LUCIA Understanding Lung Cancer related risk factors and their Impact

Horizon Europe Grant Agreement Number: 101096473

Deliverable Number	D6.9
Deliverable Title	Conclusions of common annual meeting of the 'Understanding' cluster -M36
Due date of deliverable	31.12.2025
Actual Submission Date	23.12.2025
Responsible partner	TECH
Contributors	All partners, projects within 'Understanding' cluster
Revision (draft, 1, 2, …)	1.0
Dissemination Level	PU

Start Date of the project: January 1, 2023

Duration: 48 months

Document information\ Revision History

Description \ Status	Revision date	Authors
1st Draft	26.11.25	Lopez Adrian (MELCAYA)
2nd Draft	04.12.25	Baruch Polis (LUCIA)

Executive Board Document Sign Off

Role	Partner	Signature	Date
Project coordinator	TECH	Baruch Polis	17/12/2025
WP1 Lead	ULSTER	Jonathan Wallace	22/12/2025
WP2 Lead	VICOM	Alba Garin-Muga	22/12/2025
WP3 Lead	TECH	Baruch Polis	17/12/2025
WP4 Lead	BB	Jon Eneko Idoyaga Uribarrena	18/12/2025
WP5 Lead	UHEI	Jonathan Sleeman	22/12/2025
WP6 Lead	HOPE	Marie Nabbe	23/12/2025
WP7 Lead	TECH	Liat Tsuri	17/12/2025

Executive summary

This document summarises the discussions and conclusions as presented by the participating projects. It does not introduce new content. This document reports on the third annual meeting of the Horizon Europe "Understanding (Risk Factors & Determinants)" cluster, hosted by the MELCAYA project in Barcelona the 15th of October 2025. It brought together the five cluster projects (GENIAL, LUCIA, ELMUMY, DISCERN and MELCAYA) focused on how risk factors and health determinants drive cancer development and progression. The first part of the meeting reviewed collaborative work on FAIR data management, convergence on healthcare data standards and plans for long-term interoperability with emerging EU infrastructures such as UNCAN and EUCAIM. Scientific sessions showcased cross-project results on genetic susceptibility, exposome, omics and functional models, as well as risk stratification and early diagnosis tools, including AI-based models, smart sensors and proteomic signatures.

Further sessions addressed health-policy implementation, inequalities and citizen engagement, highlighting gaps in rare cancers, the value of patient-generated evidence and methods such as social labs and Delphi surveys to co-create recommendations. The meeting concluded with presentations from UNCAN-Connect, CANDLE and EUCAIM projects, outlining how new federated European cancer data hubs will provide sustainable, secure environments for reusing cluster data and models. Overall, the cluster is progressing towards integrated, policy-relevant evidence and interoperable infrastructures that support the EU Mission on Cancer.

Contents

Acronyms & abbreviations

Term	Description
AI	Artificial Intelligence
AJCC	American Joint Committee on Cancer
AUC	Area Under the Curve
CAR-T	Chimeric Antigen Receptor T-cell
CAYA	Children, Adolescents and Young Adults
CDSIC	Clinical Data Interchange Standards Consortium
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CT	Computed Tomography
DCAT-AP	Data Catalogue Application Profile
DICOM	Digital Imaging and Communications in Medicine
DMP	Data Management Plan
DOI	Digital Object Identifier
DPO	Data Protection Officer
ECPDC	European Cancer Patient Digital Centre
eCRF	Electronic Case Report Form
EFPIA	European Federation of Pharmaceutical Industries and Associations
EGA	European Genome-Phenome Archive
EHDS	European Health Data Space
EHR	Electronic Health Records
ELSI	Ethical, Legal and Social Implications

EOSC4Cancer	European Open Science Cloud for Cancer
EPIC	European Proteomics Infrastructure Consortium
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
FDR	False Discovery Rate
FHIR	Fast Healthcare Interoperability Resources
GC-MS	Gas Chromatography-Mass Spectrometry
GDPR	General Data Protection Regulation
GWAS	Genome-Wide Association Study
HCC	HepatoCellular Carcinoma
HTA	Health Technology Assessment
IARC	International Agency for Research on Cancer
ICGC ARGO	International Cancer Genome Consortium Accelerating Research in Genomic Oncology
IP	Intellectual Property
KPI	Key Performance Indicator
MGUS	Monoclonal Gammopathy of Undetermined Significance
ML	Machine Learning
MM	Multiple Myeloma
MOFA	Multi-Omics Factor Analysis
MRI	Magnetic Resonance Imaging
NGO	Non-Governmental Organization
NGS	Next-Generation Sequencing
OMOP	Observational Medical Outcomes Partnership

PRS	Polygenic Risk Score
QC	Quality Control
RNA / DNA	Deoxyribonucleic acid / Ribonucleic acid
SME	Small and Medium Enterprise
SNP	Single Nucleotide Polymorphism
STORM	Stochastic Optical Reconstruction Microscopy
TRL	Technological Readiness Level
VOC	Volatile Organic Compounds
VPN	Virtual Private Network
WSI	Whole-Slide Image

1 Understanding (Risk Factors & Determinants) cluster projects

MELCAYA project is part of a consortium of five projects funded by the European Commission through the Horizon Europe program (HORIZON-MISS-2021-CANCER-02-03). These projects aim to achieve the first objective of the Mission on Cancer program, which is to better understand the impact of risk factors and health determinants on the development and progression of cancer. The projects within this cluster are:

- **GENIAL:** Understanding gene-environment interaction in alcohol-related hepatocellular carcinoma [1].
- **LUCIA:** Understanding Lung Cancer related risk factors and their Impact [2].
- **ELMUMY:** Elucidating risk factors and health determinants associated with the progression of monoclonal gammopathies to multiple myeloma [3].
- **DISCERN:** Discovering the causes of three poorly understood cancers in Europe (renal, pancreatic, and colorectal) [4].
- **MELCAYA:** Developing novel healthcare strategies for melanoma in children, adolescents, and young adults [5].

The primary goal of this cluster is to support the mission objective of understanding cancer, create added value, establish a policy feedback loop and increase the impact of EU funding. Further details on the EU Mission on Cancer and its objectives are provided in D8.9 Common work plan for scientific collaboration under the “Understanding” cluster.

2 Understanding (Risk Factors & Determinants) cluster annual meeting

The MELCAYA project organized the third annual meeting of the "Understanding (Risk factors & Determinants)" cluster in Barcelona (Spain) on October 15th 2025. The meeting agenda, detailed in Annex 1, started with an opening presentation by the MECLAYA project coordination team followed by the research and innovation (R&I) sessions focused on the advancements in the different areas identified for collaboration within the cluster, such as the common practices for data/material management or the cross-comparison and integration of risk stratification and early diagnosis tools.

The afternoon sessions were focused, on the one hand, on the updates regarding the addressing inequalities and citizen engagement activities of the cluster. The second part was focused on in-depth presentation about the European federated cancer research data hub initiatives, particularly, the UNCAN-Connect, CANDLE and EUCAIM projects. This deliverable provides a detailed report of all the updates and discussions, as well as the main conclusions obtained from the sessions.

3 Sessions on the identified R&I areas for collaboration within the Understanding (Risk Factors & Determinants) cluster

The first part of the meeting was focused on reporting the advancements to date of the different cluster projects on the common topics identified for research and innovation (R&I) collaboration.

3.1 Sharing and agreeing on common practices for data management

3.1.1 MELCAYA

The project collects a wide variety of data, including clinical (retrospective EHR and registry data as well as prospective recruitment), exposomal (UV/irradiance, temperature, air pollutants from public databases and sensor networks), epidemiological (incidence/prevalence from international cancer registries (GLOBOCAN, EUROCARE, etc.), volatilomics (time-series data from VOC sensors for breath and skin), genomics, epigenomics, transcriptomics and imaging (dermatoscopic and digital pathology whole-slide images) from biobanked samples. For data collection, retrospective clinical and imaging data was retrieved from hospital archives and EHR, while prospective data collection was performed through study protocols and environmental exposures downloaded from public environmental repositories and linked to patient records via time/location. Genomic data was produced in sequencing centres and linked to clinical/imaging via pseudonymized identifiers. Omics pipelines (alignment, variant calling, expression, etc.) run with standard bioinformatics workflows and outputs organized in project storage. Clinical data was cleaned, standardized and stored in structured databases. Regarding repositories and long-term access, MELCAYA plans to deposit genomic data in the European Genome-Phenome Archive (EGA) as central secure repository for sharing and joint analysis. For imaging data, the intention is to use EUCAIM as a pan-European infrastructure for radiological and, increasingly, digital pathology images. Metadata and documentation will be prepared to enable AI-ready data for CAYA melanoma research.

3.1.2 DISCERN

The project mainly collects the following types of data: large cohort and case-cohort data (such as socio-demographic, lifestyle, anthropometric, occupational or residential histories) molecular data (such as plasma metabolomics and proteomics), genomics (including whole-genome sequencing of tumour and matched blood samples) and geospatial maps for environmental exposure modelling. Data originates from existing population cohorts and nested case-cohort/series studies. Samples are processed on standard omics platforms and sequenced data generated and quality controlled in

specialized labs. Geocoding of residential histories is linked to exposure maps. All datasets ingested into the IARC Scientific IT (SIT) portal, whose interface is powered by Innovica. The cohort/case-cohort metadata is curated in Molgenis (ontology-driven, FAIR-oriented). External partners can access data through a remote analysis environment (RStudio, Jupyter or VS Code) inside SIT. No data import/export or public internet/GitHub is necessary, only access to curated software repositories (CRAN, pip, Conda). Data access governed by data-use agreements, in a way that logs and technical controls minimize data leakage. Regarding long-term data storage, the project aims to prepare metadata according to repository standards openly. Genomic data/metadata will be prepared for European repositories (e.g. EGA, ARGO) and future UNCAN integration.

3.1.3 ELMUMY

The project works with the following types of data: large, multi-centre clinical cohorts (>6,000 preclinical myeloma patients), genomics, transcriptomics and proteomics (including a protocol to extract proteins from the “waste” phase of nucleic-acid extraction) as well as functional data from experiments in biological fluids, in vitro systems, animal and in silico models. Clinical data is captured through hospital information systems and trial eCRFs, omics pipelines run by specialised labs with standard quality control procedures and functional experiments produce diverse, lab-specific data tables and image/assay outputs. A dedicated work package collects outputs from all partners into a central integration layer (conceptually a data hub). Harmonisation and annotation of clinical, omics and functional data is performed to support progression-risk modelling. The main bottleneck is heterogeneity of formats and annotations for functional data (each lab has its own system), making harmonised analysis and storage complex. The goal is to reuse data and feed it back into new clinical trials (e.g. CAR-T target discovery), creating a loop where new data can be ingested and integrated again.

3.1.4 GENIAL

The main data types used by the project are presented, including pre-existing cohort data (up to 20 years, including clinical, lifestyle, alcohol consumption and comorbidities), new molecular data (such as circulating tumour DNA, GWAS/SNP analyses, other multi-omics layers on selected patients), imaging (radiological images, digital pathology slides with associated histological information) and structured questionnaire/eCRF data on exposures. Historical cohorts already stored in institutional databases (e.g. AP-HP) and existing quality control procedures have been applied (they are now being re-documented and harmonised). New clinical and questionnaire data is captured in electronic CRFs

and then transferred to coordinating centres. Radiology images are stored as DICOM and then segmented/delineated into regions of interest with dedicated tools and stored alongside imaging metadata. Molecular labs generate ctDNA/omics data, which are documented and linked to clinical records via pseudonymised IDs. The project has a strong focus on common ontologies and terminologies to allow cross-cohort analyses. Variables from old and new datasets are mapped to standard models (e.g. CDISC/OMOP-like structures), enabling pooled analyses and gene-environment interaction studies. Data is stored within institutional infrastructures (AP-HP and partners) with project-level aggregation, to which coordinators define access procedures and documentation for future reuse.

3.1.5 LUCIA

This project works, on the one hand, with retrospective EHR data from health systems, CT images, existing omics datasets and pathological/anatomical data. On the other, they work with prospective data from around 3.000 participants, including repeated CT imaging, clinical questionnaires, lab tests, multi-omics, environmental/exposure data and sensor-derived biomarkers. Retrospective EHR and imaging are exported from hospital systems, imaging kept in DICOM and pathology/omics in formats consistent with public datasets. Prospective data from four clinical sites is ingested through defined processes such as EHR exports, study databases, device/sensor uploads, etc. Data is centralised in a secure platform hosted at Vicomtech using a data model based on OMOP + FHIR, plus support for imaging and sensor standards. Dataset registry tracks data sources, versions and access conditions. Data access is going to be performed using clinical/patient-facing apps accessible over the internet. Research access will be regulated via VPN + multi-factor authentication to a controlled environment. Role-based access control (e.g. Keycloak), logging, backup and disaster recovery will ensure compliance with the LUCIA Joint Controller Agreement. Metadata creation and DOI assignment will be supported by the platform, designed for metadata harvesting and future federation with EU-level infrastructures (e.g. EOSC4Cancer/UNCAN).

3.1.6 Common strategy on data management

The goal of this session was to outline how the projects within the cluster collaborated to align with the EU Mission on Cancer data management and FAIR principles. The group has mainly worked on the following areas:

- Common FAIR data framework, by developing and periodically updating the shared DMP chapter, capturing common positions on FAIR data, access policies and repository choices.

- Aligning on standards and quality regarding data models (OMOP, FHIR, CDISC), imaging standards (DICOM), genomic repositories (EGA, ICGC ARGO), and internal quality control procedures.
- Planning long-term repositories and legacy, including sustainable deposit of metadata and, where possible, using European research infrastructures, both existing or in development, such as UNCAN.

All projects either already use or plan to use persistent identifiers (e.g. DOIs) and to make metadata openly available at minimum, in line with EC requirements. Interoperability is being tackled through OMOP/FHIR and similar standards for clinical data (especially LUCIA and GENIAL), DICOM and digital pathology standards for imaging (GENIAL, LUCIA, MELCAYA) and FAIR genomic repositories and community standards for omics (DISCERN, GENIAL, MELCAYA, ELMUMY). All projects recognise the importance of controlled access: de-identification, access committees and secure analysis environments. Regarding shared challenges, one of the critical ones is metadata harmonisation, as each project has developed its own way of structuring metadata. There is not yet a fully shared metadata template or common ontology set across the cluster. There is also a significant heterogeneity of legacy data: long-running cohorts and older hospital systems produce different variable names, formats and coding schemes, making harmonisation to OMOP/CDISC/FHIR challenging. There is also an uneven maturity of infrastructures: some projects already operate sophisticated platforms such as SIT for DISCERN or the Health Data Platform for LUCIA, while others are still consolidating their repository and interoperability strategy.

The next steps in the common work plan involve: (1) creating a metadata harmonisation taskforce within the cluster to compare templates, converge on a minimal common metadata core and share ontology choices, (2) draft a sustainability roadmap describing how project-specific platforms (institutional SIT, LUCIA, biobank databases, etc.) will interoperate with EU-level infrastructures and (3) perform regular updates of the common DMP chapter, aligning data access and reuse procedures to facilitate responsible cross-project reuse of data.

3.2 Cross-comparison of risk factors and molecular features

3.2.1 MELCAYA

This project shared preliminary results using a cohort of 336 early-onset CAYA melanoma patients, showing that high-risk germline variants (e.g. CDKN2A, MITF, TP53, PMS2) are relatively uncommon compared to adult series, whereas moderate-risk MC1R variants are highly prevalent. Detailed

classification reveals many pathogenic or likely pathogenic MC1R alleles (those associated with the red-hair, fair-skin, freckling phenotype) and additional frameshift or loss-of-function variants that are strong candidates for reclassification as high-risk alleles. Parallel tumour sequencing demonstrated that the vast majority of conventional and spitzoid melanomas in this age group carry canonical UV-damage signatures (SBS7 and DBS1) in 80–83 % of cases, reinforcing the central role of ultraviolet exposure and its interaction with MC1R-mediated pigmentation phenotypes, even at young ages. Attempts to apply existing polygenic risk scores show only modest discrimination versus non-melanoma controls, although most cases with elevated PRS in the absence of monogenic variants suggests a substantial polygenic component that is not yet fully captured. MELCAYA therefore plans to deploy AI-based methods (e.g. FunkRVP) that integrate rare-variant information and detailed clinical metadata to better model cumulative genetic risk.

3.2.2 DISCERN

This project has integrated cancer registry data with whole-genome sequencing from kidney cancers, linking specific mutational signatures (SBS4TV) to regions with high renal cancer incidence, notably the Czech Republic and Lithuania. To understand what exposures underlie these signatures, DISCERN combines large-scale environmental mapping with molecular analyses. Long-term data on air pollution, temperature, vegetation, light-at-night, soil pollution, radiation, pesticides, water contamination and candidate toxic agents are being modelled to define “risk landscapes”. One key example is the distribution of the toxic plant *Aristolochia Clematitis*, which produces aristolochic acid. Its unexpectedly wide presence in Eastern and Central Europe suggests that exposure is not limited to herbal remedies but may occur through more general environmental routes, consistent with observed clusters of liver and renal cancers. To move from association to mechanism, DISCERN performed plasma metabolomics and experimental modelling. Mass spectrometry has revealed around 20 clusters of plasma metabolic features. One cluster characterised by TMAP, a marker of renal function, correlates strongly with the SBS4TV signature and may serve as a proxy of the underlying mutagenic exposure. In parallel, kidney organoids exposed to aristolochic acid show distinct transcriptomic and proteomic changes, including regulators that are also altered in tumours and appear linked specifically to clear-cell kidney cancer. These data start to explain why this exposure leads preferentially to particular tumour types and provide a mechanistic bridge between environmental risk factors and molecular phenotypes in patients.

3.2.3 ELMUMY

This consortium has identified five actionable molecular features that stratify progression risk using multi-omics and clinical data. First, a purely clinical classifier built from routinely collected variables outperforms existing prognostic tools. Developed initially on Greek cohorts and now being validated on German and Spanish data, this model can be implemented immediately in standard practice, providing a "low-tech" yet powerful risk stratification tool. Second, a urine peptidomics–based classifier, developed with an industrial partner, accurately distinguishes high- and low-risk patients using a completely non-invasive sample. This patented approach is particularly attractive because it avoids bone marrow aspiration and could be deployed widely once further validated. Third, the project has identified a set of flow-cytometry markers, measured in everyday diagnostic panels, that improve prediction of progression. Importantly, inter-laboratory analyses in Navarra and Athens showed excellent reproducibility, supporting their real-world applicability in clinical decision-making. Fourth, proteomic and cellular analyses have revealed a panel of progression-associated proteins, including the regulator MDM2, which appears as a recurrent player across different cancer types. While not yet directly actionable, these molecules constitute promising biomarkers and potential therapeutic targets. Finally, using cutting-edge direct stochastic optical reconstruction microscopy (dSTORM), ELMUMY partners have identified a specific surface receptor on plasma cells that is suitable as a CAR-T target. A first-in-human clinical trial based on this discovery is planned, directly linking the molecular work on progression to innovative preventive immunotherapies.

3.2.4 GENIAL

This project leverages several large cohorts and an ongoing fast-track clinical trial to perform genome-wide association studies (GWAS) and rare-variant analyses in more than 8.000 individuals with chronic liver disease and HCC. HCC has been shown to be highly heterogeneous, but largely preventable: major risk factors include chronic hepatitis B and C, alcohol use, obesity and metabolic dysfunction–associated steatotic liver disease, aflatoxin exposure and, in some regions, aristolochic acid. Less than 3 % of HCC patients carry high-penetrance Mendelian predisposition variants (e.g. in disorders like hemochromatosis or specific cancer-predisposition syndromes), but common and rare variants in genes such as PNPLA3 and TM6SF2 modulate progression from steatosis through cirrhosis to cancer, particularly in the context of alcohol and metabolic risk factors. On the tumour side, GENIAL maps how distinct exposures translate into somatic mutational signatures and structural alterations. For example, aflatoxin B1–related signatures and tobacco-associated patterns can be clearly identified in some

tumours, alongside copy-number changes in genes linked to exposure-related predisposition. In alcohol-related HCC, specific molecular features have been noted, including characteristic TERT-promoter mutations, particular mutational signatures and telomere shortening. Collectively, these analyses show that molecular diversity in HCC is tightly intertwined with the underlying risk factor profile, with implications for prognosis and for tailoring prevention strategies (vaccination, alcohol reduction, metabolic control and elimination of carcinogenic exposures).

3.2.5 LUCIA

The consortium has performed a geospatial analysis combining an environmental database (air and soil pollutants, industrial emissions, radon, urbanisation, green space, social deprivation) with clinical data from Basque Country and Liège cohorts, showing that radon exposure and low access to green space are among the strongest predictors of lung cancer, with additional contributions from particulate matter and soil contamination by nickel and antimony. To move beyond association, LUCIA has selected ten candidate carcinogens (including particulate matter and antimony-containing pollutants) and is testing their impact on transformation in vitro. Using a CRISPR-Cas12 system, the consortium can inactivate combinations of 24 lung cancer-relevant tumour suppressor genes in A549 cells and then assess how each carcinogen modifies 2D and 3D growth. This produces a “heatmap” of exposure-genotype interactions, mapping which environmental agents synergise with specific tumour suppressor losses in driving transformation. Population-based omics modelling, particularly in the ESTER cohort, has identified robust epigenetic markers of risk. Hypomethylation (and likely increased expression) of AHRR and F2RL3 is strongly associated with lung cancer risk, not only in current smokers, but also in former smokers, suggesting these loci act as long-lasting molecular scars of tobacco exposure. Functional work in a premalignant bronchial epithelial model shows that AHRR knockdown alters colony fitness and morphology, consistent with a role in regulating cellular plasticity relevant to carcinogenesis. Finally, LUCIA has identified ID1 and ID3 as key stemness-associated drivers whose expression tracks with stepwise transformation in HBEC3-derived models and correlates with poor prognosis in patient tumours. A newly developed coumarin inhibitor efficiently reduces ID1/ID3 expression at sub-micromolar concentrations and suppresses colony formation in soft agar. Ongoing in vivo work will test whether this strategy can clear premalignant lung lesions, opening a path towards chemopreventive interventions in high-risk individuals.

3.2.6 Cross-project conclusions

Across DISCERN, ELMUMY, GENIAL, LUCIA and MELCAYA, a coherent picture emerges in which cancer risk is shaped by the interaction between external exposures (pollutants, lifestyle, infections, UV,

alcohol) and molecular vulnerability (germline variants, somatic mutations, epigenetic changes and pathway deregulation). DISCERN and LUCIA demonstrate that high-resolution environmental mapping (air and soil pollution, radon, green space, toxic plants) can be quantitatively linked to mutational signatures and transformation models, identifying candidate carcinogens and exposure hotspots. GENIAL extends this to metabolic and alcohol-related liver disease, showing that relatively common genetic variants modulate the impact of lifestyle and toxins on HCC risk and biological drivers. ELMUMY and MELCAYA highlight how multi-omics and germline sequencing can yield clinically relevant risk tools, from clinically based and urine-peptidomic classifiers in myeloma to refined genetic risk profiles in young melanoma, with a shift from rare high-penetrance mutations to more polygenic and phenotype-linked susceptibility. Finally, LUCIA bridges discovery and intervention by functionally validating risk markers (e.g. AHRR, ID1/ID3) and testing pharmacological strategies to clear premalignant lesions. Together, the projects provide a strong proof-of-concept that integrating exposome data, genetic/epigenetic susceptibility and functional models can deliver actionable biomarkers, risk scores and preventive interventions aligned with the goals of the EU Mission on Cancer.

3.3 Technology, tools, knowledge and best practices for data exploitation and computational modelling

3.3.1 MELCAYA

This project has been developing a pan-European digital pathology focused on pediatric and young-adult melanoma patients. The main goal is to create a secure, interoperable platform that lets pathologists and researchers across Europe share whole-slide images (WSIs), expertise and AI tools under robust technical and organisational conditions. After a detailed requirement analysis, MELCAYA selected Halo Link as the platform backbone, a software already deployed at the University of Florence, as it can handle very large WSIs (0.5–4 GB), support multiple proprietary and non-proprietary scanner formats, provide fast web-based visualisation and allow fine-grained role-based access (view-only vs contributor). The team defined key features such as secure authentication, unique case identifiers, flexible folder/case management, annotation and measurement tools, and advanced image-analysis capabilities, and then systematically tested Halo Link against these criteria.

MELCAYA conducted a multicenter quality-control study of WSI acquisition and sharing. They collected 311 slides from 208 young melanoma patients across 11 institutions and systematically classified

defects into pre-analytical and analytical categories, grading severity as critical, major or minor. Defects were found in 144 slides, with substantial inter-institutional variability, highlighting the need for standardised protocols and feedback mechanisms. To support best practices, the group also developed a digital scanner-parameter form capturing centre characteristics (scanner brand/model, software, magnification, pixel size, colour depth, compression, QC settings and AI features). This structured metadata will be used to design future standardisation protocols and to interpret variability in AI performance across centres.

3.3.2 DISCERN

DISCERN has established a secure, centralised analytics environment at IARC (the SIT platform) that now underpins all its data exploitation and modelling work. Data arrive into a restricted "data manager" area where they are curated and harmonised, then published into shared folders accessible to partners, alongside a common code, documentation space and project-specific analysis areas. This setup lets researchers work as if on their own desktop (RStudio, Jupyter, VS Code) while keeping data fully contained within IARC infrastructure. Scientifically, DISCERN has developed and is validating pipelines for untargeted metabolomics and proteomics tailored to its complex case-cohort and case-series designs. Implemented as R packages, these pipelines allow users to systematically test multiple pre-processing strategies (normalisation, batch correction, filtering) and then quantify how robust downstream association results are across alternative workflows, a critical best practice for high-dimensional omics.

For statistical modelling, DISCERN has adapted and stress-tested methods for univariate and multivariate omics analyses in non-standard case-cohort settings, where only a fraction of cases are sampled. Many off-the-shelf tools are not directly applicable, so the group evaluates candidate models through simulation to ensure valid estimation and inference in these designs. These evaluations themselves constitute a methodological result and are being bundled into reusable code for the wider community. A key scientific contribution is a federated analysis framework for situations where cohort-level data cannot be centralised. Using EPIC proteomics data as a testbed, DISCERN compared classical meta-analysis to new federated algorithms that iteratively exchange summary statistics to approximate a pooled analysis. They show that the federated approach can reproduce pooled-data results much more closely than standard meta-analysis, especially when effect heterogeneity and high-dimensional markers are involved, providing a practical blueprint for privacy-preserving pooled modelling in future exposome and omics studies.

3.3.3 ELMUMY

This project has developed a cloud data warehouse that ingests transcriptomic, proteomic, epitranscriptomic, peptidomic and functional data from patients, cell lines and animal models into a common integration space. This warehouse hosts both processed data and metadata. It is being coupled to MLflow-style pipelines to ensure traceable and trustworthy deployment of AI models emerging from the project. The consortium has already produced several published and in-submission results using public data as a proving ground for the tools that will later be applied to ELMUMY's experimental cohorts. One major pillar of work is the state-specific biomarker discovery across the MGUS to MM spectrum. Using the concept of monotonicity, they identify genes that are consistently up- or down-regulated across these three stages, which are then used to construct ratios of monotonically expressed gene pairs as features. These gene-ratio signatures show strong discriminative power between precursor and symptomatic disease and highlight mechanistic pathways involved in progression. The methodology is designed to be directly re-usable as a predictive tool once ELMUMY's own transcriptomes are fully available.

On top of that, the project has developed a drug-repurposing pipeline for multiple myeloma. Starting from disease-specific transcriptomic signatures, the team is identifying drugs whose expression response profiles reverse these signatures, then refines candidates using structural similarity and combination-synergy modelling between approved drugs and new candidates. This has yielded stage-specific shortlists of repurposing candidates and combinations that are already being considered for experimental testing by consortium partners, turning computational modelling directly into hypotheses for the bench. Finally, ELMUMY has also begun applying multi-omics integration tools (MOFA, mixOmics and network-based approaches) to combine transcriptomic and proteomic datasets, uncovering latent factors and patient subgroups that cut across individual data types. These integrated factors are expected to form the basis for more refined AI models of progression and treatment response.

3.3.4 GENIAL

This consortium is developing a deep learning–based prediction of HCC risk in cirrhotic patients, with a strong emphasis on exploiting modern models for digital pathology. The group applies an openly available pipeline that processes WSI liver biopsies by tessellating them into small patches, extracting high-dimensional features via pre-trained foundation models, and training downstream risk-prediction networks using matched clinical outcomes. The pipeline includes rigorous five-fold cross-validation

with repeated train/validation/test splits to quantify variability in deep model performance, as well as explainability tools that generate heatmaps showing which image tiles drive predictions. A notable scientific result from their benchmarking work is that the best performing foundation model was not the one pre-trained on the largest number of images, but the one trained on the most diverse set of cancer types and tasks.

This finding, derived from comparisons across multiple recent pathology foundation models on several biomarkers and prediction tasks, suggests that data variety in pre-training is more important than the sheer volume for downstream generalisability. Building on this insight, GENIAL is now developing a liver-specific foundation model tuned to pre-cancer cirrhotic biopsies, arguing that models pre-trained only on late-stage liver cancers may capture little truly healthy or pre-neoplastic tissue. They are currently applying this framework to around 1.000 biopsies from three cohorts with different etiologies (alcohol, viral, mixed), with the goal of delivering robust, explainable HCC-risk scores that can be integrated into future clinical decision tools.

3.3.5 LUCIA

This project has created a health data platform that not only stores multimodal lung-cancer data but tightly integrates them with analytical and AI toolchains. The platform supports clinical, imaging, omics, environmental and sensor data, providing a virtual research environment for interactive exploration, advanced computational analysis and visual analytics dashboards. Most modelling work is executed directly inside this secure platform, which enforces pseudonymisation and granular access control while tracking versions of data, models and algorithms used. Several results have been produced on this infrastructure to date. On the one hand, a clinical risk-factor model that uses a simplified stacking ensemble on a small set of routinely available variables (age, smoking history, comorbidities and environmental exposures) to predict who should be prioritised for screening. The published model matches or exceeds the performance of established lung-cancer screening scores while using fewer features, making it easier to implement in routine care. Second, a time-aware EHR model based on transformer architectures that leverage longitudinal electronic health records to predict 1-year lung-cancer risk. These models have been trained and validated on large health-system datasets (e.g. Andalusia) and externally evaluated in other European hospital systems, demonstrating promising generalisability. Finally, a CT-nodule malignancy classifier that uses large-scale self-supervised pre-training on unlabelled nodules followed by fine-tuning on labelled cases. On an independent test set of 1.000 nodules, the model achieved around 90 % macro-precision and 88 % recall, providing a strong decision-support tool for radiologists, including hints at molecular subtypes.

3.3.6 Cross-project conclusions

There is a clear convergence across all projects towards secure, cloud-based environments that bring multimodal data next to scalable analytics. DISCERN's IARC SIT platform, LUCIA's Health Data Platform and the ELMUMY cloud/database/warehouse stack all expose curated datasets together with analysis tools, versioned code and model artefacts, enabling reproducible exploitation, while keeping data inside controlled infrastructures. MELCAYA extends this paradigm to very large WSI through a digital pathology hub (Halo Link), defining practical requirements for performance, security and usability across many hospitals. On the computational modelling side, projects increasingly rely on advanced AI, including foundation models and deep-learning pipelines for histopathology in GENIAL, multimodal integration tools (MOFA, mixOmics or network-based methods) and MLflow-governed AI lifecycles in ELMUMY or transformer-based EHR models plus self-supervised CT nodule networks in LUCIA. DISCERN focuses on robust, shareable R packages for omics preprocessing, high-dimensional statistical modelling and federated analyses that mimic pooled data while respecting local constraints. MELCAYA's work on slide repositories and scanner metadata highlights the importance of detailed technical documentation and QC as a prerequisite for any downstream AI. Together, these efforts define emerging best practices: co-design of infrastructure and models, strong emphasis on reproducibility and explainability, federated or privacy-preserving analytics when needed and early planning for interoperability and reuse across the cluster.

3.4 Cross-comparison of risk stratification/early diagnosis tools

3.4.1 MELCAYA

This project is working of VOC-based technologies for the detection of advanced melanoma and the development of risk-stratification algorithms. Building on evidence that melanoma lesions emit distinct VOC signatures (previously detectable even by trained dogs and by GC–MS in cell culture comparisons), the consortium is testing a breath analyser and skin patch developed by the Israel Institute of Technology (TECH) in a dedicated exploratory study. Around 100 patients with stage II–IV melanoma are going to be recruited and stratified into non-metastatic and metastatic groups. Baseline and 3-6-month follow-up measurements using breath and patch devices are being collected and benchmarked against standard imaging and, where available, liquid biopsy. By the time of the presentation, 62 patients had been recruited, 77 % of which did not have active distant metastases at first visit, while 22 % did, providing an initial cohort for evaluating the sensitivity of VOC-based

monitoring to progression. On top of this, partner Athena Tech is training risk/prognostic models on a large adult melanoma dataset with rich clinical, phenotypic and genetic information (around 21.000 cases). Early results suggest that their algorithm achieves a concordance index approximately 90 %, compared with the 75% for the standard AJCC staging system, indicating a substantial gain in prognostic discrimination, although detailed model specifications remain confidential due to industrial IP.

3.4.2 DISCERN

DISCERN has generated large-scale pre-treatment plasma proteomic data for three major cancers using the Olink platform, which measures around 1.500 proteins via dual-antibody, NGS-readout assays. The study includes around 900 colorectal, 1.000 renal and 500 pancreatic cancer patients, with blood collected at diagnosis before any surgery or systemic therapy. Across all three cancers, hundreds of proteins show statistically robust associations with all-cause mortality after stringent FDR correction. Known clinical markers have been rediscovered (e.g. CEA for colorectal cancer, KIM-1/HAVCR1 and CA9 for renal cancer), alongside novel candidates such as DCTPP1, involved in DNA integrity. For pancreatic cancer, analysis of around 60 % of the planned cohort already reveals a similar pattern of broad proteomic associations. Cross-cancer comparison shows that the protein hazard ratios for prognosis are highly correlated (around 70 %) between cancer types, indicating that many circulating proteins predominantly reflect shared systemic processes leading to death (thrombosis, heart failure, cachexia and frailty) rather than tumour-specific biology. This suggests that proteomic signatures may be particularly powerful for general prognostic stratification and for identifying targets to mitigate lethal systemic complications, even if they are less suited as purely disease-specific biomarkers.

3.4.3 ELMUMY

This project is focused on addressing risk stratification in precursor states of multiple myeloma where most individuals remain asymptomatic, with a minority undergoing disease progression. Using data from about 1.000 patients from Athens (94 progressors, the remainder non-progressors), the project builds survival models to predict both whether and when progression will occur, based on 22 baseline clinical and laboratory predictors, including bone marrow infiltration. Several statistical and machine-learning approaches were trained with stratified cross-validation, carefully addressing strong class imbalance (around 10 % progressors) via inverse probability of censoring weights and time-dependent AUC metrics. The best-performing models, particularly random forests and support-vector machines, consistently outperform the current “20/2/20” clinical risk model across internal validation, synthetic

data-based testing and an external validation cohort of 550 additional patients. Time-dependent AUCs between 12 and 36 months reach around 0.81 for random forests, clearly higher than the clinical benchmark, with concordance indices showing similarly improved discrimination. These gains persist when models are trained on synthetic data and evaluated on real patients, underscoring their robustness and potential for future clinical deployment.

3.4.4 GENIAL

This project is building on already powerful clinical algorithms that stratify cirrhotic patients under HCC surveillance into low, intermediate and high-risk groups (average annual risk 0.5-7 %), using demographics, liver function and comorbidities. The project's main scientific advance so far is to demonstrate that a simple clinical score can robustly identify a subgroup with > 3 % annual HCC incidence, which is the threshold at which MRI-based screening becomes cost-effective. This has been confirmed in the ongoing FAST-TRACK trial in France, where around 1.000 high-risk patients (selected by this algorithm) are randomised to ultrasound versus abbreviated MRI, with the observed incidence exceeding 3 %, validating the model's calibration. In parallel, GENIAL is prospectively collecting environmental exposures, constitutional DNA, circulating proteins, liquid-biopsy markers and radiomics features in approximately 4.000 patients enrolled in surveillance cohorts. Machine-learning models integrating clinical data, polygenic risk scores and serum proteins already identify "extreme phenotypes" (e.g. young patients with HCC risks comparable to much older individuals) and are expected to increase the C-index of risk prediction, thereby enriching surveillance for those most likely to benefit and enhancing detection of small, curable HCC lesions.

3.4.5 LUCIA

This project presented the work performed on developing three non-invasive, rapid diagnostic platforms for early lung cancer detection using volatile and blood biomarkers coupled to AI. The first one is a handheld breath analyser based on an array of gold-nanoparticle/polymer chemoresistors that detects multi-dimensional patterns of volatile organic compounds (VOCs) in exhaled air. Over 2.500 breath samples have been collected so far, with > 3.000 disposable kits distributed to clinical sites. Preliminary machine-learning models (CatBoost, random forests) achieve around 82 % accuracy in distinguishing lung cancer cases from controls, with results available in under five minutes at the point of care. Second, a wide-spectrum biomarker skin patch combining gas-phase sensors and microneedle electrochemical sensors has generated more than 1.000 measurements from 240 participants across four sites, with a balanced high vs low-risk datasets. This yields 85 % classification accuracy for risk

status, supporting its potential as a scalable precision-screening tool. Finally, a low-cost spectrometry-on-card platform has been developed to collect time-series optical spectra from blood on graphene/gold nanoparticle paper electrodes (>5.000 measurements so far). Although early spectra overlap substantially between samples, ongoing temporal AI modelling aims to sharpen diagnostic discrimination. All three devices feed into a unified cloud-based AI and data pipeline that integrates clinical, imaging and omics data. The convergence of smart sensors and AI is positioning LUCIA's tools for integration into future lung-cancer screening workflows, provided ongoing work confirms robustness across sites, populations and clinical contexts.

3.4.6 Cross-project conclusions

Taken together, these projects demonstrated complementary strategies for improving risk stratification and early diagnosis along the cancer continuum. GENIAL and ELMUMY illustrate how refined clinical-genetic models and survival-analysis-based AI can identify high-risk individuals within well-defined at-risk populations (cirrhosis, MGUS/MM), enabling more intensive imaging or follow-up where it is both clinically beneficial and cost-effective. LUCIA and MELCAYA bring forward a new class of smart-sensor technologies (breath analysers, skin patches and paper-based spectrometry) that can non-invasively capture complex biochemical signals and, when combined with AI, already reaching accuracies of 80-85 % in distinguishing cases or high-risk states, with clear potential for scalable screening and longitudinal monitoring. DISCERN shows that high-dimensional proteomics at diagnosis can yield powerful prognostic information across types of cancers, reflecting shared systemic pathways of cancer lethality that could be targeted to improve survival beyond tumour-directed therapy alone.

Overall, the main scientific message is that integrating multimodal data (clinical, environmental, genetic, imaging, circulating biomarkers and sensor readouts) with advanced machine-learning and survival-analysis techniques can significantly enhance risk prediction and early detection across cancer types. These efforts are converging towards more personalised, cost-effective surveillance and diagnostic pathways, directly supporting the goals of the EU Cancer Mission.

3.5 Sharing best practices on implementation of healthcare policies

3.5.1 MELCAYA

The project described the work carried out on health-policy recommendations specific for CAYA melanoma. The review of 29 national cancer control programmes showed that rare cancers such as those in CAYA received very limited attention. Key gaps include lack of definitions of centres of expertise, referral pathways for early detection, structured epidemiological monitoring and systematic

involvement of patients in policy design. Psychological support for young patients and their families is also insufficiently addressed. A patient-journey task identified multiple weaknesses, such as limited expertise in primary care (and even among some specialists), barriers to accessing innovative treatments and research protocols, structural problems such as lack of referral centres, problematic behaviours or attitudes from some professionals/insurers and a heavy emotional, social and financial burden on families. The project emphasises the need for strong patient advocacy and for patient agencies to be fully recognised as partners in care planning.

On the ethical, legal and social side, six priorities were highlighted: addressing cultural factors (e.g. tanning behaviours), improving professional training, managing over and under-diagnosis, promoting patient autonomy via transparent communication, reducing inequalities in access to care and handling the ethical/legal implications of digital early-detection tools. MELCAYA has also developed a structured framework and transferability tool for assessing digital health technologies for melanoma, covering both clinical and non-clinical dimensions and applicable from early (TRL 4–7) to late (TRL 8–10) stages of technology development. Finally, a Delphi process structured in six sections on awareness, ELSI, innovation assessment/reimbursement, data infrastructure, disease management and research, is being built to reach expert consensus on concrete policy recommendations for CAYA melanoma.

3.5.2 DISCERN

This consortium showed how patient-generated data is being used to inform cancer policy for kidney, colorectal and pancreatic cancers. A recurring example was the Global Patient Survey led by the International Kidney Cancer Coalition, which collected responses from more than 2.700 patients and carers in 46 countries. The survey revealed substantial information gaps, such as that many patients still do not know their tumour subtype, stage, survival chances, risk of recurrence or available treatment options. These deficits undermine shared decision-making and appropriate treatment choices. The survey also explored attitudes toward biomarkers. Even in kidney cancer, where predictive biomarkers are not yet in routine use, patients expressed strong trust in biomarker-guided therapy but also raised questions that must be addressed to ensure acceptability when such tools become available. Results from the survey and related work feed directly into updated scientific and policy briefs developed with professional societies, with the aim of influencing European and national authorities on cancer care standards and research priorities.

Additional work from Digestive Cancers Europe and Pancreatic Cancer Europe has focused on early-onset colorectal cancer, calling for adaptation of screening ages, survivor support and long-term

survivorship policies, whereas on pancreatic cancer it was more focused on calling for dedicated investment in research, clinical trials, databases, biobanks and specific guidelines, reflecting its extremely poor prognosis. The presentation concluded by stressing cross-cancer needs, including robust national cancer registries connected via European networks, improved epidemiological monitoring, adaptation of screening programmes for early-onset disease, better integration of genetic predisposition into early diagnosis and anticipatory research on lifestyle and environmental risks to inform prevention policies.

3.5.3 LUCIA

The project reported on the "social lab" methodology implemented to identify and prioritise barriers for creating a complex early-detection toolbox (breath analysis, skin patch, blood test and AI) in real-world lung cancer screening and care. Mixed workshops with consortium partners and external stakeholders (other screening projects, professional and patient organisations) generated and categorised 80 distinct barriers in 10 domains, later published in a white paper. A two-axis prioritisation followed: stakeholders rated each barrier's impact on successful implementation, while consortium members rated their own ability to influence it within the project timeframe. This produced a set of 38 high-impact, high-influence barriers for which the consortium developed concrete internal actions (e.g. improving clinical workflow integration, communication with professionals and patients, data quality and governance).

Another 34 barriers were judged high-impact but low-influence for LUCIA alone (e.g. systemic workforce shortages and overburdened healthcare professionals). For these, the consortium formulated recommendations targeted at policymakers, health authorities and professional bodies, and then refined them through an online workshop with external stakeholders. The team also presented an AI impact-assessment tool (Surikas), which automatically scans project documents to flag risks around bias, explainability and transparency, and proposes mitigation measures and actions. This tool, piloted in LUCIA, is intended to support responsible AI governance and alignment with emerging regulatory and ethical requirements.

3.5.4 Cross-project conclusions

Across DISCERN, LUCIA and MELCAYA, several common best practices for implementing healthcare policies emerge. First, patient-generated evidence (surveys, journey mapping) are essential to identify real information gaps, access barriers and psychosocial needs, and should systematically feed into policy briefs and national cancer plans. Second, participatory methods such as social labs, focus groups

and Delphi surveys enable multi-stakeholder co-creation of prioritised, actionable recommendations, particularly for complex AI-enabled tools and rare cancers. Third, robust data infrastructures (registries, interoperable systems) and structured HTA frameworks are critical to support equitable screening, referral and reimbursement decisions. Finally, dedicated attention to ethics, legal protections, workforce constraints and communication with patients is needed to ensure that innovative technologies and policies translate into fair, trustworthy and sustainable cancer care across Europe.

4 Session on addressing inequalities

The presentation reported on the work of the cluster group focusing on addressing inequalities in access to cancer care in Europe, whose goal is to develop policy recommendations. In 2024, the group decided to ground its work in an in-depth review of existing evidence. They started from a broad 2018 European Commission report on healthcare access and then identified three additional key sources: an EFPIA report on disparities across the cancer continuum, a concise position paper from Cancer Patients Europe (with five calls to action on prevention and early diagnosis, diagnostics and treatment innovation, survivorship and smoking prevention) and a EuroHealthNet report focused on social determinants of health and how these shape inequalities in access to cancer care. By synthesising these documents, the cluster defined three priority policy areas to structure its future work: (1) equal access to cancer prevention and early detection, (2) equal access to diagnostics and innovation and (3) equal access to treatment, care pathways, medicines and survivorship support. They intend to pay particular attention to healthcare workforce shortages and to the potential role of AI in mitigating or exacerbating inequalities.

The planned timeline includes a general awareness-raising social media campaign in November 2025, followed by three themed webinars in 2026–2027, each accompanied by a focused campaign: “From awareness to access” (equitable prevention and screening), “Breaking the diagnostics divide” (access to innovative diagnostics), and “Access to care for all” (treatment and survivorship). The insights from these webinars and stakeholder inputs will feed into drafting and disseminating the final policy recommendations and report in 2027. During discussion, participants suggested adding a recent IARC study on cancer inequalities in Europe as another important evidence source.

5 Session on citizen engagement

The working group on citizen engagement stressed that patients and citizens are not just beneficiaries of cancer research but key partners in shaping it, underlining that treatment outcomes depend heavily on patients understanding their disease and treatment options. Shared decision-making was presented as a central element of proper cancer care. When patients are genuinely involved in treatment choices, adherence improves and so do survival and quality-of-life outcomes. To support this, the team is running a European study (UK, Germany and Netherlands) on how shared decision-making is implemented and what cultural barriers exist. They are organising workshops with clinicians and scientists to promote patient advocacy and its benefits. The presentation then reviewed concrete citizen-engagement outputs from the working group: a cluster video already online, a brochure introducing the cluster and project presentations hosted on the websites of the five consortia with their partner patient organisations. A first podcast has been produced and there is agreement to create similar ones for the other projects plus a dedicated episode with a patient organisation to give stronger voice to patients.

On social media, communication officers from all consortia meet regularly to coordinate messaging, develop annual roadmaps and monitor basic KPIs (posts, views, likes). Campaigns are deliberately concentrated around key events (consortium meetings, World Cancer Day, World Cancer Research Day, World Mental Health Day, cancer awareness weeks/months or European Week Against Cancer) to maximise impact. Finally, the talk described the future working plan, including scaling up a successful social-media workshop model across all consortia, organising a “Flip the Classroom” patient-advocacy workshop at the next DISCERN annual meeting, in which patients will lead the discussion on shared decision-making with scientists and clinicians, continuing dissemination of the video and brochure, producing new podcasts, filming a final cluster video at the last meeting, exploring short interview videos with project leaders and a possible “Night of Science” event (subjected to resources available).

6 Sessions on European federated cancer research data hub initiatives

6.1 UNCAN-Connect

This project is presented as a Cancer Mission flagship initiative aiming to create a decentralised, collaborative European cancer data network to accelerate research, innovation and better care. The project has just started and brings together 53 partners from 19 countries, including universities,

hospitals, research infrastructures, SMEs, NGOs and patient organisations, covering expertise from oncology and data science to governance, cybersecurity and AI. Its core concept is to develop a federated architecture composed of a Central Platform and multiple National Cancer Data Nodes connected to diverse institutional and thematic data holders. Rather than centralising data, UNCAN-Connect enables secure, trusted secondary use of sensitive health data "at source", supported by a common data governance, compliance and operational framework. This framework aims to define shared values and a trust model, ensure alignment with the European Health Data Space (EHDS) and AI regulations, and provide common guidelines for data processing, provenance, interoperability and reproducibility across Member States.

Technically, the project will develop tools and workflows for local data processing within cancer data nodes, and a central platform offering services for data exploration, analysis, computing and federated learning via trusted research environments. The platform is conceived as a generic health data-sharing infrastructure validated in real-world oncology use cases. Eleven use cases across six cancer areas (pediatric cancers, pancreatic adenocarcinoma, lymphoid malignancies, lung cancer, ovarian cancer and prostate cancer) form the "engine" of the project. They address key questions along the cancer care spectrum: early prediction of chemotherapy-related fever and cardiovascular impairment in children, biomarker-based early detection of pancreatic cancer, AI-supported prediction of relapse and late complications in lymphoid malignancy survivors, multimodal risk prediction, early detection and treatment response modelling in lung cancer, personalised decision-analytic modelling and AI-enhanced robotic screening in ovarian cancer and better selection of men for prostate biopsy based on multimodal data and AI. These use cases provide heterogeneous data (clinical, imaging, omics, wearables and preclinical models) to drive design, stress-test interoperability and demonstrate added value of trusted data sharing.

A strong emphasis was placed on stakeholder engagement and long-term sustainability. Dedicated work packages will develop strategies to engage national and pan-European stakeholders (research infrastructures, clinicians, policymakers, patients and industry), create formal partnerships and build capacity through training and collaborative learning. Communication and dissemination activities aim to make outputs visible, understandable and actionable for the broader cancer community. Impact will be continuously assessed by a patient-led team across 10 countries, looking at how well UNCAN-Connect enables re-use of cancer data, what challenges and benefits arise and how lessons can inform policy and infrastructure planning at national and EU level. Overall, UNCAN-Connect is positioned as a

foundational ecosystem for secure, interoperable cancer data use in Europe, co-created with end users and validated through clinically meaningful use cases spanning prevention, diagnosis, treatment and survivorship.

6.2 CANDLE

This project is focused on building a network of National Cancer Data Nodes to support two flagship EU platforms: the UNCAN.eu research platform and the European Cancer Patient Digital Centre (ECPDC). The idea is to create a hub and node structure: a central EU platform connected to national nodes that organise data, tools and communities within each Member State. The consortium brings together 40 partners from 20 Member States (31 beneficiaries and 9 affiliated partners) and 6 major European research infrastructures (covering omics, biobanks, translational research, exposome, clinical research and imaging). It closely collaborates with UNCAN-Connect, the EUCP project (building the patient portal), ECHoS (Cancer Mission hubs), the Genome Data Infrastructure and the 1+ Million Genomes initiative, as well as several Joint Actions on comprehensive cancer centres, personalised cancer medicine and cancer registries.

The importance of national nodes is highlighted, particularly the fragmented cancer data landscape, which leads researchers to reinvent data solutions in isolation or the lack of technologies that are interoperable. The European Health Data Space (EHDS) offers a unique opportunity to create a blueprint to align infrastructures across countries. CANDLE is planning to use this to bring together national communities, registries, cancer mission hubs, comprehensive cancer centres and other initiatives under a coherent data structure. Overall, National Cancer Data Nodes should fulfil the following core functions:

1. Act as a link and support structure for national cancer data holders, reducing their burden.
2. Improve data quality, ideally already at source.
3. Maintain national metadata catalogues interoperable with EU-level catalogues.
4. Provide common data variables for cancer datasets.
5. Offer secure processing environments for data analysis and tools.

CANDLE surveyed 22 countries and identified several possible national-node models: a single empowered agency (e.g. Finland), a node built on national research-infrastructure nodes (e.g. HealthRI plus the cancer registry in the Netherlands), leading cancer institutes in smaller countries, regional constellations where health systems are decentralised (e.g. Spain, Germany) and cancer registries as a

starting point in less-developed settings. The project is now running workshops and surveys with national representatives to capture user needs, map technical and organisational "maturity," and co-design tailored roadmaps and resource kit for building functional nodes. CANDLE positions itself as data infrastructure in the middle of an ecosystem, serving five key user groups: researchers, patients, clinicians, research infrastructures and policymakers. It offers infrastructure solutions, training and skills guidance, support for AI users, and help with policy dialogue to ensure nodes are aligned with Commission priorities and EHDS. Although only in month 5, CANDLE has already held joint workshops, launched early analyses and been invited into several external use cases and initiatives, showing that its national-node concept is already gaining strength across Europe.

6.3 EUCAIM

This project was presented as a flagship Cancer Mission project for building a pan-European federated infrastructure for cancer imaging data. Its goal is to overcome the fragmentation and short-lived nature of previous AI for Health Imaging projects (such as CHAIMELEON, EuCanImage, INCISIVE, ProCancer-I, etc.) and create a sustainable atlas of cancer imaging in Europe that supports AI-driven improvements in cancer diagnosis and treatment. The project (running from 2023–2026), started with 76 partners and now includes 95 partners from 17 countries plus more than 200 stakeholders. It is co-funded by the European Commission and coordinated from La Fe Hospital in Valencia. A first version of the platform was released in 2024, and a data-holder onboarding began in 2025, with full federated operation and expansion planned by 2027. The core aims of the project are to address fragmentation by aggregating up to 60 million anonymised cancer images and at least 100.000 cases (KPI already met), making these FAIR, ethically compliant datasets accessible to clinicians, researchers and innovators and providing a secure federated data warehouse for observational studies and the development/benchmarking of AI tools for precision oncology.

The architecture offers two options for data holders: (1) a central hub model, in which datasets are transferred to one of two reference nodes (UPV Valencia or Health-RI in the Netherlands) or (2) a federated local node model in which institutions keep data on-premises, set up as a local node, supported technically by EUCAIM. Depending on whether they transfer or federate, organisations sign a data transfer agreement or data sharing agreement, with the project's legal team guiding all ethical and legal requirements (GDPR, EHDS alignment, DPO reports, ethics approvals, etc.). Datasets must follow FAIR principles and specific technical standards such as DICOM, a common data model for clinical data, DCAT-AP for metadata and a dedicated EUCAIM hyperontology. Around 80 tools are

available for de-identification, harmonisation, annotation, FAIRness assessment, data quality and curation across the full data preparation pipeline.

To ensure sustainability beyond the project, EUCAIM is being transformed into an EDIC (European Digital Infrastructure Consortium), with Spain as host country and at least 13 Member States already supporting the initiative. Formal EDIC approval is expected around late 2026, securing long-term operation. The current platform status includes a dashboard with a public catalogue, documentation, helpdesk and training (Moodle, videos, handbooks). Registered users can explore datasets via federated queries (e.g. specific tumour type, age, Gleason score), submit access requests through a Negotiator tool to the EUCAIM Access Committee, which checks alignment with each data holder's conditions (e.g. non-commercial use). Data are then prepared and made available in secure environments without local download. EUCAIM also categorises datasets and data warehouses into three maturity tiers (1–3) via questionnaires and supports data holders to progress towards full compliance with EUCAIM and EHDS standards. Multiple support teams (engagement, technical, training, FAIR implementation and data population monitoring) help stakeholders become data holders, data users, software providers or associated organisations, with benefits including regulatory readiness, ethical-legal support, visibility and enhanced collaboration.

7 Conclusions

The third annual meeting of the "Understanding (Risk Factors & Determinants)" cluster confirmed substantial progress towards the Mission on Cancer objective of clarifying how risk factors and health determinants shape cancer onset and progression. All five projects are now generating rich multimodal datasets (clinical, imaging, omics, exposome, sensor data) and have converged on shared FAIR data principles, common standards (OMOP/FHIR, DICOM, EGA, ARGO) and secure analysis environments, laying the groundwork for sustainable cross-project data reuse.

Scientifically, the cluster has moved from descriptive associations to increasingly mechanistic and clinically actionable insights. Cross-comparisons link environmental exposures to mutational signatures and functional models, identify robust molecular features of progression and refine risk stratification and early-detection tools across very different tumour types. Parallel work on digital pathology, AI-ready infrastructures and federated or privacy-preserving analytics is turning these findings into practical pipelines that can be replicated beyond individual projects.

Equally important, the cluster has embedded implementation, inequalities and citizen engagement into its agenda. Joint work on health-policy recommendations, social-lab methodologies, shared decision-making and patient-generated evidence demonstrates a commitment to ensuring that scientific advances translate into fairer, more responsive cancer care. The close dialogue with emerging European data infrastructures (UNCAN-Connect, CANDLE, EUCAIM) positions the cluster's outputs to feed directly into long-term, federated ecosystems for cancer data and AI. Overall, the meeting showed a maturing, highly collaborative cluster that is delivering both cutting-edge science and concrete pathways for impact on prevention, early diagnosis, policy and patient experience across Europe.

8 Annexes

8.1 Annual cluster meeting agenda

Funded by the European Union

EU MISSIONS CANCER

Understanding (Risk Factors & Determinants)

Cluster 3rd Annual Meeting

15th October 2025 in Barcelona (Spain)

AGENDA

Hosted by: MELCAYA project, coordinated by Fundació de Recerca Clínic Barcelona-Institut d'Investigacions Biomèdiques August Pi i Sunyer.

Place: Room 4.1 of the Faculty of Medicine, University of Barcelona (Google Maps, see logistical information)

Link to join online: https://us02web.zoom.us/meeting/register/c6uPawHxRZ26fokdMcwdJw

Wednesday 15th of October 2025

(09:20-9:30) Welcome and introductory remarks

Speaker: Susana Puig (MELCAYA)

Research & innovation sessions

(09:30-10:15) Sharing and agreeing on common practices for data management

Speakers: Adrián López (MELCAYA), Ali Farnudi (DISCERN), Makis Zoidakis (ELMUMY), Étienne Audureau (GENIAL) and Alba Garin (LUCIA)

(10:15-10:30) Update about common strategy on data management

Speaker: Anushka Sachdev (LUCIA)

(10:30-11:15) Cross-comparison of risk factors and molecular features

Speakers: Stephan Ossowski (MELCAYA), Marc Gunter (DISCERN), Makis Zoidakis (ELMUMY), Patricia de la Cruz (GENIAL) and Jonathan Sleeman (LUCIA)

(11:15-11:30) Coffee break

(11:30-12:15) Technology, tools, knowledge and best practices for data exploitation and computational modelling

Speakers: Roberta Gugliotta (MELCAYA), Vivian Viallon (DISCERN), George Spyrou (ELMUMY), Laura Zigutyte (GENIAL) and Alba Garin (LUCIA)

(12:15-13:00) Cross-comparison of risk stratification/early diagnosis tools

Speakers: Susana Puig (MELCAYA), Karl Smith-Byrne (DISCERN), Nestoras Karathanasis (ELMUMY), Pierre Nahon (GENIAL) and Baruh Polis (LUCIA)

(13:00-13:30) Sharing best practices on implementation of healthcare policies

Speakers: Laura Sampietro (MELCAYA), Olivier Exertier (DISCERN) and Ivett Jakab (LUCIA)

(13:30-14:30) Lunch break

(14:30-14:45) Update about addressing inequalities

Speaker: Veronika Vsetickova (GENIAL)

(14:45-15:00) Update about citizen engagement

Speaker: Olivier Exertier (DISCERN)

(15:00-17:00) European Federated Cancer Research Data Hub

Speakers: Kadi-Liis Veiman (UNCAN-Connect) and Lifang Liu (CANDLE)

(17:00-17:15) Coffee break

(17:15-18:00) European Cancer Imaging Initiative

Speaker: Silvia Flor (EUCAIM)

Final remarks

References

[1] https://cordis.europa.eu/project/id/101096312

[2] https://cordis.europa.eu/project/id/101096473

[3] https://cordis.europa.eu/project/id/101097094

[4] https://cordis.europa.eu/project/id/101096888

[5] https://cordis.europa.eu/project/id/101096667