehensive patient demographics, symptoms, and treatment outcomes. Clinicians can interact with GenDAI's Discovery UI (T4.4), adjusting prediction thresholds based on confidence scores and interpretability outputs, enabling crucial human-in-the-loop validation and integration of domain expertise.
4. **Integration into Diagnostic Pipelines:** Validated biomarkers are logged into the Model Registry, ensuring traceability and reproducibility within the clinical workflow (Section 2.5.2.F). Biomarkers that meet clinical validation thresholds are recommended for further wet-lab validation, addressing the need for robust validation before clinical adoption.

3.9.3 UI Models

The Data Analyst runs specialised algorithms or machine learning tools on aggregated, large-scale datasets to identify novel biomarkers. Such algorithms digest data and produce data. Moreover, the algorithms often do not run one shot but perform processes and repeat activities. The visualisation of the status of the computation is important to let the user know in advance how the analysis is going, also in order to take actions, if needed, before the end of the process. The biomarker discovery can benefit from visualisations, which provide the existing data but have the potential to reveal missing data and interesting patterns. Examples of tools and techniques to find interesting patterns are reported in the state of the art of Section 2.6. The user interface for AI models is designed to meet user requirements for interaction and fine-tuning (Section 2.5.2.F):

- **Model Monitoring Dashboard:** A dashboard that displays the performance metrics of the AI models, including accuracy, precision, and recall, offering a transparent view of model performance.

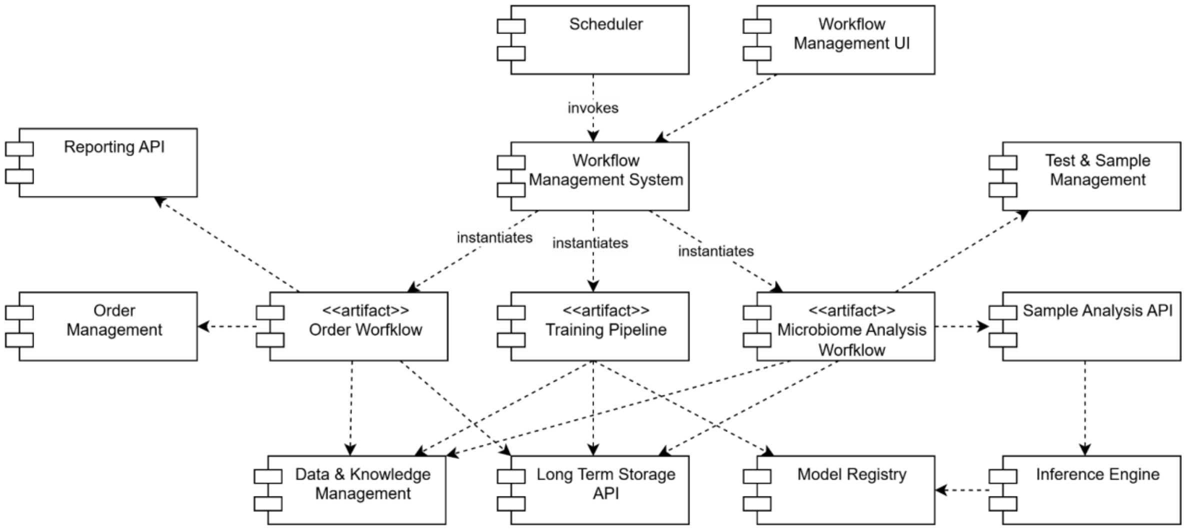
- **Explainability Interface:** An interactive interface that provides deep insights into the model's predictions, showing which features contributed to the decision and enabling clinicians to understand the *why* behind the results, directly addressing the demand for interpretability (Section 2.5.2.C).
- **Training UI (T4.4):** A simple user interface designed to allow users (like those at IMB) to fine-tune or retrain specific layers of the developed models for their own personal needs or updated sample collections. This facilitates personalised genomic analysis through the split-fed learning framework (T4.2), thereby increasing the potential for novel discovery and ensuring the tool's clinical utility and adaptability.

3.9.4 Architecture Models

The architecture for AI models and training is integrated into the overall system architecture, with components for data ingestion, model training, and inference. The AI models are deployed within the **Workflow Processing** layer, where they interact with the **Analysis APIs** to process data and generate results. The **Persistence** layer stores the trained models and their metadata, ensuring that they can be reused and audited as needed.

3.9.5 Final Component Model

As this constitutes the last use case of our use context, the component model can be finalized as follows.



Compared to the previous section, a **training pipeline** that handles data preprocessing, model training, and evaluation has been added. This training pipeline fills the model registry, which can then be used during inference.

The complete component model can thus be described as follows.

A scheduler regularly invokes a workflow management system to ensure that diagnostic and analysis processes are started and supervised at the right times. The workflow management system is responsible for instantiating specific workflows, such as an order workflow or a microbiome analysis workflow. The order workflow coordinates with the order management component to handle incoming laboratory orders and also interacts with a reporting API when results are ready to be presented. Meanwhile, the microbiome analysis workflow is launched whenever stool samples undergo sequencing or other lab procedures. It communicates with a test and sample management component to obtain the status or details of samples being processed and calls into a sample analysis API to run the required computational pipelines. The microbiome analysis workflow and the order workflow both have access to data and knowledge management services, which aggregate scientific databases, reference information, or learned insights about stool samples and diagnostic parameters. They also utilize a long-term storage API to archive or retrieve large data sets, ensuring that raw and processed data are persistently maintained and can be revisited for quality control, compliance, or future research. Through these connections, the entire system is able to schedule and execute workflows, manage orders, orchestrate diagnostics, analyze stool samples, generate findings, and then communicate the results back to users through reporting interfaces.

Explainability & Clinical Integration

A major challenge in AI-driven biomarker discovery is ensuring that predictions are interpretable and clinically actionable. GenDAI addresses this by:

- Using SHAP values to quantify feature importance enables experts to understand why a particular microbiome feature was classified as a biomarker.
- Providing confidence intervals for AI-generated predictions, allowing clinicians to adjust diagnostic thresholds dynamically.
- Enabling expert feedback loops, where biomarker predictions can be validated or overridden based on clinician insights.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union

4. Conclusions and Next Steps

This deliverable, D1.1, successfully concludes the work of Tasks T1.1 and T1.2 within Work Package 1, fulfilling all its stated objectives. The work has resulted in a comprehensive and robust foundation for the GenDAI project, establishing the necessary strategic and technical groundwork for all subsequent development.

Through a rigorous methodology combining desk research, socio-economic modeling, and qualitative interviews with domain experts, a detailed set of user requirements has been identified and compiled. These requirements meticulously document the key challenges and needs of the clinical diagnostics ecosystem across critical domains, including clinical workflows, data processing, security, artificial intelligence, and interactive visualization.

Crucially, these validated user needs have been translated into a complete suite of formal conceptual models. This includes Use Context and Use Case Models, Information Models, Component and Architecture Models, and UI Models, which together form a concrete and coherent technical blueprint for the GenDAI platform. This blueprint is complemented by a formal ethical proposal and management plan, ensuring responsible innovation is embedded in the project from its inception.

The primary value of this deliverable lies in establishing a common, agreed-upon framework that provides a single source of truth for all project partners. This de-risks future development by creating a clear and traceable link from clinical needs to technical implementation, ensuring the final product will be fit-for-purpose and aligned with the validated needs of its intended users. The outcomes of D1.1 fully meet the objectives set forth in the project's Description of Action.

With the completion of D1.1, the GenDAI project transitions from the foundational observation and theory-building phases to the systems development and implementation phases. The conceptual models and formal specifications detailed herein are not static; they will serve as the direct and essential inputs for the project's subsequent technical work packages.

The practical development and integration work will now proceed as follows:

- The architecture, component, and data processing models will directly inform the implementation of the GenDAI Diagnostics Workflow in WP2.
- The data management and security requirements, including the specifications for Persistent Unique Identifiers (PUIs) and long-term archiving, will steer the development of the GenDAI Safe & Cloud Computing Platform in WP3.
- The defined AI requirements for handling high-dimensional data, ensuring explainability, and enabling model training will guide the development of biomarker discovery models in WP4.
- The UI and visualization models will be implemented to create the user-friendly GenDAI Interactive Reporting tools in WP5.

In summary, this deliverable provides the essential roadmap and technical specifications that will steer the hands-on development phases, ensuring the GenDAI platform successfully meets the validated needs of its intended clinical and research users.

This project is funded by the European Union, project GenDAI, Grant Agreement No 101182801

Funded by the European Union