Longitudinal clinical, physiological and molecular profiling of female metastatic cancer patients: protocol and feasibility of a multicenter high-definition oncology study

Leonardo D. Garma[1†], Sonia Pernas[2†], David Vicente Baz[3], Rosario García Campelo[4], Josefa Terrasa[5], Berta Nasarre[6], Ruth Vera García[7], Begoña Bermejo[8], Santiago González Santiago[9], Bartomeu Fullana[2], Mario Rodríguez[2], Manuel Fernandez Bruno[4], Cristina Reboredo Rendo[4], Antonia Perelló[5], Desirée Jiménez[1,6], Juan Antonio Guerra[6], Virginia Arrazubi[7], Susana de la Cruz[7], Juan Miguel Cejalvo[8], Cristina Tebar[8], Elena Nuñez Berrueco[10], Berta Hernández-Marín[11], Jacobo Rogado[11],Pablo M Olmos[12], Leire P. Arbaizar[12], Yolanda Madrid[13], Jon Sanz-Landaluze[13], Gustavo Moreno-Martín[13], Aída Morillas[1], Antonio Lopez-Alonso[1], Alfonso Rodríguez-Patón[10], Ramón Colomer[1,11,14,15] Antonio Artés[12,16,17,18], Miguel Quintela-Fandino[1*]

[1]Breast Cancer Clinical Research Unit, Centro Nacional de Investigaciones Oncologicas (CNIO), Madrid, Spain
[2]Institut Catala d'Oncologia (ICO)-Institut d'Investigació Biomèdica de Bellvitge
[3]Hospital Virgen de la Macarena, Sevilla, Spain
[4]Medical Oncology Service, University Hospital A Coruña (XXIAC-SERGAS), A Coruña, Spain
[5]Department of Medical Oncology, Hospital Son Espases, Palma de Mallorca
[6]Hospital Universitario de Fuenlabrada, Spain
[7]Department of Oncology, University Hospital of Navarra, Instituto de Investigación Sanitaria de Navarra, IdISNA, Navarra, Spain.
[8]Medical oncology, Hospìtal Clínico Universitario de Valencia, Biomedical Research Institute INCLIVA, Medicine Department, Universidad de Valencia
[9]Hospital San Pedro de Alcántara, Cáceres, Spain
[10]Departamento de Inteligencia Artificial. Universidad Politécnica de Madrid (UPM)
[11]Medical Oncology Division, Hospital Universitario La Princesa, Madrid, Spain
[12]Dept Signal Theory & Communications, Universidad Carlos III de Madrid, Spain
[13]Universidad Complutense de Madrid (ROR 02p0gd045). Facultad de Ciencias Químicas. Ciudad Universitaria, 28040 Madrid, España.
[14]Department of Medicine, Universidad Autónoma de Madrid (UAM), Madrid, Spain
[15]Chair of Personalised Precision Medicine, Universidad Autonoma de Madrid (UAM – Fundación Instituto Roche), Madrid, Spain
[16]Evidence-Based Behavior, Spain
[17]Instituto de Investigación Sanitaria Gregorio Marañón, Spain
[18]CIBERSAM, Spain

[†]Equal contribution
*Corresponding author: mquintela@cnio.es

Abstract

Background:
Despite significant advances in precision oncology, a large proportion of patients receiving genomically matched therapies fail to respond. Emerging evidence highlights that cancer

outcomes are influenced by multiple non-genetic factors, including microbiome composition, metabolic environment, physiological parameters, and lifestyle habits. High-Definition Oncology (HDO) aims to integrate multi-dimensional, longitudinal patient data—spanning clinical, molecular, physiological, and behavioral domains—to enable truly individualized cancer care.

Methods:

We launched a prospective, multicenter observational study (HDO study; NCT0659050628) to longitudinally profile 300 female patients with newly diagnosed metastatic breast, lung, or colorectal cancer. Eleven data modalities are collected, including tumor and germline genomics, germline epigenomics, gut microbiome, blood and stool metabolome and proteome, exposome analysis, continuous wearable-based physiological monitoring, digital footprint, medical imaging, and patient-reported outcomes. This manuscript provides a comprehensive description of the study design, data acquisition workflows, sample processing pipelines, quality control procedures, and analytical strategies.

Results:

We report feasibility data from the first 10% of accrual (n=30). Patients completed 100% of scheduled clinical visits, 97.4% of planned plasma collections, 80.7% of stool samples, and 100% of tumor biopsies. Physiological monitoring captured activity, heart rate, sleep, and blood oxygen saturation data during 95.0%, 84.2%, 90.6%, and 70.7% of total patient-days, respectively. Biological samples met quality control criteria across all omic layers. Patient engagement with mobile apps for pain and emotion reporting exceeded 80%.

Conclusions:

The HDO study demonstrates the feasibility of deep, longitudinal, multi-modal patient profiling in metastatic cancer. This foundational protocol sets the stage for future analyses aiming to define disease trajectories, molecular taxonomies of outcomes, and patient digital twins.

Introduction

The discovery of the first oncogene-addiction driving mutations[1,2] displaced the “one size fits all” paradigm. The incorporation of genomic techniques has changed the standard of care in many cancer types: lung or colorectal cancer patients undergo gene-panel testing, including determinations of KRAS, BRAF, EGFR, HER2, ALK, ROS, RET, MSI, mutational burden and PD-L1 to allocate the treatment among several therapeutic options[1,3,4]. Breast cancer patients can be tested by gene expression platforms to determine the intrinsic subtype that will help decide therapy[5]. NGS panels increasingly often aim to match patient genomic profiles with therapeutic interventions[6,7]. However, the limitations of this approach indicate that additional factors are involved in clinical efficacy: in the former examples, about 25% of EGFR-mutant lung cancer patients or 40% of KRAS wild-type colorectal cancer patients do not respond to their matching therapies osimertinib[8] or cetuximab[9]. Similarly, immunogenic tumors are controlled by immunotherapy in less than 50% of the cases[10]. The predictive accuracy of genomic platforms is imperfect in hormone-positive cancers[11], malignancy in which also the most powerful therapeutic approaches (endocrine therapy or CDK4/6 inhibitors) are administered with imperfect or no biomarkers at all[12,13]. A significant proportion of HER2-positive tumors do not respond to anti-HER2 agents[14], and in triple-negative breast cancer there is an alarming paucity of biomarkers.

Although genomics plays a key role in cancer care, there is evidence that features of different nature influence the development and clinical course of oncologic diseases. For example, the composition of gut microbiota influences the chance of developing several cancer types[15,16], and can modulate the probability of response to immune checkpoint inhibitors[17,18]. Diet and activity can influence tumor biology and treatment outcome[19,20], and medications of routine use (omeprazole, metformin, antidepressants) strongly modify the microbiota or have pharmacokinetic interactions with anticancer compounds[21]. A large fraction of the molecular species isolated in human plasma are the result of our environmental exposure and has shown a relevant impact in cancer outcomes[22,23]. Whether changes in physiologic parameters such as heart rate, activity, or sleep patterns are causally or consequentially linked to cancer outcomes, and whether they constitute actionable targets, is debatable; however, there is evidence that high heart rate variability predict better prognosis[24,25] in cancer patients; and the stability of circadian rhythms in colorectal cancer patients has been linked to the length of their survival[26]. Therefore, complementing genomic analysis with other molecular and clinical dimensions can provide a better understanding of the disease and its clinical course, helping to overcome the current limitations of cancer treatments.

A unique "Patient Data Universe" (PDU) influences or informs about cancer clinical trajectory and treatment suitability and efficacy for each unique patient. Current and evolving technologies (wearable devices, smartphones and mobile apps) and analytic techniques (machine learning, deep learning, natural language processing, or LLMs) allow capturing and harnessing more and better this information in order to guide truly individualized treatment decisions that can be holistic in nature (i.e., delivering care not only in the pharmacologic sphere, but also a holistic therapeutic plan including habits change, nutritional, exercise, social and/or cognitive recommendations). This "high-definition" view or cancer care is in stark contrast with current guideline-based, evidence-based recommendations, which are issued for an "average patient" that does not really exist. Such hypothetical patient behaves according to the average results observed in populations of variable numbers (from small trials to large epidemiological studies) and is not defined by any particular feature or characteristic (i.e., even if we selected a patient with the average age, gender, tumor size and stage, co-morbidities and toxic habits, and the top 2-3 mutations considered in a typical cancer trial, and gave them the tested trial medication, that would certainly not warrant tumor response and disease control). Within our vision of practicing what we have termed High-Definition Oncology (HDO), a concept spanning from the conceptual frame of high-definition medicine[27], we integrate remote and continuous physiological monitoring, patient-reported outcomes, digital footprints (including location tracking), emotional and clinical data, and multi-omics profiling. By applying advanced data integration and analytical methods, we aim to gain a deep, granular, and longitudinal understanding of the factors shaping a cancer patient's clinical course. Our ultimate goal is to optimize therapeutic strategies for each patient, enhance clinical decision-making by "augmenting" the clinician's capabilities, detect favorable or unfavorable trajectories, and generate early warnings when interventions are needed.

Conversely and as opposed to current practice, we aim for true personalized care—tailored to the unique biological, clinical, and lifestyle factors of each patient—guided by longitudinal, multiparametric data capture and continuous patient monitoring. To this end, we have launched the HDO observational study.

The HDO study aims to profile longitudinally and with high depth 300 female patients with treatment-naïve advanced lung, colorectal and breast cancers using a multi-dimensional approach encompassing 11 data modalities: electronic health records (EHR), medical imaging, smartphone and activity tracker sensor data (physiological variables: sleep, heart rate, blood oxygen saturation, steps, physical activity; and digital footprint: location tracking -time spent at home, unique locations, etc.- and smartphone use). patient reported outcomes (quality of life

questionnaires, adverse effects, diet adherence, emotional status, visual analogue pain scale), tumor genome (whole exome sequencing), germline genome (genotyping array), germline epigenome (microarray-based profiling), gut microbiome (shotgun metagenomics), blood plasma proteome, blood plasma metabolome, fecal metabolome and exposome (heavy metals concentrations in stool and blood samples).

This study focuses exclusively on women and includes metastatic cases of the most common cancer types in females: breast, lung, and colorectal cancer. Our goal is to uncover both cancer type-specific characteristics and shared molecular and clinical features across different cancers ('pan-cancer' features). By identifying cross-cutting molecular signatures and clinical patterns, we aim to define a broader framework for cancer behavior.

These pan-cancer features could reveal the untapped potential of variables (e.g. inflammatory states, sleep patterns or abundance of specific microbial genera) as tools for assessing patient status and improving prognostic accuracy. By restricting the cohort to women, the study aims to reduce variability associated with sex- and gender-based differences and enhance the precision of the findings.

The study follows patients from the diagnosis of metastatic disease until death, ensuring a comprehensive and longitudinal analysis of disease evolution. This long-term perspective is critical for the following study objectives: 1) defining disease trajectories and their key drivers, identifying patterns in how the cancer progresses under different therapeutic strategies; 2) developing a molecular taxonomy of clinical outcomes, classifying patients based on molecular and clinical characteristics that predict their prognosis and response to treatment; and 3) generating patient digital twins, leveraging AI and longitudinal data to create virtual models that can simulate disease progression and optimize individualized treatment decisions.

We define the disease trajectory based on the dynamic changes in both its clinical status and the patient's quality of life over time. This includes clinical transitions such as progressive disease, stable disease, partial recovery or total recovery, as well as the occurrence of toxicity events or comorbities. Additionally, it encompasses variations in well-being measured through ECOG performance status, patient-reported quality of life (QoL), and pain levels. Tracking these trajectories allows for a comprehensive understanding of each patient's condition, moving beyond the limitations of current clinical practices.

By molecular taxonomy of outcomes**,** we refer to the classification of patients based on distinct molecular profiles despite having clinically indistinguishable outcomes. While patients may share the same disease status or response to treatment, their underlying molecular characteristics (e.g. genetic mutations, epigenetic modifications, or metabolic profiles) can vary significantly. Identifying these molecular subgroups allows for a deeper understanding of disease mechanisms, potentially revealing new therapeutic targets and refining personalized care approaches.

Patient digital twins are virtual models built from the multi-modal data collected from patients. These dynamic simulations replicate an individual's disease trajectory, allowing us to predict outcomes, test therapeutic strategies, and identify early intervention opportunities. By leveraging advanced analytics and machine learning, patient digital twins are expected to enable a highly personalized approach to cancer care, optimizing treatment decisions and improving patient outcomes.

By integrating these elements, the study aims to move beyond conventional clinical classifications and develop a more precise, data-driven approach to cancer care.

This report provides a comprehensive overview of the study design and methodological framework employed in the HDO study. Given the scale and complexity of this study—particularly

with the planned depth of sampling and data integration—we also assess its feasibility based on preliminary findings from the first 10% of patient accrual (N=30). This includes an evaluation of patient engagement, adherence to data collection protocols, and the overall feasibility of capturing high-resolution, longitudinal datasets during their first year of enrollment in the study.

Methods

Study design

This is a prospective, observational study involving 9 Spanish hospitals (Institut Catala d'Oncologia (ICO)-Institut d'Investigació Biomèdica de Bellvitge; Hospital Virgen de la Macarena; University Hospital A Coruña (XXIAC-SERGAS); Hospital Son Espases; Hospital Universitario La Princesa; Hospital Universitario de Fuenlabrada; Hospital Universitario de Navarra; Hospital Clínico Universitario de Valencia; Hospital San Pedro Alcántara de Cáceres) and 3 research institutions (Centro Nacional de Investigaciones Oncologicas (CNIO); University Charles III of Madrid (UC3M) ; Polytechnical University of Madrid (UPM)) which aims to recruit a total of 300 female patients with advanced breast (BC; N=100), lung (LC; N=100) or colorectal (CRC; N=100) cancers. The first patient was enrolled in February 2023 and the recruitment and monitoring will continue until the last patient's demise. The study is coordinated by the Breast Cancer Clinical Research Unit (BCCRU) at CNIO, supported by two data analysis and artificial intelligence research groups at the UC3M and UPM and a Patient Advocacy Organization (CRIS Contra el Cancer). The protocol for this study was approved by the ethical committee of the Fuenlabrada Hospital (protocol HDM-21/11). The study was registered at www.clinicaltrials.gov as NCT06590506[28]. The project is conducted in accordance with the Declaration of Helsinki and Good Clinical Practice standards.

Accrual

The inclusion criteria for the study require participants to be female with a histologically confirmed diagnosis of locally advanced or metastatic breast cancer (BC), lung cancer (LC), or colorectal cancer (CRC) that is not curable by loco-regional treatment. Patients must be treatment-naïve for metastatic disease, have an ECOG performance status of less than 2 (0 or 1, i.e., bed-ridden patients are excluded), and be able to use both a physiological monitoring device (smartwatch) and a smartphone, including the study's designated applications. All patients are required to sign an informed consent form complying with the ICH E6R2 guidelines. For the BC cohort, eligibility is further restricted to hormone receptor-positive (HR+), HER2-negative cases that qualify for treatment with endocrine therapy and CDK4/6 inhibitors. In contrast, patients in the LC and CRC cohorts are accepted regardless of their specific treatment regimen. Previous treatments for the metastatic setting such as palliative analgesic radiotherapy for bone metastases or cranial radiotherapy—whether for symptomatic brain metastases or at the investigator's discretion—are permitted, as long as inclusion occurs after such symptoms are controlled.

In order to accrue patients that are representative of real-world practice, we have not excluded patients on the basis of co-morbidity, life expectancy or co-medications. In addition, patients enrolled in therapeutic clinical trials were candidates to the HDO study as long as they were not receiving a treatment blinded to the investigators.

Patient training visit

Besides the engagement with serial tumor, blood and fecal sampling, a key aspect for study success is the extent to which the patients are familiar with the wearable and tracking devices required for this trial. Upon enrollment, each patient is assigned a unique numerical identifier and registered in a monitoring platform called eB2 MindCare[29].

eB2 MindCare includes a patient-facing mobile application and a research dashboard for study coordinators and researchers. The application automatically collects a variety of data, including patients' emotions (inputted at will) and daily pain levels (prompted daily), and administers periodic health questionnaires (EORTC QLQ-30, GHQ28, PRO-CTCAE at baseline, month 1, and every 3 months thereafter, typically coinciding with hospital visits). Additionally, dietary data related to adherence to the Mediterranean diet is also collected periodically via the app. The MindCare application further gathers passive data such as phone usage patterns and activity derived from the smartphone's sensors. The necessary permissions are granted upon installation, allowing seamless and secure data integration directly with eB2 databases.

The monitoring via the eB2 MindCare app is complemented by physiological tracking using a wearable device (Vivosmart 4 activity tracker, Garmin, USA). Patients receive detailed instructions and documentation for operating the Garmin wearable and installing its companion app, Garmin Connect, on their smartphones. The Garmin wearable monitors physiological parameters continuously, including blood oxygen saturation during sleep periods. To minimize data gaps, patients are specifically instructed to recharge the device during daytime hours, ensuring continuous data collection overnight. Typically, the rechargeable battery lasts up to 5 days and requires less than 30 minutes for a full charge.

The Garmin Connect app transfers the wearable's physiological data to the eB2 MindCare platform, facilitating comprehensive physiological and behavioral monitoring. Study coordinators are instructed to promptly notify eB2 data managers if a patient experiences technical difficulties, changes or loses their phone, or uninstalls the app. This ensures quick resolution, preserving data integrity and study continuity.

Patients are informed of the study’s tentative schedule and their upcoming scheduled clinical visits.

Samples and data collection

The study aims to obtain clinical and molecular profiles from the patients at regular intervals from the diagnosis until disease progression or complete recovery. Patients are scheduled for clinical consultation and baseline sample and collection visit before the first planned treatment dose of their systemic treatment. Subsequent timepoints are 1 month after the first treatment dose, and every three months thereafter.

At the baseline visit, a tumor biopsy, blood sample, and stool sample are collected from the patient. Clinical data are then captured using a customized electronic case report form (CRF) designed specifically for the study, which includes 24 clinical fields, 18 demographic fields, and 96 bloodwork fields. These fields are populated by dedicated data entry staff, who gather information from the electronic health records (EHR), clinician input, and lab results from bloodwork and biochemistry. All CRF fields are completed during the baseline visit, while subsequent visits only involve updates to the clinical and bloodwork data.

Genomic profiles of the patients and their tumors, blood DNA methylation, gut microbiome, blood plasma proteome, blood plasma metabolome, and fecal metabolome are derived from the baseline samples. On subsequent visits, additional blood and stool samples are collected to generate further molecular data —blood DNA methylation, gut microbiome, blood plasma proteome, blood plasma metabolome, and fecal metabolome— (Figure 1).

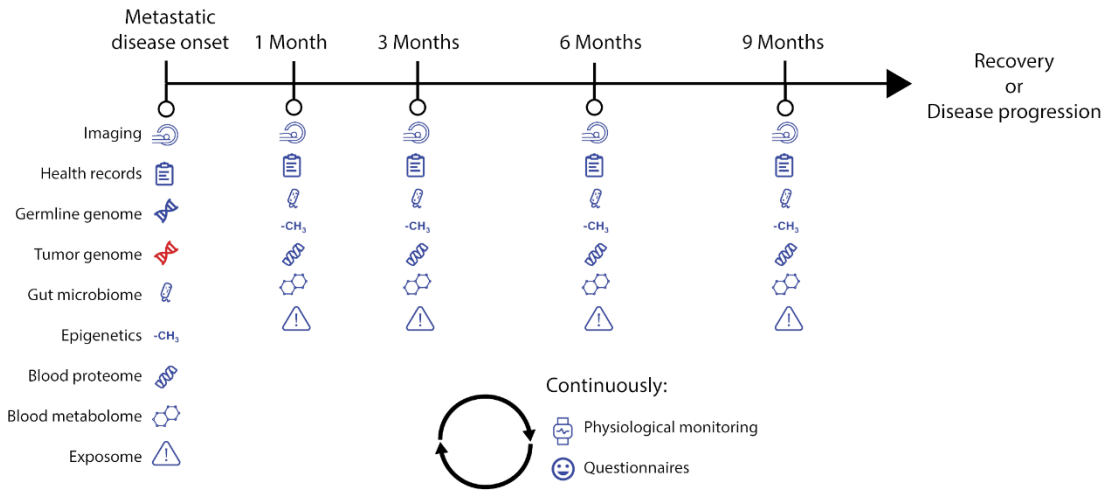


Figure 1. Schematic overview of the longitudinal data collection strategy in a clinical trial following metastatic disease onset.
Data types are collected at metastatic disease onset and at subsequent intervals of 1, 3, and every 3 months thereafter until recovery or disease progression. Data sources include imaging, health records, germline genome, tumor genome, gut microbiome, epigenetics, blood proteome, blood metabolome, and exposome. Germline and tumor genome data are collected only at baseline (metastatic disease onset), while other data types are collected repeatedly at designated time points. In parallel, physiological monitoring data and patient-reported outcomes are collected continuously. This multi-modal and longitudinal approach enables comprehensive characterization of disease trajectory and therapeutic outcomes.

Sample processing

Routine blood parameters are determined at the study site laboratories. A total of 4 4-ml EDTA tubes are collected for hematology analysis, while 1 tube of 3 ml venous blood is obtained for gasometry. A 4 ml coagulation-citrate tube is used for coagulation testing, and several 5 ml biochemistry tubes are collected for biochemical assays, including markers of renal, hepatic, and cardiovascular function. Further samples are collected at the study sites and sent to the central laboratory for additional analysis: two 10-ml BD Vacutainer CPT tubes are collected for peripheral blood mononuclear cell (PBMCs) extraction, and two 10-ml BD Vacutainer EDTA tubes are used for plasma collection. These samples are aliquoted and stored in 16 cryotubes (1.5 ml each) for shipping and long-term storage at -80ºC.

The procedure for stool sample collection is as follows: The patient collects the sample in a sterile feces collection container, filling it approximately halfway. The sample must be collected no more than 24 hours prior to processing and should be stored in a refrigerator at 4ºC by the patient. If the sample is collected at the hospital, it must be kept refrigerated until processing and should not be left at room temperature for more than 4 hours. Upon arrival at the hospital, three 5 ml Eppendorf tubes are labeled with the patient's information and the sample collection date.

Approximately 200 mg of stool is then transferred into each of three labeled Eppendorf tubes. The tubes are placed in a plastic bag and stored at -20ºC until they are sent to the central laboratory for further analysis.

The preferred materials for tumor sampling, listed in order of priority, are as follows: A) at least 3 frozen biopsy cylinders taken at the time of metastasis diagnosis, B) FFPE (formalin-fixed, paraffin-embedded) block, C) remaining block from the metastasis diagnosis not specifically collected for the project (assumed to be FFPE), D) at least 30 slides from the metastasis block with an H&E stained section, and E) archived tumor block from the initial diagnosis.
The sample collection protocol emphasizes that it is crucial to obtain at least three cores, as some cores may contain only coagulated tissue or fat, making them unsuitable for analysis. A pathology report confirming the presence of tissue in the cylinders must be provided, with the anonymized report including the patient's study code.
For labeling, each paraffin block or slide box is properly labeled with the sample details, including the acquisition date (whether archived or new) and the patient's study code. Frozen tumor samples in cryotubes are stored at -80ºC, while FFPE blocks or slides are stored at room temperature in a plastic bag or box, respectively, until they are shipped to the central laboratory.

Case Report Form

During the first clinical visit, the oncologists participating in the study collect demographic data, clinical history, disease evaluation, current treatments, pregnancy and childbirth history, surgical history, comorbidities, and allergies from both the EHR and the patient. Bloodwork results (96 variables) are also added to the case report form (CRF).

At each subsequent visit, the CRF is updated with new clinical and radiological evaluations, as well as the latest bloodwork results. Anonymized CRF records are stored in a secure, ad-hoc server (ShareCRF, Spain), where oncologists can add and/or modify records through interactive forms. This system ensures that patient data is continuously updated and securely stored for analysis throughout the study.

Medical Imaging

At each visit, patients' tumors are examined using the imaging technique most appropriate for their clinical features (CT, MRI, PET, PET-CT, radiography, or ultrasound). The radiologist's evaluation of the results is then added to the CRF. Anonymized raw images are collected as part of the study's data and securely stored in an ad-hoc server for future analysis.

Hematoxylin and eosin-stained whole-slide imaging

In parallel with the processing of tumor samples, hematoxylin and eosin-stained (H&E) tumor slides are scanned at the central laboratory. Whole slides are acquired using a slide scanner (AxioScan Z1, Zeiss) with a Plan-Apochromat 20x/0.8 M27 objective, and images are captured using Zen Blue Software (V3.1 Zeiss). The whole slide imaging (WSI) data is then stored on the central laboratory’s server in czi format.

Smartwatch Sensor Data

The patients are instructed to use the wearable devices throughout the duration of the study to collect sensor data. These devices log various parameters, including physical activity (activity type and duration), number of steps, stress levels, heart rate, and oxygen saturation. Through the

eB2 MindCare mobile application, patients can also register their emotions and pain levels at any time during the study. The raw data from the devices is securely logged to eB2 MindCare platform for further analysis.

Patient Reported Outcomes

A subjective evaluation of the patients' quality of life (QoL) and adverse effects is collected using the EORTC QLQ-C30[30], GHQ28[31] and PRO-CTCAE[32] questionnaires. The patients' adherence to the Mediterranean diet is examined using a 14-item questionnaire[33]. These questionnaires are filled in during each clinical visit and are also continuously available to the patients through the eB2 MindCare mobile app.
Partial and global summary scores are calculated for EORTC QLQ-C30, GH28 and diet questionnaires using their corresponding manuals[31,33,34], yielding 15, 5 and 1 features respectively.
The patients' pain levels are evaluated during each visit using a visual analogue scale (VAS), ranging from 0 to 10[35]. Patients can also add additional pain values at any time using the MindCare mobile application.
Patients' emotions are recorded through the MindCare app, which provides a pre-defined list of 20 emotions characterized by positive, negative, or neutral valence. All the patient-generated data is securely logged to the eB2 server and is then copied locally for further analysis.

Tumor Genome

The biopsy samples obtained during the first visit are used for whole exome sequencing of the tumors. DNA is extracted from the FFPE samples using the DNeasy® Blood & Tissue Kit (QIAGEN, USA). DNA concentration is measured using the Qubit 1X dsDNA HS Assay Kit (ThermoFisher Scientific, USA) with the Qubit 4 Fluorometer (ThermoFisher Scientific, USA). DNA integrity is assessed using the Genomic DNA Assay Kit (Revvity, USA) on the LabChip GX Touch (Revvity, USA).

Samples with a total DNA content above 120 ng are used to prepare sequencing libraries with the KAPA HyperPlus Kit (Hoffmann-La Roche, Switzerland). The libraries are then sequenced using the NovaSeq X Plus (Illumina, USA) sequencer at a loading concentration of 150 pM in a 10B Flowcell (2x150bp). Raw fastq files generated by the sequencing service are received and stored locally.

Variant calls are generated from the raw data using the Varca pipeline[36], which implements the GATK´s best practices[37]. Putative variants are then filtered, keeping only those passing the pipeline quality control filters, with a frequency lower than 1% in the general population (gnomAD[38] and 1000 genomes project[39]), affecting primary transcripts and labeled as modifiers, moderate or high impact variants by Ensembl's VEP[40].

Germline Genome

A genomic profile of the patients is created using the blood samples obtained during the first medical visit. DNA is extracted from the PBMCs using the DNeasy® Blood & Tissue Kit (QIAGEN, USA). The total DNA mass is quantified using the Qubit 1X dsDNA HS Assay Kit (ThermoFisher Scientific, USA) in the Qubit 4 Fluorometer (ThermoFisher Scientific, USA). Samples with less than 250 ng of DNA are discarded, while the remaining samples are used for genotyping with the

Infinium™ Global Screening Array-24 v3.0 BeadChip (Illumina, USA). Raw IDAT files generated by the genotyping service at the central laboratory are stored locally.
The IDAT files are processed using Illumina´s proprietary GenomeStudio v2.0.5 software[41] to obtain variant call files. The genotype calls are then filtered based on the distribution of their reported quality metrics (GenotypeCall Score > 0.2 and Call Frequency > 0.9).

Germline epigenome

The blood samples collected during each visit are used to profile the patients' epigenome. DNA is extracted from PBMCs using the same protocol as for genotyping. Sample quality is assessed using the same process as for the germline genome, with samples containing less than 600 ng of DNA being discarded. DNA methylation is then quantified using the EPICv2 methylation array (Illumina, USA). Raw IDAT files generated by human genotyping unit at the central laboratory are stored locally.
The raw data is processed using the SeSAME R library[42]. The intensity values are pre-processed using QCP parameters (quality-based masking probes of poor design, inference of channel for Infininum-I prbes, p-value masking using oob). Then methylation beta values are calculated for each probe.

Gut Microbiome: metagenomics

DNA extracted from the stool samples collected at each visit throughout the study is used to profile the microbiome communities in the patients' gut. DNA extraction is performed using QIAGEN QIAamp® PowerFecal® Pro DNA Kits (QIAGEN, USA). The DNA mass per sample is determined using the Qubit 1X dsDNA HS Assay Kit (ThermoFisher Scientific, USA) in the Qubit 4 Fluorometer (ThermoFisher Scientific, USA). Samples with less than 120 ng of DNA are discarded.
DNA samples (0.5 ng) are simultaneously fragmented and tagged with sequencing adapters using the Nextera XT DNA Library Preparation Kit (Illumina, USA). Adapter-ligated libraries are completed by PCR (13 cycles) and extracted using a double-sided SPRI size selection. The libraries are then applied to an Illumina flow cell for cluster generation and sequenced on an Illumina instrument following the manufacturer's recommendations. DNA libraries are sequenced using a NovaSeq X sequencer (Illumina, USA) with a target depth of 2 Gbases per sample. The raw fastq files generated by CNIO's sequencing service are stored locally.
The sequencing data is processed using Kraken 2[43] to detect and discard the host DNA and obtain microbial taxonomies. We then use Bracken[44] to estimate the relative abundances of the detected taxa. In parallel, HumaNn is used to perform a functional profiling of the microbiome. For this, microbial reads are aligned to cluster representatives in the UniRef90 database and subsequently mapped to gene ontology (GO) terms to identify functional pathways associated with the microbial communities.

Blood Plasma Proteome

Sample Preparation

Plasma samples are diluted 1:10 with lysis buffer containing Tris-HCl (pH 8.5) and 4% SDS. To monitor digestion efficiency, four non-human proteins—Concanavalin A, GroEL, Nitroreductase, and Protein A—are added as controls. The samples are heated at 90°C for 15 minutes, cooled, and centrifuged to remove debris. The supernatant is then treated with TEAB buffer (pH 8.5) and 10 mM TCEP to reduce disulfide bonds, followed by alkylation with 20 mM CAA at room temperature in the dark to prevent disulfide bond reformation.

Protein Cleanup and Digestion

Automated protein cleanup and purification are performed using the SP3 beads protocol on the King Fisher Apex system. Sequential washing steps are applied to remove contaminants before enzymatic digestion. Proteins are digested at a 1:20 protein-to-enzyme ratio using LysC and Trypsin, with overnight incubation at 37°C to ensure efficient peptide generation. Post-digestion, samples are acidified with formic acid (pH < 2), and iRT peptides are added at a final concentration of 50 nmol/µL for cross-sample normalization before loading onto EvoTips.

LC-MS/MS Analysis

Peptides are analyzed using an Evosep One system coupled to an Orbitrap Astral mass spectrometer (Thermo Fisher Scientific). Samples are loaded onto Evotips and separated using the standard “Whisper Zoom” gradient with a throughput of 60 samples per day (SPD) on a 15 cm, 75 µm i.d. Aurora Elite TS column packed with 1.7 µm C18 beads (IonOpticks). The column temperature is maintained at 55°C. Peptides are ionized at 1.5 kV using an EASY-Spray source with a capillary temperature of 280°C. The mass spectrometer is operated in data-independent acquisition (DIA) mode, with MS1 spectra acquired at a resolution of 240,000. Precursor spectra are interspersed with a 1-second period, and ion peptides are fragmented using higher-energy collisional dissociation (HCD) at a normalized collision energy of 27. The AGC target percentages are set to 500% for both Full MS (maximum IT of 3 ms) and DIA MS/MS (maximum IT of 3 ms). A 2 m/z precursor isolation window is used with optimized window placement between 380-980 m/z. Precursor scan range is 390-980 m/z, and fragment scan range is 150-2000 m/z. Basal pooled samples spiked with the same non-human protein controls serve as quality controls to assess digestion efficiency.

Data Processing and Analysis

Raw files are processed using DIA-NN 1.9.2 in library-free mode against the mouse protein database (UniProtKB/Swiss-Prot, one protein per gene, 21,990 sequences). Precursor m/z range is set from 379 to 981, with default settings applied elsewhere. Protein group intensities are obtained by summing precursor quantity values from the report.tsv file, filtered with Lib.Q.Value and Lib.PG.Q.Value < 0.01. Protein group intensity data are further processed in DAPAR[45]. Normalization is performed using LOESS, and missing values are imputed using the SLSA algorithm[46] for partially observed values and DetQuantile for values missing across entire conditions.

Plasma and fecal metabolome

Plasma and stool samples are shipped to General Metabolics (USA) for sample pre-processing and analysis.

Plasma samples pre-processing

The procedure begins by thawing the serum or plasma samples on ice for 30-60 minutes. The plasma or serum is then vortexed for approximately 15 seconds and centrifuged to remove any bubbles. A 20 µL aliquot of serum/plasma is transferred to a labeled 96-well plate, and 180 µL of 80% methanol in water is added to the sample. The plate is capped, placed on a shaker, and vigorously agitated for 15 seconds. The plate is incubated at 4°C for one hour. Following incubation, the samples are centrifuged at 4,000 rpm for 15 minutes to pellet the precipitate. The supernatant (100-400 µL) is then transferred to a shipping plate and stored at -20°C which is then shipped to General Metabolics (USA) for analysis.

Stool samples pre-processing

Between 30 and 80 mg from each sample are used, aiming to minimize the variation (ideally within a 2-3-fold range across all samples). The sample weight is measured by first weighing a pre-labeled 2 mL microtube, adding the frozen fecal material, and calculating the difference in mass. Homogenization is performed using a bead beater, after which the tubes are re-equilibrated at -20°C before proceeding with extraction. The extraction solvent, a mixture of 80% methanol and 20% water, is added to the homogenized fecal samples in two separate aliquots of 700 µL each. Each tube is vortexed for 15 seconds and briefly centrifuged to ensure proper mixing. After both extractions, the supernatant is pooled and transferred to labeled microtubes, and the samples are centrifuged again to remove any remaining solids. The final supernatant is collected and stored at -80°C. After completing the extraction, the samples are ready for shipping. Samples are then shipped overnight on dry ice.

Metabolic profiling

Metabolome profiles of the sample extracts are acquired using flow-injection mass spectrometry, following a method adapted from Fuhrer et al 2011. The instrumentation consists of an Agilent 6550 iFunnel LC-MS Q-TOF mass spectrometer in tandem with an MPS3 autosampler (Gerstel) and an Agilent 1260 Infinity II quaternary pump. The running buffer used is 60% isopropanol in water (v/v) buffered with 1 mM ammonium fluoride. Hexakis (1H, 1H, 3H-tetrafluoropropoxy)-phosphazene) (Agilent) and 3-amino-1-propanesulfonic acid (HOT) (Sigma Aldrich). The isocratic flow rate is set to 0.150 mL/min. The instrument is run in 4GHz High Resolution, negative ionization mode. Mass spectra between 50 and 1,000 m/z are collected in profile mode. 5 uL of each sample are injected twice, consecutively, within 0.96 minutes to serve as technical replicates. A pooled sample, from 92 different samples is injected periodically throughout each batch to be used as an anchor reference. Samples are acquired randomly within plates. The

results received from General Metabolics consist of raw intensity data, with assigned putative metabolite identities. These results are stored locally at CNIO.
The raw intensity values are normalized by log-transforming the data and then applying a probabilistic quotient normalization (PQN): first the median value of each pooled reference sample is computed, then each metabolite intensity on each sample is divided by the corresponding median from the reference, and then each sample is divided by the median of all its quotients.

Blood plasma and stool samples heavy metals determination

The total contents of metals implicated in breast cancer (Pb, , W, and As)[47] as well as elements with putative protective effects (Se, Zn, and Cu) are measured in stool and blood plasma samples using inductively coupled plasma mass spectrometry (ICP-MS) after acid digestion in a microwave oven (MSP 1000, CEM, Matheus, NC). Approximately 400 mg of blood plasma and 100 mg of lyophilized stool are weighed into Xpress vessel, and .750 mL of concentrated nitric acid (Scharlab, Spain) along with 0.250 mL of 30% (v/v) hydrogen peroxide (Dismadel, Spain) are added. The mixtures are digested at 150ºC for 20 min using a maximum power (800 W). After digestion, acidified solutions are diluted to a final volume of 5.0 mL with ultra-pure water, filtered through a 0.22 µm syringe nylon filters (Dismadel, Spain) and analyzed.

Elemental concentrations are determined using an Agilent 7700x ICP-MS instrument (Agilent Technologies, Tokyo, Japan) by the following isotopes: ^{206}Pb, ^{184}W, ^{75}As, ^{78}Se, ^{66}Zn and ^{63} Cu, under continuous acquisition mode. The experimental conditions are: RF power of 1550W, sample depth of 8mm, plasma gas flow rate of 15.0 L·min^{-1}, auxiliary gas flow rate of 0.90 L·min^{-1} and micromist nebulizer with a flow rate of 1.01 L·min^{-1}. For Se determination, a reaction cell with H_2 as reaction gas (4.1 L min^{-1}) is used. Quantification is performed by external calibration with the isotope mentioned above.

Privacy and data handling considerations

The project data encompasses -omic information (including genomics, transcriptomics, metagenomics, proteomics, and metabolomics), as well as physiological monitoring data, patient-reported outcomes, anonymized clinical records, laboratory test results, and medical imaging data.
The CNIO holds ultimate ownership of the data related to this project. To ensure compliance with applicable laws and regulations, CNIO has appointed a Data Protection Delegate in accordance with EU Law 2016/679, which is mandatory for institutions handling large quantities of data, particularly health-related data. Mr. Igor Pinedo, a lawyer contracted through DAC Beachcroft firm, serves as the Data Protection Delegate for CNIO, as required under Article 37.1.6 of the EU Data Protection laws. His role includes informing and advising CNIO staff regarding data-related matters, overseeing legal compliance, cooperating with authorities on data protection, and acting as a liaison between CNIO and regulatory bodies.

To ensure data security, the study includes a robust security plan aligned with the General Data Protection Regulation (GDPR) and European guidelines on data security for Big Data projects

(Supplementary File 1). This security plan is focused on principles of Privacy by Design and by Default, traceability, and data portability. To further safeguard the data, external auditors have been appointed to monitor any potential data breaches or incidents that may result in data escaping CNIO's control, while a specialized IT firm has been contracted to protect the data from cyberattacks. Any data breaches will be promptly reported to the relevant authorities and legal bodies.

The activities of the project, as well as the intended use of the data generated by the consortium, are reviewed by the Research Ethics Committees of both CNIO and the Carlos III Health Institute. EU Regulation 2016/679, which came into effect on 25 May 2018, governs the protection of individuals with regard to personal data processing and its free movement. The project adheres to the principles outlined in this regulation, ensuring full compliance with data protection laws throughout the study's lifecycle.

The study's Data Management Plan (DMP) follows the FAIR (Findable, Accessible, Interoperable, Reusable) principles (Supplementary File 1). This DMP ensures the study complies with ethical standards and legal requirements. It ensures that the principles of data anonymization are strictly enforced, so no data can be linked back to an individual's identity using reasonable methods. Additionally, the plan facilitates collaboration with the CNIO Biobank to establish a framework for accessing and co-governing multi-omic data. This data will be encoded and stored in standardized formats such as FASTQ, mxXML, and DICOM. The DMP also ensures that all data will be deposited in open-access repositories while safeguarding personal data protection and upholding ethical standards.

The eB2 MindCare app has the CE marking as a Class I medical device in accordance with Regulation (EU) 2017/745 (MDR).

Results

Recruitment and study adherence

Between February 16, 2023, and October 13, 2023, we recruited 30 patients. A 10% accrual rate was reached 239 days after the trial commenced. Initially, the trial began at a single hospital, with two additional study sites joining after two months, followed by three more at the four-month mark. Eight months after the kick-off, a total of nine hospitals were participating. As of March 2025, the current accrual rate stands at 5.9 patients per month.

These first 30 patients were 13 breast, 13 lung and 4 colorectal metastatic cancer cases, with a mean age of 59.8 years (median 58.5 years, range 28-76 years). At the 12-month mark (study report data cut-off date), these patients had been enrolled for a median of 217 days (range 29-365 days). One patient voluntarily withdrew consent (70 days) and 2 patients died after 29 and 181 days on trial, respectively (Figure 2).

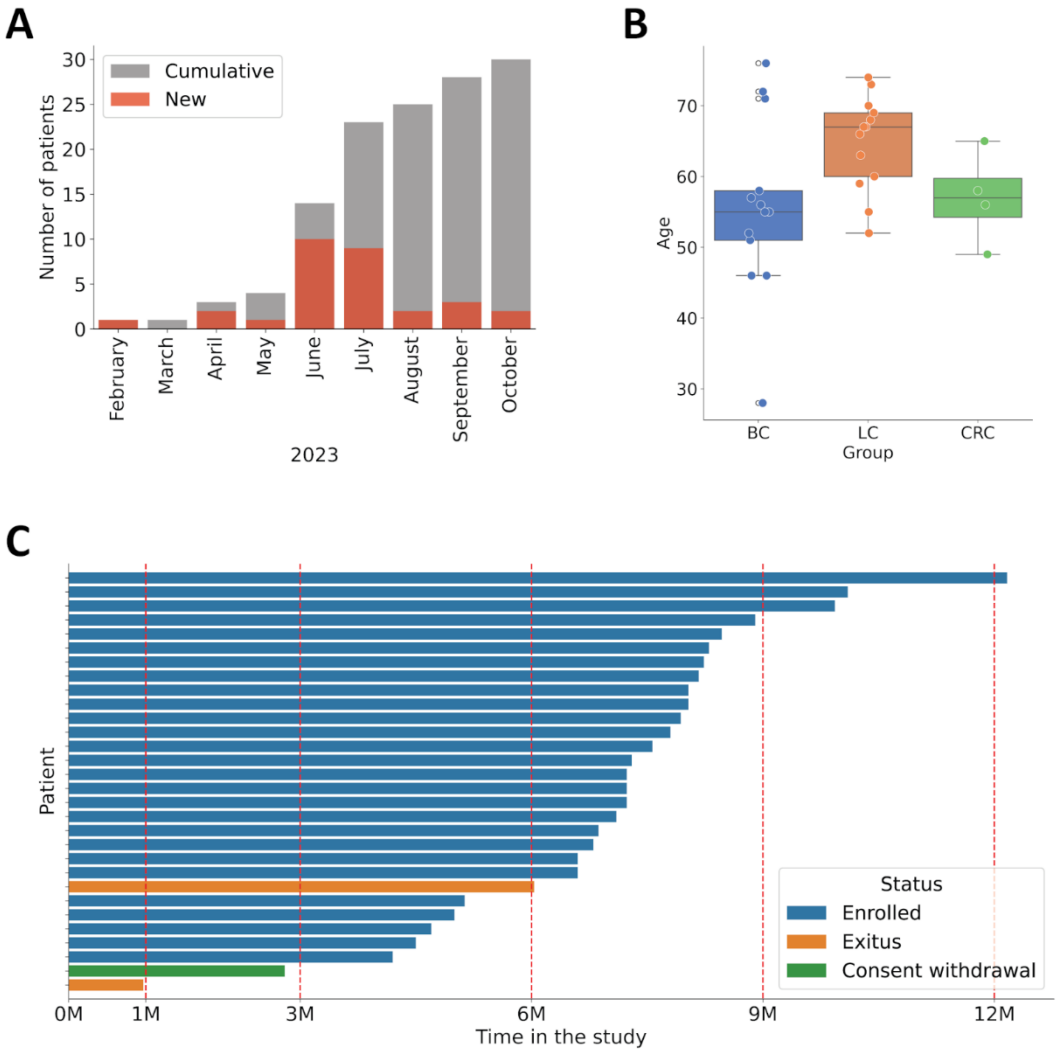


Figure 2. Overview of patient enrollment, demographics, and participation status in the study. (**A**) Monthly and cumulative patient enrollment over time from February to October 2023. Bars indicate the number of newly enrolled patients each month (red) and the cumulative total (gray). (**B**) Age distribution of participants by diagnostic group: breast cancer (BC), lung cancer (LC), and colorectal cancer (CRC). Boxplots show the median, interquartile range, and individual data points. (**C**) Patient retention and study participation timeline. Each horizontal bar represents an individual patient's duration in the study. Bars are color-coded by status: actively enrolled (blue), deceased (exitus, orange), or withdrawn consent (green). Red dashed lines indicate key time points at first month and every 3-month intervals.

Data and sampling completeness

All patients completed 100% of their due clinical visits. During these visits, 98.8% of the clinical laboratory tests, 100% of the medical history updates, 94.5% of social evaluation parameters and 94.05% of clinical evaluations (radiological evaluation, events -toxicty, treatment changes…-, ECOG) were completed. Out of all the planned samples, 97.37% of plasma, 80.7% of stool and 100% of the tumor samples were collected at the hospitals and shipped to the central laboratory (Figure 3A). Patients filled 81.14% of the planned questionnaires (Figure 3B); 28 patients reported pain values through the mobile app, with an average of 1.02 values per day (median 1, range 1-6). Similarly, out of the 30 patients, 28 imputed emotions through the eB2 application, with an average of 1.03 emotions per day (median 1, range 1-5) (Figure 3C).

Physiological monitoring captured activity (95.01%), heart rate (84.16%), sleep (90.64%), and blood oxygen saturation (70.73%) over total patient-days (Figure 3D). The observed variation in data capture rates reflects both technical and user-related factors. Activity data, based on accelerometry, demonstrated the highest capture rate due to its robustness and lower sensitivity to wear conditions. In contrast, heart rate and sleep data rely on continuous optical sensing, which may be affected by intermittent wear, improper device positioning, or motion artifacts. The notably lower capture rate for blood oxygen saturation may be attributed to its greater susceptibility to signal noise, the requirement for stable conditions (e.g., during sleep), and potential limitations in sensor performance. These discrepancies may also be influenced by device synchronization issues, user compliance, and periods of non-wear.

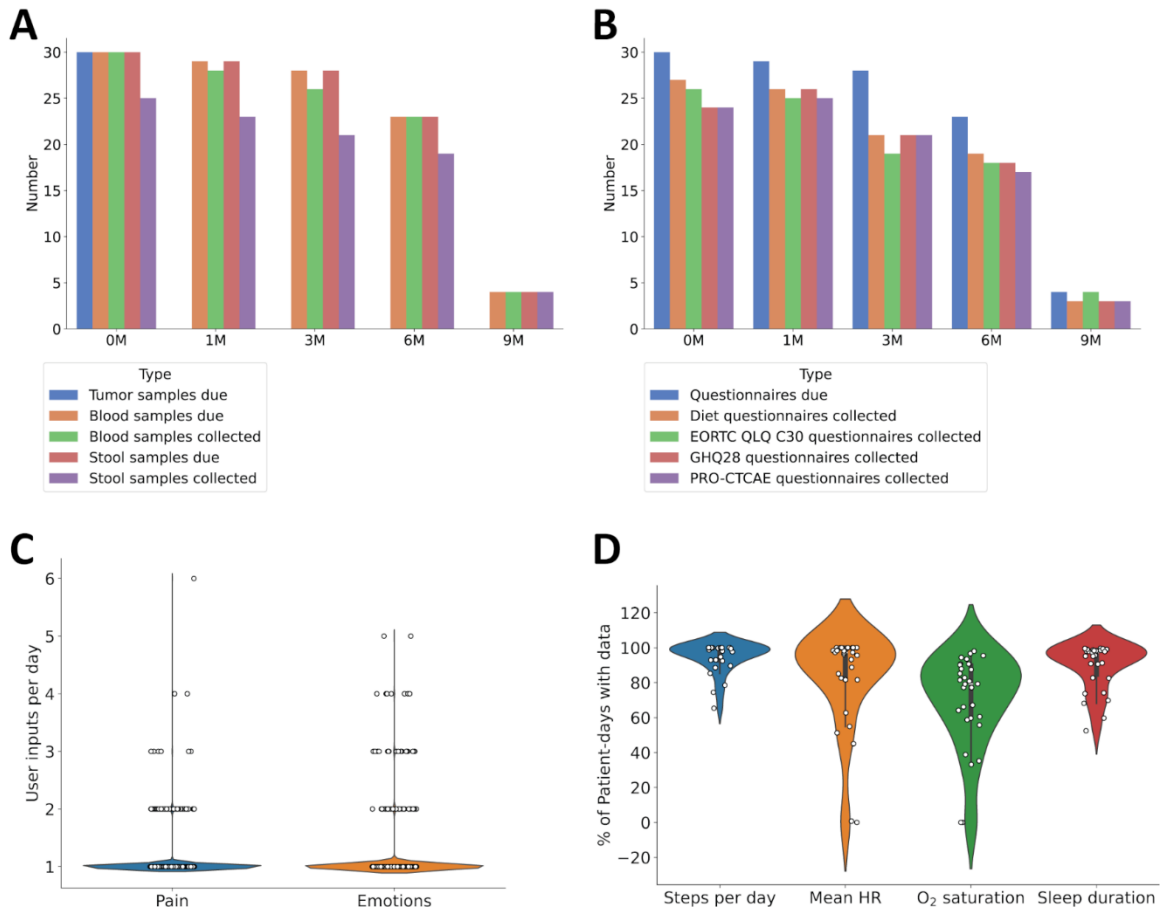


Figure 3. Data collection and digital health engagement.
(**A**) Longitudinal sample collection overview showing the number of tumor, blood, and stool samples due and collected at baseline (0M) and at 1, 3, 6, and 9 months. (**B**) Questionnaires completed over time, displaying the number of questionnaires due and collected, including diet, EORTC QLQ-C30, GHQ28, and PRO-CTCAE questionnaires at the same intervals as in **A**. (**C**) Distribution of daily user inputs related to pain and emotions. Each dot represents a patient on a specific date. (**D**) Distribution of the percentage of patient-days with available data for physiological monitoring metrics: steps per day, mean heart rate (HR), oxygen (O_2) saturation, and sleep duration. Each dot represents an individual patient.

Data Quality

All (100%) the biological samples collected at the central laboratory met the pre-established quality control criteria: a minimum concentration of 100 ng of DNA for tumor sample DNA WES, a minimum of 250 ng for genotyping and DNA methylation profiling, and a minimum of 100 ng for metagenomics sequencing. Tumor, blood and stool samples yielded 11.9 (0.29-70), 10.2 (0.4-59) and 9.0 (1.5-27) ug of DNA, respectively.
The total ion current, the median ion current and the coefficient of variation indicated that also all the plasma and stool samples were suitable for metabolomics profiling.

Data analysis considerations

The analysis of data collected in this study involves integrating various types of -omic data, clinical parameters, and patient-reported outcomes. As outlined in the Methods section, each -omic data modality will undergo specific processing methods, while statistical analysis will be employed to handle clinical data and patient-reported outcomes. Image processing will utilize advanced deep-learning techniques (e.g., HistoQC[48], CellVit[49], MONAI[50], 3D Slicer[51]).
Initially, each data modality will be analyzed independently to explore differences across clinical outcomes, patients, and conditions. However, integrating these diverse data modalities presents a significant challenge, as standardizing and harmonizing data for downstream analyses requires careful consideration of platform-specific biases and technical variance across datasets. Fortunately, the rise of multi-omic approaches, especially in single-cell profiling studies, has spurred the development of numerous multi-modal integration tools (e.g., MOFA+[52], mixomics[53] or SSGATE[54]) , which we will leverage to perform holistic analyses that incorporate all available data simultaneously.

To analyze time-series data, we will employ models such as recurrent neural network, long short-term memory networks, and transformer-based architectures for capturing temporal dependencies and latent trajectories in multi-omic and clinical datasets. Dimensionality reduction techniques, including principal component analysis, autoencoders, variational autoencoders (VAEs), will aid in extracting key patterns while mitigating the curse of dimensionality. Causal inference methodologies, such as Bayesian networks analysis, will be used to assess relationships between molecular features and clinical outcomes. Representation strategies, including co-occurrence networks and correlation-based approaches, will facilitate the exploration of interdependencies among omics layers and clinical phenotypes.

The PDTs will be developed as data-driven statistical models, leveraging deep learning to predict the future progression of disease trajectories based on integrated clinical, molecular, and behavioral profiles. Ultimately, these models will enable the simulation of potential treatment and intervention scenarios, allowing for the selection of the most effective therapeutic strategies.

Given the study's complexity and the longitudinal nature of data collection, missing data could present a significant challenge; missing values might arise from various sources such as participant dropouts, missed appointments, or issues with wearable device data collection[55]. Addressing missing data appropriately will be essential to ensure unbiased conclusions. Recent studies suggest that techniques like lightGBM[56] and probabilistic frameworks[57] could be the most suitable approaches to impute missing values on clinical data[58], may be well-suited for imputing missing clinical data, while molecular data may require specific imputation methods depending on the type of data.[59].

The extensive and multi-faceted nature of data collection—spanning genomics, proteomics, metabolomics, microbiome profiling, and physiological and clinical monitoring—means that each patient contributes a significant amount of detailed information. This comprehensive molecular profiling requires considerable resources, time, and expertise, making recruitment and profiling of a large patient cohort a considerable challenge. Furthermore, the complexity of the study necessitates advanced analytical methods and adequate statistical power to detect meaningful associations between molecular features and clinical outcomes. Small sample sizes in studies of this nature can reduce statistical power, making it harder to identify true associations and increasing the risk of type I or type II errors. Achieving an optimal balance between sample size and statistical validity is therefore crucial when designing such multifaceted studies.

The validation of putative markers and therapeutic targets remains a crucial challenge, as ensuring reproducibility across different cohorts and external datasets will require rigorous cross-validation strategies, bootstrapping methods, and independent replication studies to confirm the robustness of our findings.

Discussion

While genomic profiling has revolutionized targeted therapies, its predictive power is limited, as evidenced by the substantial proportion of patients who do not respond to genomically matched treatments[8–11,14]. This limitation, coupled with increasing evidence supporting the significant role of non-genetic factors in cancer progression and treatment resistance[17–26], underscores the necessity of incorporating these factors into cancer care.

The High-Definition Oncology (HDO) study represents an attempt to advance our capabilities to provide personalized cancer care through disease trajectories segmentation, molecular taxonomies of outcomes and patient digital twins.

This holistic approach aligns with the emerging paradigm of high-definition medicine, which emphasizes individualized patient profiles encompassing genetic, environmental, and lifestyle factors[27].

The robustness of the HDO study hinges on the quality and integrity of the collected data. Implementing standardized protocols for data collection, processing, and storage is paramount to ensure accurate and consistent data capture across all modalities, thereby enhancing the reliability of longitudinal analyses.

Preliminary findings from the first 10% of patient accrual (N=30) indicate that the study is progressing steadily, with high patient engagement and adherence to data collection protocols. The completeness of both clinical data and sample collection is remarkably high across all modalities, ensuring a robust and comprehensive dataset for longitudinal analyses. These encouraging early results suggest that the HDO study is well-positioned to generate clinically meaningful insights and establish a new benchmark for holistic cancer research.

Our study's extensive use of diverse data sources, enabling deep, longitudinal patient profiling, represents a significant evolution from early high-definition medicine studies, which primarily focused on integrating specific techniques such as DNA sequencing and medical imaging[27].

In comparison with large-scale studies such as The Cancer Genome Atlas (TCGA), which characterized thousands of tumor samples across multiple cancer types[60], or the UK Biobank[61] and the "All of Us" Research Program[62], which involve hundreds of thousands of participants, our study focuses on depth over breadth. Large-scale studies like TCGA prioritize broad genomic and clinical data collection to identify patterns across large populations, while here we delve deeply

into individual patient profiles, potentially uncovering nuanced interactions between clinical, molecular, and behavioral factors. Additionally, our study features a rich, ad-hoc case report form (CRF) that is specifically designed to comprehensively characterize the clinical status of patients, allowing for more detailed and tailored clinical data. In contrast, large-scale studies often rely on heterogeneous clinical reports that were not initially intended for these types of studies, leading to potential gaps or inconsistencies in the information available for each patient. This balance between sample size (N) and depth of data is crucial; smaller studies like ours can provide detailed insights into disease mechanisms and personalized medicine approaches, whereas larger studies offer statistical power to generalize findings across populations. In this regard, we expect our study to contribute to identifying the optimal trade-off between the number of samples and the depth of profiling, offering a framework for designing future studies that maximize both individual-level resolution and broader applicability.

Handling sensitive patient information necessitates stringent data privacy and security measures. The comprehensive data security and management plans of this study are designed to address these challenges effectively.
The HDO study's holistic approach positions it at the forefront of personalized oncology research. By addressing challenges related to data quality, privacy, and data analysis, the study aspires to uncover novel insights that transcend traditional one-size-fits-all models, ultimately paving the way for individualized cancer treatments that consider the whole patient rather than solely their tumor genetic profile. We anticipate that the HDO study will set a new standard for the design of future holistic research initiatives.

Funding

This study was performed at Centro Nacional de Investigaciones Oncológicas and funded by the Instituto de Salud Carlos III and the European Union (NextGenerationEU/PRTR), project number PMP22/00032. Leonardo Garma is beneficiary of a MSCA fellowship funded by the European Union (HORIZON-MSCA-2023-PF-01–01, grant agreement 101155328 HD-BRECA).

Competing Interests

References

1. Lynch, T. J. *et al.* Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib. *N. Engl. J. Med.* **350**, 2129–2139 (2004).
2. Heinrich, M. C. *et al.* Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. *J. Clin. Oncol.* **21**, 4342–4349 (2003).

3. Shaw, A. T. *et al.* Lorlatinib in advanced ROS1-positive non-small-cell lung cancer: a multicentre, open-label, single-arm, phase 1–2 trial. *Lancet Oncol.* **20**, 1691–1701 (2019).
4. André, T. *et al.* Pembrolizumab in microsatellite-instability–high advanced colorectal cancer. *N. Engl. J. Med.* **383**, 2207–2218 (2020).
5. Cardoso, F. *et al.* 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).
6. Massard, C. *et al.* High-throughput genomics and clinical outcome in hard-to-treat advanced cancers: Results of the MOSCATO 01 TrialGenomics to improve cancer outcome. *Cancer Discov.* **7**, 586–595 (2017).
7. Le Tourneau, C., Borcoman, E. & Kamal, M. Molecular profiling in precision medicine oncology. *Nat. Med.* **25**, 711–712 (2019).
8. Mok, T. S. *et al.* Osimertinib or platinum–pemetrexed in EGFR T790M–positive lung cancer. *N. Engl. J. Med.* **376**, 629–640 (2017).
9. Van Cutsem, E. *et al.* Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N. Engl. J. Med.* **360**, 1408–1417 (2009).
10. Lee, C.-H. *et al.* Lenvatinib plus pembrolizumab in patients with either treatment-naive or previously treated metastatic renal cell carcinoma (Study 111/KEYNOTE-146): a phase 1b/2 study. *Lancet Oncol.* **22**, 946–958 (2021).
11. Sparano, J. A. *et al.* Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *N. Engl. J. Med.* **379**, 111–121 (2018).
12. Mouridsen, H. *et al.* Superior efficacy of letrozole versus tamoxifen as first-line therapy for postmenopausal women with advanced breast cancer: results of a phase III study of the International Letrozole Breast Cancer Group. *J. Clin. Oncol.* **19**, 2596–2606 (2001).
13. Rugo, H. S. *et al.* Management of abemaciclib-associated adverse events in patients with hormone receptor-positive, human epidermal growth factor receptor 2-negative

advanced breast cancer: Safety analysis of MONARCH 2 and MONARCH 3. *The oncologist* **26**, e53–e65 (2021).

14. Cortés, J. *et al.* Trastuzumab deruxtecan versus trastuzumab emtansine for breast cancer. *N. Engl. J. Med.* **386**, 1143–1154 (2022).

15. Loke, Y. L., Chew, M. T., Ngeow, Y. F., Lim, W. W. D. & Peh, S. C. Colon Carcinogenesis: The Interplay Between Diet and Gut Microbiota. *Front. Cell. Infect. Microbiol.* **10**, (2020).

16. Ranjbar, M. *et al.* The dysbiosis signature of Fusobacterium nucleatum in colorectal cancer-cause or consequences? A systematic review. *Cancer Cell Int.* **21**, 194 (2021).

17. Baruch, E. N. *et al.* Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science* **371**, 602–609 (2021).

18. Matson, V. *et al.* The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science* **359**, 104–108 (2018).

19. Jones, L. W. *et al.* Exercise and Risk of Cardiovascular Events in Women With Nonmetastatic Breast Cancer. *J. Clin. Oncol.* **34**, 2743–2749 (2016).

20. Matthews, C. E. *et al.* Amount and Intensity of Leisure-Time Physical Activity and Lower Cancer Risk. *J. Clin. Oncol.* **38**, 686–697 (2020).

21. Liu, J. *et al.* Integration of epidemiologic, pharmacologic, genetic and gut microbiome data in a drug–metabolite atlas. *Nat. Med.* **26**, 110–117 (2020).

22. Engdahl, E. *et al.* Bisphenol A Inhibits the Transporter Function of the Blood-Brain Barrier by Directly Interacting with the ABC Transporter Breast Cancer Resistance Protein (BCRP). *Int. J. Mol. Sci.* **22**, 5534 (2021).

23. Lillo, M. A., Nichols, C., Seagroves, T. N., Miranda-Carboni, G. A. & Krum, S. A. Bisphenol A Induces Sox2 in ER+ Breast Cancer Stem-Like Cells. *Horm. Cancer* **8**, 90–99 (2017).

24. Kloter, E., Barrueto, K., Klein, S. D., Scholkmann, F. & Wolf, U. Heart Rate Variability as a Prognostic Factor for Cancer Survival – A Systematic Review. *Front. Physiol.* **9**, (2018).

25. Zhou, X. *et al.* Heart rate variability in the prediction of survival in patients with cancer: A systematic review and meta-analysis. *J. Psychosom. Res.* **89**, 20–25 (2016).

26. Innominato, P. F. *et al.* Prediction of overall survival through circadian rest-activity monitoring during chemotherapy for metastatic colorectal cancer. *Int. J. Cancer* **131**, 2684–2692 (2012).

27. Torkamani, A., Andersen, K. G., Steinhubl, S. R. & Topol, E. J. High-definition medicine. *Cell* **170**, 828–843 (2017).

28. High Definition Medicine for Solid Tumors Oncology (HDM-PMP22).

29. eB2 Mindcare.

30. Aaronson, N. K. *et al.* The European Organization for Research and Treatment of Cancer QLQ-C30: a quality-of-life instrument for use in international clinical trials in oncology. *JNCI J. Natl. Cancer Inst.* **85**, 365–376 (1993).

31. Goldberg, D. P. & Hillier, V. F. A scaled version of the General Health Questionnaire. *Psychol. Med.* **9**, 139–145 (1979).

32. Basch, E. *et al.* Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *J. Natl. Cancer Inst.* **106**, dju244 (2014).

33. Martínez-González, M. A. *et al.* A 14-item Mediterranean diet assessment tool and obesity indexes among high-risk subjects: the PREDIMED trial. (2012).

34. The EORTC QLQ-C30 scoring manual. (2001).

35. Scott, J. & Huskisson, E. Vertical or horizontal visual analogue scales. *Ann. Rheum. Dis.* **38**, 560 (1979).

36. Varca.

37. der Auwera, V. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinforma.* **43**, 1 (2013).

38. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

39. Siva, N. 1000 Genomes project. *Nat. Biotechnol.* **26**, 256–257 (2008).

40. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 1–14 (2016).

41. GenomeStudio 2.0.5.

42. Zhou, W., Triche Jr, T. J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* **46**, e123–e123 (2018).

43. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).

44. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

45. Wieczorek, S. *et al.* DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics. *Bioinformatics* **33**, 135–136 (2017).

46. Bø, T. H., Dysvik, B. & Jonassen, I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.* **32**, e34–e34 (2004).

47. Koual, M. *et al.* Environmental chemicals, breast cancer progression and drug resistance. *Environ. Health* **19**, 117 (2020).

48. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Inform.* **3**, 1–7 (2019).

49. Hörst, F. *et al.* Cellvit: Vision transformers for precise cell segmentation and classification. *Med. Image Anal.* **94**, 103143 (2024).

50. Cardoso, M. J. *et al.* Monai: An open-source framework for deep learning in healthcare. *ArXiv Prepr. ArXiv221102701* (2022).

51. Pieper, S., Halle, M. & Kikinis, R. 3D slicer. in *2004 2nd IEEE international symposium on biomedical imaging: nano to macro (IEEE Cat No. 04EX821)* 632–635 (IEEE, 2004).

52. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 1–17 (2020).

53. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).

54. Lv, T., Zhang, Y., Liu, J., Kang, Q. & Liu, L. Multi-omics integration for both single-cell and spatially resolved data based on dual-path graph attention auto-encoder. *Brief. Bioinform.* **25**, bbae450 (2024).

55. Austin, P. C., White, I. R., Lee, D. S. & van Buuren, S. Missing data in clinical research: a tutorial on multiple imputation. *Can. J. Cardiol.* **37**, 1322–1331 (2021).

56. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, (2017).

57. Van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

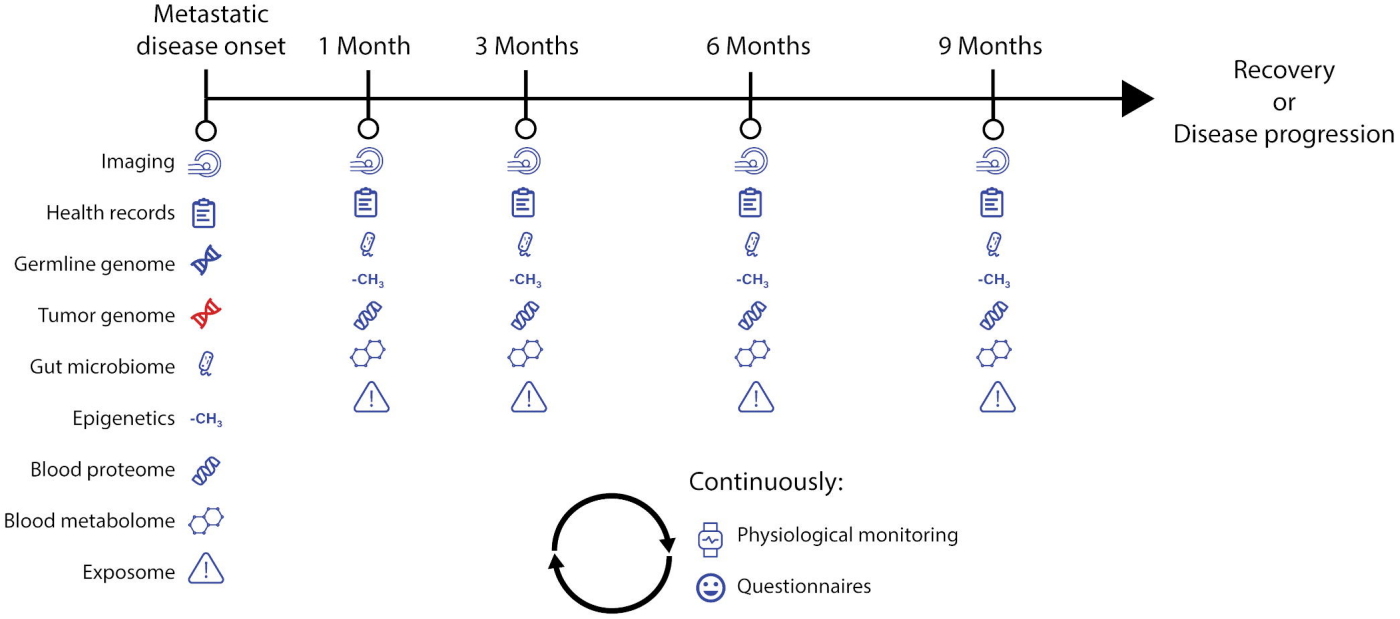
58. Luo, Y. Evaluating the state of the art in missing data imputation for clinical data. *Brief. Bioinform.* **23**, bbab489 (2022).

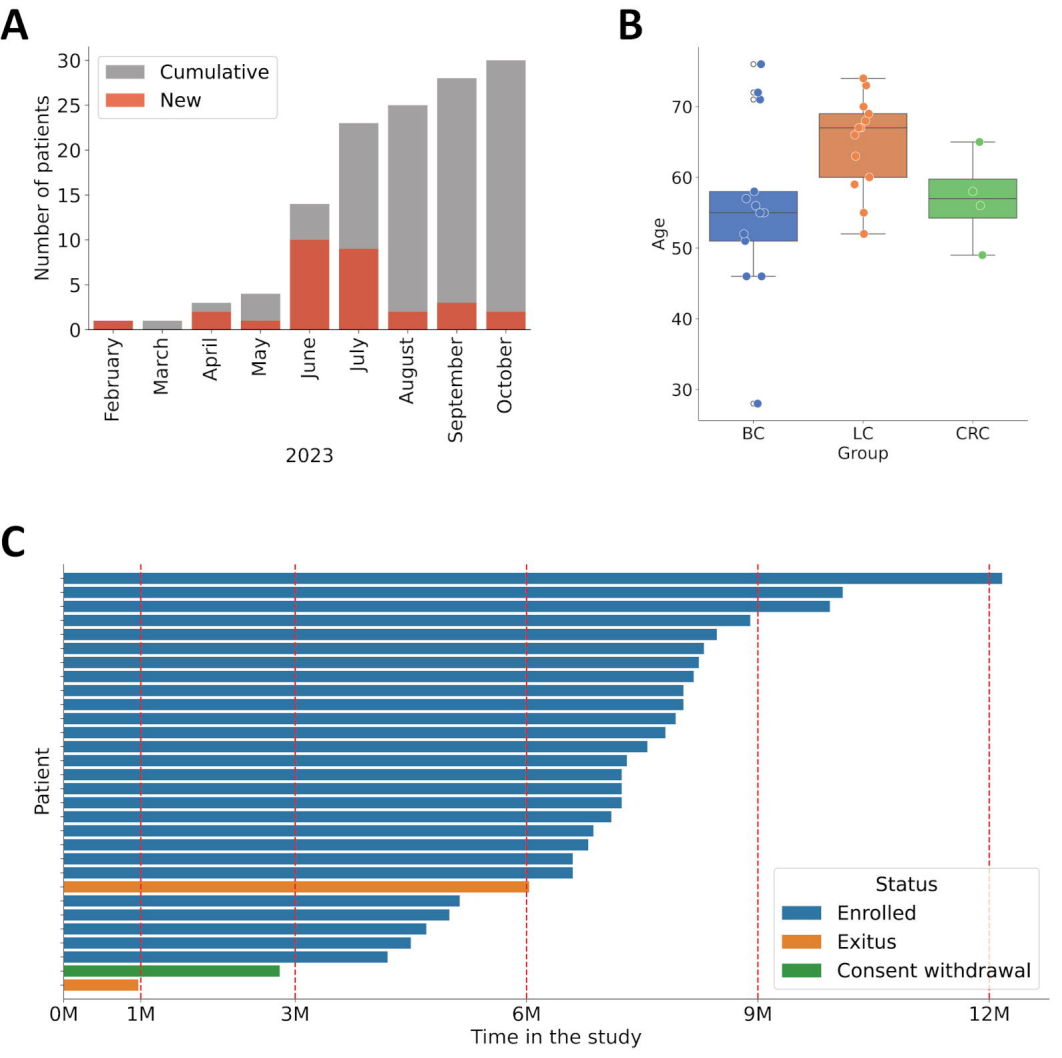
59. Flores, J. E. *et al.* Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front. Artif. Intell.* **6**, 1098308 (2023).

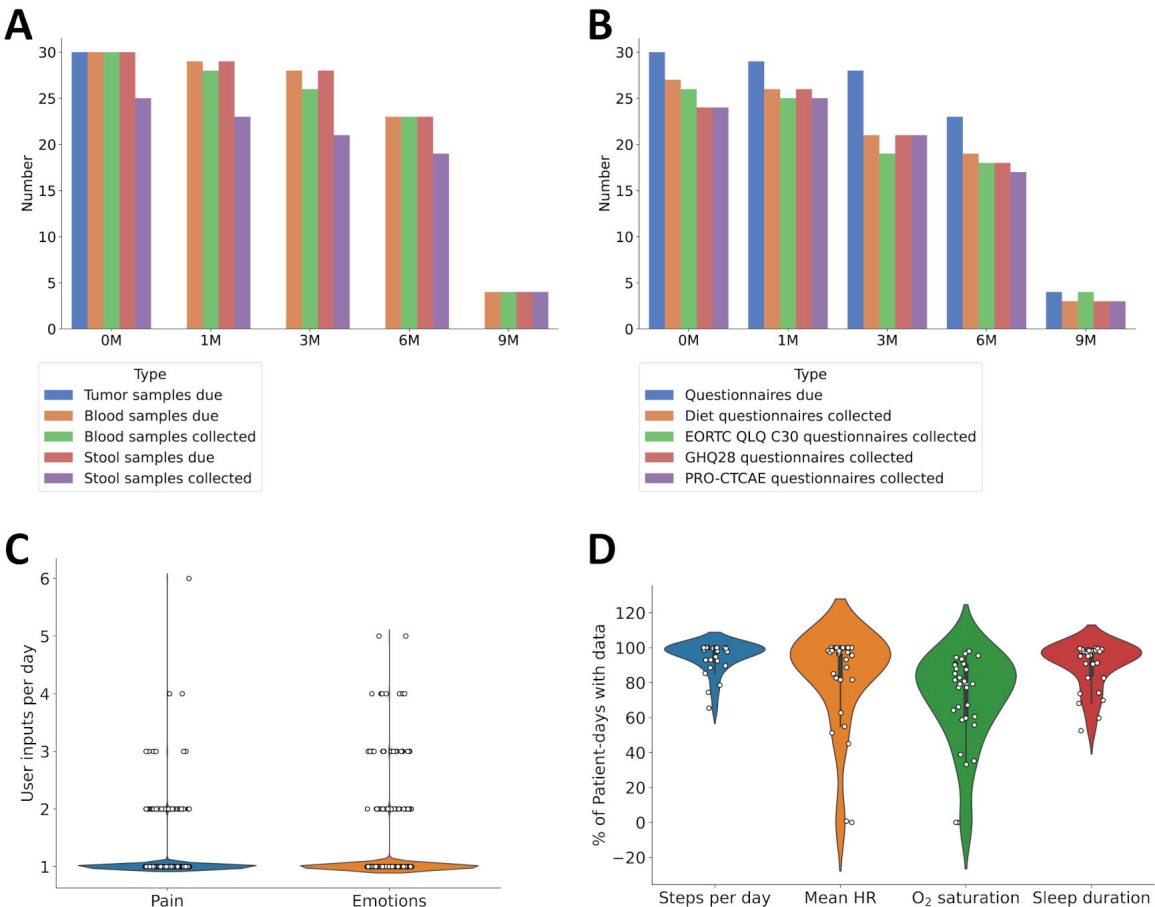
60. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

61. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

62. Investigators, A. of U. R. P. The “All of Us” research program. *N. Engl. J. Med.* **381**, 668–676 (2019).


Metastatic disease onset
1 Month
3 Months
6 Months
9 Months
Recovery or Disease progression
Imaging
Health records
Germline genome
Tumor genome
Gut microbiome
Epigenetics
-CH3
Blood proteome
Blood metabolome
Exposome
Continuously:
Physiological monitoring
Questionnaires


A
Cumulative
New
Number of patients
0
5
10
15
20
25
30
February
March
April
May
June
July
August
September
October
2023
B
Age
30
40
50
60
70
BC
LC
CRC
Group
C
Patient
Status
Enrolled
Exitus
Consent withdrawal
0M
1M
3M
6M
9M
12M
Time in the study


A
Number
0M
1M
3M
6M
9M
Type
Tumor samples due
Blood samples due
Blood samples collected
Stool samples due
Stool samples collected
B
Number
Type
Questionnaires due
Diet questionnaires collected
EORTC QLQ C30 questionnaires collected
GHQ28 questionnaires collected
PRO-CTCAE questionnaires collected
C
User inputs per day
Pain
Emotions
D
% of Patient-days with data
Steps per day
Mean HR
O_2 saturation
Sleep duration