MISTRAL DATA CATALOGUE

VERSION 1

NOVEMBER 6, 2026

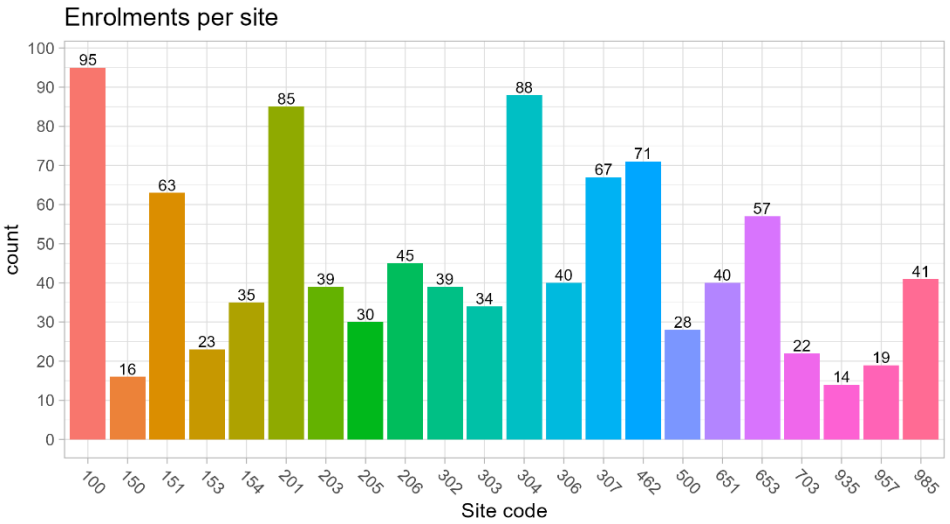
Index

1. Introduction....3
2. REDCap data....5
2.1 Participant journal data (historical and follow-up)....5
2.2 Visit data....6
3. Protocol deviations....16
4. Sample overview data....16
5. Querying process....18
6. Microbiome data....18
Quality Assurance Data and Metagenomic Pipeline Outputs....18
7. Biomarker data (ELISA cytokines)....24
IL-6....24
CRP....24
D-dimer....25
8. Metabolomics data....Error! Bookmark not defined.
9. Proteomics data....26
10. Resistome data....Error! Bookmark not defined.
11. Host genetics....26
12. Coding tables....26
Coding table 1: XXXXX EXAMPLE....Error! Bookmark not defined.
13. Appendices....Error! Bookmark not defined.
Appendix I....Error! Bookmark not defined.

14. References.....27

1. Introduction

The purpose of the MISTRAL data catalogue is to provide information on the data collected for the MISTRAL project Work Package 4 (WP4). MISTRAL WP4 was part of the MISTRAL EU Horizon 2020 funded project, centrally coordinated by IrsiCaixa (Barcelona, Spain), which aims to explore the impact of the human microbiome on clinical outcomes, specifically cardiovascular events, in people living with HIV (PLWH). WP4 is nested within the EuroSIDA study, an observational cohort of PLWH that has been collecting data from clinical sites across Europe since 1994. Data for MISTRAL WP4 have been collected from almost one thousand PLWH across 22 clinical sites in Europe and Israel (Figure 1). Collected data include clinical data, lifestyle questionnaire data, data from routine blood results, omics data from collected blood and faecal samples (metagenomics, proteomics, metabolomics, and potentially genomics data), as well as data on inflammatory biomarkers.



100	CHU Saint-Pierre (BE)	**304**	University Hospital Cologne (DE)
150	Hvidovre Hospital (DK)	**306**	University Hospital Bonn (DE)
151	Rigshospitalet (DK)	**307**	ICH Study Center (DE)
153	Odense Universitetshospital (DK)	**462**	Ospedale San Paulo (IT)
154	Aarhus universitetshospital, Skejby (DK)	**500**	Centre Hospitalier de Luxembourg (LU)
201	Mortimer Market Centre, Central and Northwest London NHS Foundation Trust (UK)	**651**	Hosp. Universitari Germans Trias i Pujol (ES)
203	Western General Hospital (UK)	**653**	Hospital de la Santa Creu i Sant Pau (ES)
205	Royal Sussex County Hospital (UK)	**703**	Hadassah Hospital (ISR)
206	Imperial College (UK)	**935**	Faculty Hospital Bulovka (CZ)
302	University Clinic Hamburg Eppendorf (DE)	**957**	Wojewodzki Szpital Specjalistyczny (PL)
303	J.W.Goethe University Hospital (DE)	**985**	University Hospital of Infectious Diseases (HR)

Figure 1. Enrolment into the MISTRAL study per clinical site

Two site visits approximately 10-24 months apart were planned for all participants in MISTRAL. During these site visits, the participant submits a stool sample and has blood drawn for both plasma and whole blood, as well as completes a lifestyle questionnaire. In addition to the participant site visits, all sites conduct annual follow-up by filling out an annual clinical follow-up form from the clinical data in participants' journals.

Participants enrolled in the study are either:

- EuroSIDA participants (participants already enrolled in EuroSIDA): the patient ID (PID) is the same as the PID used in EuroSIDA (7 digits, with first three digits indicating the site number).
  Enrolment: For EuroSIDA participants enrolment data is extracted from already collected EuroSIDA annual follow-up data (i.e., data from the participant's journal that are entered into EuroSIDA follow-up forms in REDCap) prior to or during the first MISTRAL site visit. As previously mentioned, annual follow-up is conducted with data being updated in REDCap by each site. For EuroSIDA participants this entails sites filling out the EuroSIDA follow-up form between October-December each year.
- Non-EuroSIDA participants (participants followed at EuroSIDA sites, but not enrolled in EuroSIDA): the PID follows the same structure as in EuroSIDA but start from 'xxx-8xxx'. E.g., the first non-EuroSIDA participant in MISTRAL from site 100 is 100-8001, then 100-8002, etc.
  Enrolment: For non-EuroSIDA participants, a REDCap MISTRAL enrolment form is completed by each site. The MISTRAL enrolment form contains baseline and historical participant journal data (similar to EuroSIDA follow-up forms). As for EuroSIDA participants, annual follow-up is conducted between October-December each year, with data being updated in REDCap by each site. For non-EuroSIDA participants this entails sites filling out MISTRAL follow-up forms, which are identical to EuroSIDA follow-up forms.

For more information of the project, please see the MISTRAL protocol and the MISTRAL questionnaire/visit forms, as well other study documents on the MISTRAL website.

2. REDCap data

2.1 Participant journal data (historical and follow-up)

The MISTRAL enrolment form (for non-EuroSIDA participants only), EuroSIDA follow-up forms, and MISTRAL follow-up forms contain baseline data of the participant, as well as historical and follow-up data. Historical data include information such as, e.g., date of first positive HIV test, first visit at the clinic, CD4 count measurement at ART initiation, and lowest CD4 count. As mentioned above, follow-up data are updated each year between October and December by the sites' clinical staff based on participants' journal data.

All participant journal data are collected in 17 tables, then cleaned, downloaded and exported to the following database

Table name	Table characteristics
tblART	Type of antiretroviral drug, start and stop dates, and reason for stopping
tblART_LAI	Data on injection dates for long-acting injectable antiretroviral therapy
tblBAS	Basic information such as demographics, basic clinical information and date of HIV diagnosis
tblCEP	Type and date of clinical events and procedures including serious non-AIDS conditions
tblDIS	Type and date of AIDS defining diseases and malignanices (CDC Category C conditions)
tblLAB	Type, date, value, and unit of laboratory tests. Part of routine clinical care, provided annually
tblLAB_BP	Date, diastolic and systolic values and unit of blood pressure measurements. Part of routine clinical care, provided annually
tblLAB_CD4	Date and value of CD4 measurements
tblLAB_HCV_RES	Information on hepatitis C virus (HCV) genotype and subtype
tblLAB_RES	Background information on HIV resistance tests, laboratory, library, kit, software and type of test
tblLAB_RNA	Date, value, detection limit and type of viral assay
tblLAB_VIRO	Test results for viro-/serological tests (hepatitis etc.)
tblLTFU	Data on death and drop-out
tblMED	Type, start and stop dates for other medication/treatments
tblPREG	Information about pregnancies started or completed since 1st of January 2016
tblVIS	Visit related information such as visit date and weight
tblVIS_SUBS	Information on tobacco, alcohol, and substance abuse

Table 1. Description of tables for all MISTRAL participant journal data.

Note: in tblBAS, the date of birth was changed to 15th of the month for all participants. For some of the participants, the birth month was unknown and for these, the month is listed as 06 (June).

2.2 Visit data

At enrolment each participant completes the first site visit (visit 1). Each participant returns 10-24 months after visit 1 to complete the second site visit (visit 2). Both visits are entered by the sites in REDCap

A total of 991 participants have completed visit 1 and 795 returned to complete visit 2. A number of participants missed visit 2 due to them withdrawing consent or being lost to follow-up.

Each visit contains the following forms in REDCap:

Patient Baseline Data (visit 1 only): Confirms the participant meets inclusion and exclusion criteria and provides information on consent to potentially human genomics analysis
Questionnaire: Unlike participant journal data, the MISTRAL questionnaire is completed by clinical staff together with the participant. The questionnaire contains information on lifestyle factors and data related to the samples collected during the visit (Table 2).
Samples: During each site visit, participants deliver a stool and blood is drawn (plasma and whole blood). More information can be found in Section 4. Each sample is split into several aliquots. The label IDs for these aliquot tubes are entered into REDCap in this section.

The REDCap codebook can be requested by contacting the MISTRAL coordinating centre at mistral.rigshospitalet@regionh.dk

Table 2. Fields in MISTRAL questionnaire.

Field name	Field Label	Format	Field values	Notes
		Gender & Date of Birth		
[bas_patient]	Patient ID	Integer		Unique participant ID number (Cohort Patient ID)

				Can be linked between tables. Participants have only one participant ID number.
[es_bas_gender]	Gender	Radio	1 2 3 4 5 9	Male Female Transgender male Transgender female Other Unknown
[es_bas_birth_d]	Date of birth	Date_ymd		
		Stool		
[vis_stool_samp_col_d_1]	Date of defecation for current stool sample	Date_ymd		
[vis_stool_samp_rec_d_1]	Stool sample receival date at clinic	Date_ymd		
[vis_stool_samp_t_1]	Stool consistency for current stool sample	Dropdown	Type_1 Type_2 Type_3 Type_4 Type_5 Type_6 Type_7	Type 1 Type 2 Type 3 Type 4 Type 5 Type 6 Type 7
[vis_stool_samp_time_1]	Time of defecation for current stool sample	Text (time)		
[vis_stool_aver_14d_t]	Average stool consistency based on the Bristol Stool	Dropdown	Type_1 Type_2 Type_3 Type_4 Type_5	Type 1 Type 2 Type 3 Type 4 Type 5

	Form Scale in the last 14 days		Type_6 = Type 6 Type_7 = Type 7
[vis_stool_aver_14d_fr]	Average stool frequency (times per day) within last 14 days	Radio	red_61 = Less than every 3rd day red_62 = Every 2nd to every 3rd day red_63 = 1-3 x daily red_64 = 4-5 x daily red_65 = > 5 x daily unk = Unknown
Diet, vitamins, probiotics and alcohol			
[vis_diet_t_omni]	Omnivore: both plant- and animal-based diet	Checkbox	omni = vis_diet_t_omni___omni
[vis_diet_t_pesc]	Pescatarian: plant- and fish/seafood-based diet	Checkbox	pesc = vis_diet_t_pesc___pesc
[vis_diet_t_vege]	Vegetarian: plant-based diet including dairy products and eggs	Checkbox	vege = vis_diet_t_vege___vege
[vis_diet_t_vega]	Vegan: solely plant-based diet (i.e. no animal products or by-products)	Checkbox	vega = vis_diet_t_vega___vega
[vis_diet_t_lact]	Lactose-free diet	Radio	0 = No 1 = Yes 9 = Unknown
[vis_diet_t_glut]	Gluten-free diet	Radio	0 = No 1 = Yes

			9	Unknown
[vis_diet_rmeat_1m_fr]	Red meat consumption (days per week)	Radio	0	0
			1_3	1-3
			4_5	4-5
			6_7	6-7
			9	Unknown
[vis_diet_fruitveg_1m_fr]	Average number of portions of fruit and/or vegetable consumption per day	Radio	0	0
			1_2	1-2
			3_5	3-5
			6_99	> 5
			9	Unknown
[vis_diet_dairy_1m_fr]	Average number of portions of dairy/milk consumption per day	Radio	0	0
			1_2	1-2
			3_5	3-5
			6_99	> 5
			9	Unknown
[vis_diet_fiber_1m_fr]	Average number of portions of fiber/whole grains consumption per day	Radio	0	0
			1_2	1-2
			3_5	3-5
			6_99	> 5
			9	Unknown
[med_id_multiv_1m_y]	Regular multi-vitamin tablet consumption in the last month	Radio	0	No
			1	Yes
			9	Unknown
[med_id_vitd_1m_y]	Regular vitamin D tablet consumption in the last month	Radio	0	No
			1	Yes
			9	Unknown
[med_id_vitb_1m_y]	Regular vitamin B tablet	Radio	0	No
			1	Yes

	consumption in the last month		9	Unknown	
[med_id_probio_fr]	Consumption of probiotics	Radio	daily_w	Daily or 4-6 times a week within the last week	
			daily_m	Daily or 4-6 times a week within the last month	
			1_3_w_w	1-3 times a week within the last week	
			1_3_w_m	1-3 times a week within the last month	
			none	None	
			unk	Unknown	
[med_id_probio_t_yog]	Probiotic type: Yogurt and yogurt-like products	Checkbox	yog	med_id_probio_t_yog___yog	Yogurt and yogurt-like products (Kefir, fermented milk, etc.)
[med_id_probio_t_komb]	Probiotic type: Kombucha	Checkbox	komb	med_id_probio_t_komb___komb	Kombucha
[med_id_probio_t_supp]	Probiotic type: Probiotic supplements	Checkbox	supp	med_id_probio_t_supp___supp	Probiotic supplements
[vis_subs_alco_1m_fr]	Average number of alcoholic beverages per week	Text (number, Min: 0, Max: 50)			
Treatment					

[med_id_antibio_y]	Antibiotics within the last three months (excluding topical treatment (e.g., applied to the skin))	Radio	0 No 1 Yes 9 Unknown	
[med_id_antibio_1]	Type of antibiotic (1)	Dropdown (autocomplete)		
[med_id_antibio_sd_1]	Start date Antibiotic (1)	Date_ymd		
[med_id_antibio_pv_1]	On antibiotic (1) at most recent visit	Radio	0 No 1 Yes 9 Unknown	
[med_id_antibio_ed_1]	Stop date Antibiotic (1)	Date_ymd		If treatment is ongoing, stop date is left blank. Up to 15 antibiotics can be listed.
[med_id_antifun_y]	Antifungal medication within the last three months (excluding topical treatment (e.g., applied to the skin))	Radio	0 No 1 Yes 9 Unknown	
[med_id_antifun_1]	Type of antifungal medication (1)	Dropdown (autocomplete)		
[med_id_antifun_sd_1]	Start date Antifungal (1)	Date_ymd		
[med_antifun_pv_1]	On antifungal (1) at most recent visit	Radio	0 No 1 Yes 9 Unknown	

[med_id_antifun_ed_1]	Stop date Antifungal (1)	Date_ymd		If treatment is ongoing, stop date is left blank. Up to 10 antifungal can be listed.
[med_id_antiinf_y]	Anti-inflammatory medication taken within three months from stool sample collection date	Radio	0 No 1 Yes 9 Unknown	
[med_id_antiinf_1]	Type of anti-inflammatory medication (1)	Dropdown (autocomplete)		
[med_id_antiinf_sd_1]	Start date Anti-inflammatory (1)	Date_ymd		
[med_antiinf_pv_1]	On anti-inflammatory (1) at most recent visit	Radio	0 No 1 Yes 9 Unknown	
[med_id_antiinf_ed_1]	Stop date Anti-inflammatory (1)	Date_ymd		If treatment is ongoing, stop date is left blank. Up to 10 anti-inflammatory can be listed.
[med_ppump_y]	Proton pump inhibitors within the last three months	Radio	0 No 1 Yes 9 Unknown	(omeprazole, pantoprazole, esomeprazole, rabeprazole, lansoprazole, or dexlansoprazole)
[med_bc_1m_y]	Oral hormonal birth control or post-menopausal hormone replacement	Radio	0 No or Not Relevant 1 Yes 9 Unknown	

	therapy within the last month			
[med_bc_sd_1]	Date for latest start of oral hormonal birth control or post-menopausal hormone replacement therapy	Date_ymd		
[med_bc_ed_1]	Date for latest stop of oral hormonal birth control or post-menopausal hormone replacement therapy	Date_ymd		If the treatment is ongoing, please leave stop date blank
[med_trans_1m_y]	Transgender hormone replacement within the last month	Radio	0 No or Not Relevant 1 Yes 9 Unknown	
[med_trans_1m_sd_1]	Date for latest start of transgender hormone replacement	Date_ymd		
[med_trans_1m_ed_1]	Date for latest stop of transgender hormone replacement	Date_ymd		If the treatment is ongoing, please leave stop date blank
[med_oest_1m_y]	Oestrogen treatment	Radio	0 No or Not Relevant 1 Yes 9 Unknown	

	within the last month			
[med_oest_1m_sd_1]	Date for latest start of oestrogen treatment	Date_ymd		
[med_oest_1m_ed_1]	Date for latest stop of oestrogen treatment	Date_ymd		If the treatment is ongoing, please leave stop date blank
[med_test_1m_y]	Testosterone treatment within the last month	Radio	0 No or Not Relevant 1 Yes 9 Unknown	
[med_test_1m_sd_1]	Date for latest start of testosterone treatment	Date_ymd		
[med_test_1m_ed_1]	Date for latest stop of testosterone treatment	Date_ymd		If the treatment is ongoing, please leave stop date blank
		Sex life		
[vis_anal_y]	Do you receive anal sex?	Radio	0 No 1 Yes 9 Unknown	
[vis_anal_past_3d_y]	Received anal sex in the last 3 days	Radio	0 No 1 Yes 9 Unknown	
		Diseases and surgery		
[dis_gastro_y]	Clinically diagnosed with specific gastrointestinal disease	Radio	0 No 1 Yes 9 Unknown	(ulcerative colitis, Crohn's, irritable bowel syndrome, chronic constipation, chronic diarrhoea, etc.)

[cep_gastro_5y_y]	Major surgery of the gastrointestinal tract in the past five years	Radio	0 No 1 Yes 9 Unknown	(exception of cholecystectomy and appendectomy)
[dis_bowelresec_y]	Had any major bowel resection (surgical removal of part of the colon or rectum) at any time	Radio	0 No 1 Yes 9 Unknown	
[dis_bowelresec_d_1]	Date of latest bowel resection	Date_ymd		
[dis_glut_y]	Gluten intolerance	Radio	0 No 1 Yes 9 Unknown	
[dis_glut_dia]	Diagnosis type of gluten intolerance	Radio	clin Diagnosed by a clinician self Self-reported unk Unknown	
[dis_glut_d_1]	Date of gluten intolerance diagnosis/report	Date_ymd		
[dis_lact_y]	Lactose intolerance	Radio	0 No 1 Yes 9 Unknown	
[dis_lact_dia]	Diagnosis type of lactose intolerance	Radio	clin Diagnosed by a clinician self Self-reported unk Unknown	
[dis_lact_d_1]	Date of lactose intolerance diagnosis/report	Date_ymd		
Lifestyle and Infant History				

[vis_travel_3m_y]	Traveled outside of current resident country within the last three months	Radio	0 No 1 Yes 9 Unknown
[vis_pet_1m_y]	Had a pet with fur or feathers living in household within the last month	Radio	0 No 1 Yes 9 Unknown
[vis_exercise_1m_fr]	Average exercise per week (in hours) the last month	Text (number, Min: 0, Max: 70)	
[newborn_breastfdpr_y]	Primarily breastfed as an infant	Radio	0 No 1 Yes 9 Unknown
[delivery_child_deliv_m]	Birth delivery method	Radio	5 Vaginally 12 Cesarean section 9 Unknown

3. Protocol deviations

Protocol deviations includes, for example, participants who did not deliver all types of samples (stool, plasma, and whole blood), participants who did not complete the lifestyle questionnaire, and participants who had their visit 2 outside of the defined timeframe. Reasons for these deviations are given, as well as how these might affect analysis.

4. Sample overview data

During each site visit, as mentioned, a stool sample is delivered, and blood is drawn. From the stool sample a total of 6 aliquots should be prepared and from the blood, 6 plasma, and 4 whole blood aliquots should be prepared, resulting in a total of 32 sample aliquots per

participant for both site visits (16 per site visit) (Table 3). The MISTRAL Study Laboratory Manual, with information on sample processing, can be found here.

Visit 1	Visit 2 (10-24 months after visit 1)
6 stool aliquots	6 stool aliquots
6 plasma aliquots	6 plasma aliquots
4 whole blood aliquots	4 whole blood aliquots
Total: 16 sample aliquots per participant	Total: 16 sample aliquots per participant

Table 3. Sample overview per participant per site visit

Note: Not all participants were able to deliver all requested aliquots (Figure 2).

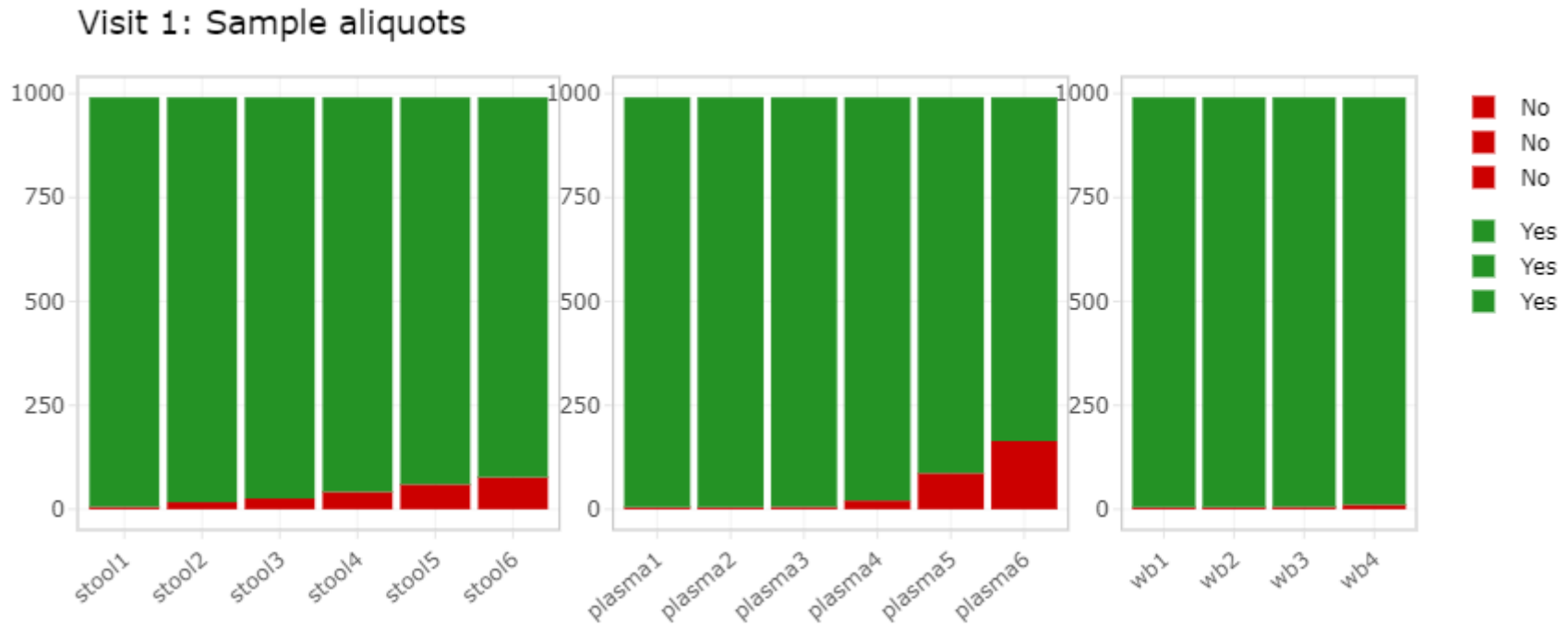


Figure 2. Overview of sample aliquots collected for visit 1.

5. Querying process

Throughout the data collection process, the REDCap data have been queried extensively. As part of the MISTRAL study, an R code was set up to create an HTML-based, interactive data quality report that can be used to monitor real-time incoming data across multiple REDCap forms and additional data sources. The report provides a general overview of the data and highlights discrepancies between the incoming data sources. Further, it checks for outliers and missingness, and monitors trends relevant for future data collection, all to make the quality assurance process easier and more systematic.

6. Microbiome data

One stool aliquot from each MISTRAL participants has been used to conduct shotgun metagenomic analysis. First, DNA has been extracted using DNeasy PowerSoil Kit (Qiagen), following manufacturer's instructions. The DNA product has been analysed for quality control purposes, using TapeStation D5000. Shotgun metagenomic libraries have been constructed from the extracted DNA using Nextera XT DNA Library Preparation Kit (Illumina). To verify the size of PCR enriched fragments, the template size distribution has been checked by running on an Agilent Technologies 2100 Bioanalyzer using a DNA 1000 chip. To achieve the highest quality of data, prepared libraries have been quantified using qPCR according to the Illumina qPCR Quantification Protocol Guide in order to create optimum cluster densities across every lane of every flow cell. Libraries have then been sequenced using Novaseq X Platforms (Illumina), by Macrogen, Korea, obtaining an output of 2Gb per sample.

Quality Assurance Data and Metagenomic Pipeline Outputs

Preprocessing

Sequencing data will be preprocessed using the in-house metagenomic pipeline, implemented in NGLess[1]. The quality control steps will include:

- Trimming reads at positions where the average Phred quality score over four consecutive bases falls below 20.
- Removing low-quality bases from the ends of reads using a quality score threshold of 20.

- Discarding reads shorter than 100 bases after trimming.
- Filtering out human reads by mapping to the HG19 human reference genome using BWA-MEM2[2], with a minimum alignment identity of 80% across 90 bases.

Microbial Profiling

- Taxonomic profiling will be performed using both mOTUs2[3] and MetaPhlAn4[4].
- Functional gene profiling will involve mapping reads to the Integrated Gene Catalog (IGC)[5] using BWA-MEM, retaining reads with a minimum alignment of 60 bases and ≥90% identity.
- Gene richness will be estimated by sampling 4,000,000 gene names from each sample and counting the number of unique genes. This calculation will only be applied to samples with more than 4 million reads.
- Microbial diversity will be assessed at both the mOTU and genus levels using the phyloseq R package [6]. Diversity metrics will be computed from rarefied mOTU abundance tables to control for variation in sequencing depth and reduce library size bias.
- Microbial evenness measures will be calculated at the mOTU level using custom R codes.

Antibiotic Resistance Gene (ARG) Profiling

- ARG detection will be performed by mapping reads to the CARD database[6] using KMA[7].
- Read counts will be normalized by gene length, followed by normalization to a 16S rRNA-based bacterial abundance estimate.
- For this, reads will be mapped to the SILVA bacterial SSU rRNA[8] database (arb-silva.de_2020-11-19_id917499_tax_silva_no_eu). 16S rRNA genes from the same genome will be grouped, read counts will be length-normalized, summed, and multiplied by 100 to generate normalization factors. The resulting values may exceed 100 if certain bacteria carry multiple copies of an ARG.arb-silva.de_2020-11-19_id917499_tax_silva_no_eu.fasta

Gut Metabolic Module (GMM) Profiling

- GMM profiling will be conducted by length-normalizing IGC gene count profiles and mapping them to KEGG ontology terms[9], based on the file IGC_catalog-v1.0.0.emapper.annotations-v2.tsv.
- These values will then be normalized to 16S rRNA abundance (as with ARGs), and GMM profiles will be generated using omixer-RPM[10].

Metagenomic Assembly and Metagenomic Assembled Genomes

- Metagenomic libraries will be individually assembled to generate metagenome-assembled genomes (MAGs).
- MAGs will be refined, reassembled, dereplicated, and assigned taxonomic classifications.
- High-quality MAGs will be selected based on quality assurance thresholds of ≥70% genome completeness and ≤10% contamination.
- Gene prediction will be performed on the high-quality MAGs.
- Reference protein sequences will be sourced from existing literature, and Hidden Markov model (HMM) searches will be used to identify homologous genes within the MAGs.
- Homologs with >99% sequence similarity to reference proteins will be used to construct phylogenetic trees.
- Clades clustering with the reference gene sequences will be identified as putative functional genes.
- Forward metagenomic reads will be mapped to the selected functional genes using bwa mem2.
- Hits with >99% sequence identity and alignment length >20 amino acids will be filtered.
- These filtered reads will be used to generate abundance tables showing functional gene read counts across all fecal samples.

Table 4. Microbiome and metagenomics data sources, types, formats, and processing levels used in the MISTRAL study.

Type	Data file	Description	Number of data points per sample
QC	Read_stats.csv	Read statistics	1) Number of total reads 2) Number of reads after QC 3) Reads after human removal
QC	fqstats.txt	Read statistics	Metadata for multiple FASTQ files: 1) encoding 2) file 3) gcContent 4) maxSeqLen 5) minSeqLen 6) nonATCGFraction 7) numBasepairs 8) numSeqs
QC	Stats_mapping_to_human.txt	Read statistics	1) Number of reads used for the mapping to human db 2) Number of reads aligned to human db 3) Number of reads aligned to human db uniquely.

QC	Stats_after_mapping_to_human.txt	Read statistics	Number of reads after the filtering of human reads
QC	mapping_to_igc.txt	Read statistics	1) Number of reads used for the mapping to IGC 2) Number of reads aligned 3) Number of reads aligned uniquely
Taxonomy	kingdom_count.txt	Number of reads mapping to each kingdom	5 Kingdoms
Taxonomy	kingdom_relative_abundance.txt	Relative abundance on the kingdom level	5 Kingdoms
Taxonomy	family_counts.txt	Number of reads mapping to each family	741 Families
Taxonomy	family_relative_abundance.txt	Relative abundance on the family level	741 Families
Taxonomy	order_counts.txt	Number of reads mapping to each order	396 orders
Taxonomy	order_relative_abundance.txt	Relative abundance on order level	396 orders
Taxonomy	phylum_counts.txt	Number of reads mapping to each phylum	146 phyla
Taxonomy	phylum_relative_abundance.txt	Relative abundance on phylum level	146 phyla
Taxonomy	class_counts.txt	Number of reads mapping to each class	235 classes
Taxonomy	class_relative_abundance.txt	Relative abundance on class level	235 classes
Taxonomy	Genus_count.txt	Number of reads mapping to each genus	2544 genera
Taxonomy	Genus_relative_abundance.txt	Relative abundance on genus level	2544 genera
Taxonomy	Motus_and_specl_count.csv	The number of raw reads mapping to each mOTU. mOTU stands for MG-based operational taxonomic units and corresponds roughly to species https://motu-tool.org/	14213 mOTUs/species

Taxonomy (external metagenomic module)	Motus_and_specl_rarified_count.csv	Number of reads mapping to each mOTU subsampled by the minimum library size. Library: Total number of reads that mapped to Bacterial and Archaeal domains. The subsampling code was constructed in R	14213 mOTUs/species
Taxonomy	Motus_and_specl_relative_abundance.txt	TSS: total sum scaling, also referred to as "relative abundance", the abundances were normalized by dividing the corresponding sample library size.	14213 mOTUs
Taxonomy	all_samples.all_kingdoms_all_levels.metaphlan4.tsv	Metaphlan results for all domains including all taxonomic sublevels	1610 taxonomical units
Taxonomy	all_samples.only_bacteria_all_levels.metaphlan4.tsv	Metaphlan results for the Bacterial Domain and all other taxonomic sublevels	1162 taxonomical units
Taxonomy	Metaphlan results on other levels	Custom R code is available to extract abundance data at all taxonomic levels	
Diversity	Species_diversity.csv	Diversity measures on mOTU level	8 measures Observed;Chao1;se.chao1;ACE;se.ACE;Shannon; Simpson;InvSimpson # Observed means the number of species detected. # All measures were calculated using the R phyloseq package
Diversity	Species_diversity_rarified.csv	Diversity measures based on the rarified taxa table "Motus_and_specl_rarified_count.csv" on mOTU level	8 measures Observed;Chao1;se.chao1;ACE;se.ACE;Shannon; Simpson;InvSimpson # Observed means the number of species detected. # All measures were calculated using the R phyloseq package
Diversity	Genus_diversity.csv	Diversity measures on genus level	The same measures as for species

Diversity	Gene_richness.csv	Normalized gene richness (Number of different IGC genes detected in each sample pr 4M reads)		1
Functional	ARO.normalized_profile.csv	Normalized ARG abundances	921 AROs	
Functional	Modules.tsv	Normalized GMM abundances	103 GMMs	
Functional	IGC_genes.profile.txt	Number of reads mapping to each IGC gene	9.879.874 Genes	
Functional	IGC_genes.normed_profile.txt	Number of reads mapping to each IGC gene normalized by gene length (Reads/per bp)	9.879.874 Genes	
Functional (external metagenomic module)	IGC_genes.profile_TSS.txt	The number of reads mapped to each IGC gene was normalized by dividing the corresponding sample read library size.	9.879.874 Genes	
Functional (external metagenomic module)	IGC_genes.profile_rarified.txt	Number of reads mapping to each IGC gene rarified based on read library size	9.879.874 Genes	
Mapping	GMM_mappings.txt	The mapping between the GMM identifier (MFXXXX) and the name of the module	103 GMMs	
Assembly	SampleID.contigs.fa	Contigs assembled from high-quality metagenomic reads		
Assembly	SampleID.scaffolds.fa	Scaffolds assembled from constructed contigs		
Genomes	SampleID.bin.X.fa	MAG assemblies meeting ≥70% completeness and ≤10% contamination thresholds		
Genomes	.gff or .tsv	Gene annotations for high-quality MAGs		

Genomes	.fasta (protein sequences), .nwk or .tree (Newick tree format)	Functional gene homologs and phylogenetic trees
Genomes	.tsv or .csv (tabular read count matrix)	Abundance tables with read counts per sample for each identified functional gene

7. Biomarker data (ELISA cytokines)

Inflammatory biomarkers interleukin-6 (IL-6), C-reactive protein (CRP) and D-dimer were measured in plasma samples from visit 1.

IL-6

Reported in pg/ml. Detection limit is <0,8 pg/ml

Table 5. IL-6 Measurement Classification.

Range (pg/ml)	Classification
< 0.8	Below the detection limit
0.8 – 2	Uncertain quantification
> 2	Clinically significant

CRP

Reported in mg/l. Detection limit is <1 mg/l.

D-dimer

Reported in ng/ml. Detection limit is <215 ng/ml.

8. Metabolomics data

Metabolomics profiling was conducted using the Metabolon Global Metabolomics Platform, which utilizes ultra-high performance liquid chromatography coupled with tandem mass spectrometry (UPLC-MS/MS). Instrumentation included a Waters ACQUITY UPLC system and a Thermo Scientific Q-Exactive Orbitrap mass spectrometer with heated electrospray ionization (HESI-II). Several chromatographic methods were applied to maximize metabolite coverage: reverse phase (RP) UPLC-MS/MS in both positive and negative ion modes, and hydrophilic interaction chromatography (HILIC) UPLC-MS/MS in negative ion mode.

A total of 468 plasma samples were analyzed in the current study (Project ID: IRCX-0102-24ML+), which were subsequently merged with a previous Metabolon project (IRCX-0101-24ML+), resulting in a combined dataset of 962 experimental samples. All samples were human plasma, stored at -80°C upon receipt, accessioned into the Metabolon Laboratory Information Management System (LIMS), and tracked throughout processing.

Table 6 Outline of the merged dataset

Project	Description	N
IRCX-0102-24ML+	*Current Study*	468
IRCX-0101-24ML+	Previous Study	494

The present dataset contains 1,685 biochemicals, including 1,396 compounds of known identity (named biochemicals) and 289 unnamed compounds. After merging with the previous dataset, the combined set consists of 1,321 biochemicals, of which 1,076 are named and 245 unnamed. Compound identification was based on three independent criteria: retention index (RI) within a narrow window, accurate mass match within ±10 ppm, and MS/MS spectral match scores.

Quality control procedures included pooled reference plasma samples (MTRX), a client matrix pool (CMTRX), process blanks (PRCS), and the addition of internal and recovery standards to each sample. Instrument variability was assessed using the relative standard deviation

(RSD) of internal standards, with a median RSD of 4%. Overall process variability was estimated using endogenous biochemicals detected across all CMTRX samples, yielding a median RSD of 8%.

Peak areas were quantified based on the area under the curve, and data normalization was applied for inter-day analytical consistency when needed. Specifically, a run-day block correction method was used, in which compound medians were registered to one (1.00) within each batch, and sample values were proportionally adjusted. This normalization ensures comparability across multi-day acquisition runs. The selection of a normalization method for downstream MISTRAL analyses is still under discussion.

All pre-informatics processing was managed through Metabolon's integrated system architecture, comprising a LIMS for sample tracking, proprietary software for data extraction and peak identification, and tools for quality control and compound verification. The infrastructure is supported by an Oracle 10.2.0.1 database, ensuring traceability and reproducibility throughout the analysis pipeline. Reported values are relative abundances.

9. Proteomics data

Coming soon

10. Resistome data

Coming soon

11. Host genetics

Only consent from some participants (about 94%). List of participants who have consented.

Coming soon.

12. Coding tables

The study uses HICDEP and ATC codes.

13. References

1. Coelho, L. P. *et al.* NG-meta-profiler: Fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome* **7**, 1–10 (2019).

2. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

3. Milanese, A. *et al.* Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat. Commun.* **10**, (2019).

4. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).

5. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

6. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2019).

7. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics* **19**, 307 (2018).

8. Yilmaz, P. *et al.* The SILVA and 'all-species Living Tree Project (LTP)' taxonomic frameworks. *Nucleic Acids Res.* **42**, 643–648 (2014).

9. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

10. Darzi, Y., Falony, G., Vieira-Silva, S. & Raes, J. Towards biome-specific analysis of meta-omics data. *ISME J.* **10**, 1025–1028 (2016).