The Milieu Intérieur study - an integrative approach for study of human immunological variance.

Stéphanie Thomas, Vincent Rouilly, Etienne Patin, Cécile Alanio, Annick Dubois, Cécile Delval, Louis-Guillaume Marquier, Nicolas Fauchoux, Seloua Sayegrih, Muriel Vray, et al.

Clinical Immunology (2015) **157**, 277–293

available at www.sciencedirect.com

Clinical Immunology

www.elsevier.com/locate/yclim

The Milieu Intérieur study — An integrative approach for study of human immunological variance☆

Stéphanie Thomas[a,b,c], Vincent Rouilly[a,d], Etienne Patin[e,f], Cécile Alanio[a,b,c], Annick Dubois[g], Cécile Delval[g], Louis-Guillaume Marquier[h], Nicolas Fauchoux[h], Seloua Sayegrih[h], Muriel Vray[i], Darragh Duffy[a,b,c], Lluis Quintana-Murci[e,f,*], Matthew L. Albert[a,b,c,j,], for The *Milieu Intérieur* Consortium[¶]**

The Milieu Intérieur Consortium is composed of the following team leaders: Laurent Abel[1], Andres Alcover, Philippe Bousso, Pierre Bruhns, Ana Cumano, Marc Daëron, Cécile Delval, Caroline Demangel, Ludovic Deriano, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Antonio Freitas, Odile Gelpi, Ivo Gomperts-Boneca, Serge Hercberg[2], Olivier Lantz[3], Claude Leclerc, Hugo Mouquet, Sandra Pellegrini,

Abbreviations: AbdoCM, abdominal circumference; ALP, alkaline phosphate; ALT, alanine aminotransferase; ANSM, Agence Nationale de sécurité du médicament et des produits de santé; AST, aspartate aminotransferase; BASO, basophil; β-HCG, beta-human chorionic gonadotropin; BILI, bilirubin; BMI, body mass index; BUN, blood urea nitrogen; Ca, calcium; Cl, chloride; CLT, clinical laboratory test; CMV, cytomegalovirus; CREAT, creatinine; CRF, case report form; CRO, clinical research organization; CRP, C-reactive protein; CVD, cardiovascular disease; df, degree of freedom; DYSBP1, diastolic blood pressure; EBV, Epstein–Barr virus; ECG, electrocardiogram; eCRF, electronic case report form; EOS, eosinophil; GFR, glomerular filtration rate; GGT, gamma-glutamyl transpeptidase; GLUC, glucose; HAS, human serum albumin; HBV, hepatitis B virus; hCG, human chorionic gonadotropin; HCO3, bicarbonate; HCT, hematocrit; HCV, hepatitis V virus; HDL, high density lipoprotein; HGB, hemoglobin; HIV, human immunodeficiency virus; HSA, human serum albumin; HTLV, human T cell lymphotropic virus; IgA, Immunoglobulin A; IgE, Immunoglobulin E; IgG, Immunoglobulin G; IgM, Immunoglobulin M; INSEE, Institut National de la statistique et des études économiques; IQR, interquatile range; K, potassium; LabEx, Laboratoire d'Excellence; LDL, low density lipoprotein; LYMPH, lymphocyte; MAMP, microbial-associated molecular pattern; MCH, mean corpuscular hemoglobin; MCV, mean corpuscular volume; mM, glucose level; MONO, monocyte; Na, sodium; NEUTR, neutrophil; P, phosphorus; PBMC, peripheral blood mononuclear cells; PCA, principal component analysis; PHOS, phosphorus; PLT, platelet; RBC, red blood cell; RT, room temperature; SNP, single nucleotide polymorphisms; SOPs, standard operating procedures; SYSBP1, systolic blood pressure; TCHOL, total cholesterol; TG, triglyceride; TPROT, total protein; TRIGLY, triglyceride; UA, urinalysis; UAC, uric acid; V0, visit 0; V1, visit 1; V2, visit 2; WBC, white blood cell.

☆ One sentence summary: This report presents the first demographic data from the *Milieu Intérieur* Consortium, which has established a 1000-person healthy population-based study, for assessing the genetic and environmental determinants of human immunologic variance.

* Correspondence to: L. Quintana-Murci, Unit of Human Evolutionary Genetics, CNRS URA3012/Institut Pasteur, 28, rue du Dr. Roux, 75724 Paris Cedex 15, France.

** Correspondence to: M.L. Albert, Unit of Dendritic Cell Immunobiology, Inserm U818/Institut Pasteur, 25, rue du Dr. Roux, 75724 Paris Cedex 15, France. Fax: +33 1 45 68 85 48.

E-mail addresses: quintana@pasteur.fr (L. Quintana-Murci), albertm@pasteur.fr (M.L. Albert).

¶ unless otherwise indicated, partners are located at Institut Pasteur, Paris.

http://dx.doi.org/10.1016/j.clim.2014.12.004

Stanislas Pol[4], Lars Rogge, Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Vassili Soumelis[3], Frédéric Tangy, Eric Tartour[5], Antoine Toubert[6], Marie-Noëlle Ungeheuer, Lluis Quintana-Murci[§], Matthew L. Albert[§]

[1] *Hôpital Necker, France*
[2] *Université Paris 13, France*
[3] *Institut Curie, France*
[4] *Hôpital Côchin, France*
[5] *Hôpital Européen George Pompidou, France*
[6] *Hôpital Saint-Louis, France*

[a] *Center for Human Immunology, Institut Pasteur, Paris, France*
[b] *Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, Paris, France*
[c] *INSERM U818, France*
[d] *Center for Bioinformatics, Institut Pasteur, Paris, France*
[e] *Laboratory of Human Evolutionary Genetics, Department of Genomes & Genetics, Institut Pasteur, Paris, France*
[f] *CNRS URA3012, France*
[g] *PIRC, Institut Pasteur, Paris, France*
[h] *Biotrial, Rennes, France*
[i] *Unit of Emerging Diseases Epidemiology, Institut Pasteur, Paris, France*
[j] *INSERM UMS20, France*

Received 17 July 2014; accepted with revision 1 December 2014
Available online 3 January 2015

KEYWORDS
Healthy donor;
Cohort design;
Immune phenotypes;
Baseline serologic data;
Metabolic syndrome;
CMV

Abstract The *Milieu Intérieur* Consortium has established a 1000-person healthy population-based study (stratified according to sex and age), creating an unparalleled opportunity for assessing the determinants of human immunologic variance. Herein, we define the criteria utilized for participant enrollment, and highlight the key data that were collected for correlative studies. In this report, we analyzed biological correlates of sex, age, smoking-habits, metabolic score and CMV infection. We characterized and identified unique risk factors among healthy donors, as compared to studies that have focused on the general population or disease cohorts. Finally, we highlight sex-bias in the thresholds used for metabolic score determination and recommend a deeper examination of current guidelines. In sum, our clinical design, standardized sample collection strategies, and epidemiological data analyses have established the foundation for defining variability within human immune responses.

1. Introduction

Susceptibility to infections, disease severity, and response to medical therapies or vaccines are highly variable from one individual to another. Medical practices and public health policies typically take a 'one size fits all' model for disease management and drug development. This approach ignores individual heterogeneity in immune responses that likely impacts the response to therapy or the efficiency and development of side effects secondary to vaccine or treatment administration. Due to the complexity of immune responses at the individual and population level, it has been challenging thus far to define the borders of a healthy immune system as well as the parameters (genetic, epigenetic, and environmental) that drive its naturally-occurring variability. In particular, such assessments require large sample sizes, consensus for defining "healthy", and standardized protocols for sample recruitment. In this context, the *Milieu Intérieur* Consortium initiated in September 2012 a cross-sectional healthy population-based study called *"Genetic & Environmental Determinants of Immune Phenotype Variance: Establishing a Path Towards Personalized Medicine (ID-RCB Number: 2012-A00238-35)"*.

The overall aim of the *Milieu Intérieur* study is to assess the factors underlying immunological variance within the general healthy population. The primary objective is to

[§] co-coordinators of the Milieu Intérieur Consortium. Additional information can be found at: http://www.pasteur.fr/labex/milieu-interieur.

define genetic and environmental factors that contribute to the observed heterogeneity in immune responses. This will be realized by characterizing and integrating (i) every-day life habits through an extensive questionnaire; (ii) genomic variability using genome-wide SNP genotyping and whole-exome sequencing; (iii) metagenomic diversity based on sequence analysis of bacterial, fungal and viral populations in fecal and nasal samples; (iv) induced transcriptional and protein signatures by whole microbes, microbial-associated molecular pattern (MAMP) agonists, medically relevant cytokines, or stimulators of the T cell response; and (v) variability in levels of circulating immune cell populations based on flow cytometry. The secondary objective is to establish a cell bank, including EBV-transformed B cell lines and fibroblasts from genetically annotated healthy individuals for use in mechanistic studies. To achieve the above-mentioned objectives, a total of 1000 healthy volunteers, descendants of mainland French persons for at least three generations, split equally by sex (1:1 sex ratio) and stratified across five-decades of life were recruited.

Herein, we present the socio-demographic and biological parameters that define our healthy donor cohort. Through unbiased statistical approaches, we identified known sex- and age-associated phenotypes, thus confirming the overall integrity of the data and validating our population sample as a reference for the healthy French population. Additional analyses provided new insight into the definition and risk factors of metabolic syndrome. Finally, we identified dependent and independent variables among the collected meta-data, results that will be applied to future association studies. This unique healthy donor population study may ultimately serve as a control reference sample for future disease based studies.

2. Materials and methods

2.1. Study objectives

In the context of a French scientific initiative, financed through the Investissement d'Avenir as part of a *Laboratoire d'Excellence* (LabEx) research program, the *Milieu Intérieur* Consortium was developed with the objective to define the determinants of human immune variance.

2.2. Clinical protocol and implementation

The clinical study was approved by the *Comité de Protection des Personnes* — Ouest 6 (Committee for the protection of persons) on June 13th, 2012 and by the French *Agence nationale de sécurité du médicament* (ANSM) on June 22nd, 2012. The study is sponsored by the Institut Pasteur (Pasteur ID-RCB Number: 2012-A00238-35), and was conducted as a single center study without any investigational product. The protocol is registered under ClinicalTrials.gov (study# NCT01699893).

Our strategy to define the parameters of a healthy population included the gathering of a working group composed of experts representing different clinical (medical biology, regenerative medicine, allergy, pediatrics, nutrition, psychiatry, lab medicine) and scientific (immunology, genetics, epidemiology, methodology, sociology, gut microbiota) specialties to help establish the criteria for qualifying an individual as a "healthy" donor while preserving the feasibility of recruitment and permitting robust statistical analysis. Specifically, this working group discussed the general eligibility criteria to pre-screen subjects (age, sex, BMI, self-reported ancestry, relatedness with the other subjects) and identified specific exclusion criteria that may impact the immune system and/or study procedures (e.g., chronic diseases known to involve the immune system, subjects with skin disorders that would compromise skin biopsy, etc.). Known medical, physiological, and behavioral factors with potential to affect immune cell activities or the microbiota environment were thoroughly reviewed and retained on the basis of their impact on the objective of our project, while preserving the feasibility of enrollment. We further considered the prevalence of donor characteristics, excluding those phenotypes that are below 1% in the population (e.g., peanut allergy), to ensure sufficient power for association studies. Efforts were made to avoid the selection of individuals following too conservative criteria (i.e., "super healthy" population), as this would compromise the underlying purpose of the study. A Scientific Advisory Board helped to develop and refine the study protocol, donor information and consent forms. They also provided oversight for ensuring consistency in screening, enrollment, body site sampling, and compliance with regulatory and data management requirements.

Laboratory protocols were standardized and staff members were trained in sample preparation protocols. Two risk assessments audits were conducted during the training period to refine sample handling and technical protocols. The clinical study opened at the investigator site (Biotrial, Rennes, France) on September 7th, 2012 and the first sample was collected on September 17th, 2012. All subjects provided informed consent prior to enrollment in the study. Subjects received compensation for their participation.

2.3. Subject screening and recruitment

A pre-existing donor database composed of ~110,000 donors was used for pre-screening potential participants in accordance with the study criteria. Additional advertising and website recruitment campaigns were launched in order to complete strata not sufficiently represented in the donor database. Eligibility was assessed by telephone interview and confirmed during a preliminary information meeting about the objectives of the research. Interested participants that met pre-screening criteria returned for the enrollment visit (referred to as V0). During V0, eligibility criteria were assessed in two stages: first, based on demographical data and clinical examination; and second, by analysis of blood and urine samples that were sent for clinical laboratory testing (Table S1). During the course of their participation in the *Milieu Intérieur* project, subjects were informed and encouraged to participate in a non-interventional French nutritional survey, *Etude Nutrinet-Santé* (www.etude-nutrinet-sante.fr) [25].

Upon receiving the clinical laboratory results, and confirming that all inclusion and exclusion criteria were respected, all subjects were invited to return for the inclusion visit (referred to as V1). Based on a defined randomization strategy, 500 subjects participated in a second visit (referred

to as V2) for repeat sampling (Fig. 1). V0 and V1 were scheduled with a 4–14 day interval; and V2 took place 14–42 days after V1. Of those that were randomized for repeat collections, 340 donors consented for a skin biopsy at V1 (n.b. the number of subjects with skin biopsy at V1 was restricted due to technical constraints). The financial compensation for participating in V0 was 50€, 150€ for V1, 100€ for V2 and 50€ for the skin biopsy.

2.4. Cross-sectional study

The *Milieu Intérieur* sample is composed of 1000 healthy volunteers, descendants of mainland French persons for at least three generations, stratified according to sex with a 1:1 ratio (500 subjects by sex); and age (5 decades of age: [20–29], [30–39], [40–49], [50–59] and [60–69] years, with 200 subjects per stratum). Subjects were randomized for a single or repeated collection (50% per stratum returned for V2). All donors were recruited by Biotrial Inc., a clinical research organization (CRO) based in Rennes, France. From September 17th, 2012 to August 8th, 2013, a total of 1238 donors were screened and 1012 healthy donors were enrolled. Twelve donors withdrew, so the final sample collection was composed of 1000 persons.

2.5. Inclusion/exclusion criteria

The study design concerned the definition of "healthy" in accordance with the goal to maximize our ability to associate genetic and epigenetic variation with defined phenotypes. This was achieved by establishing a detailed list of inclusion and exclusion criteria that ensured the recruitment of volunteers with a minimally perturbed immune system. Briefly, donors could not have evidence of, or report a history of neurological or psychiatric disorders, or severe/chronic/recurrent pathological conditions. Other exclusion criteria included: history or evidence of alcohol abuse, recent use of illicit drugs (including cannabis), recent vaccine administration, and recent use of immune modulatory agents. To avoid the influence of hormonal fluctuations in women during the peri-menopausal phase, only pre- or post-menopausal women were included. To avoid the presence of population structure in our study population (i.e., highly variable genetic backgrounds due to different ancestry), which would impact upon the power to detect genotype-to-phenotype associations, we restricted our study to individuals of European-descent, i.e., French citizens whose ancestry for three generations was of Metropolitan French origin (i.e., the subject's parents and grandparents were born in continental France).

2.6. Physical and clinical laboratory testing

After initial evaluation for recruitment criteria, additional physical examination and clinical laboratory testing were performed at visit V0 in order to fully include the donors. Donor BMI was restricted to ≥ 18.5 and ≤ 32 kg/m^2. 20 mL of blood sample (collected at V0 and V2, for repeat sampling) was used for clinical chemistry, hematologic and serologic assessments. A urinary human chorionic gonadotropin (hCG) test was performed on female donors, and urine toxicology screens for cannabinoid use, proteinuria and glycosuria were conducted on all donors. All clinical laboratory assessments were performed at the certified *Laboratoire de biologie médicale, Centre Eugene Marquis* (Rennes, France).

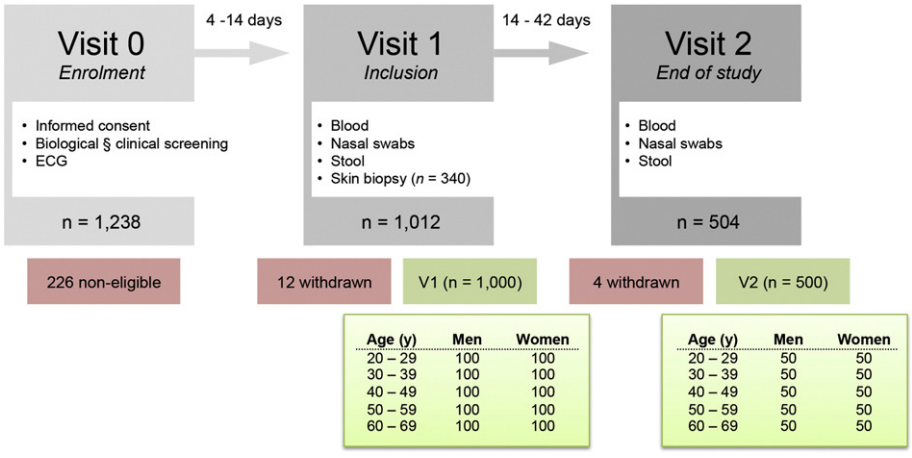


Figure 1 Schematic representation of donor recruitment for the *Milieu Intérieur* study. To include 1000 healthy persons stratified according to sex (500 men, 500 women) and age (200 donors per decade of life, 20–69 years of age), we enrolled a total of 1238 individuals at visit 0 (Enrollment). Of those screened, 226 donors were considered non-eligible for reasons of consent withdrawal (n = 54), past medical history (n = 67), identification of an exclusion criteria during the onsite physical examination (n = 54), or during laboratory testing (n = 51) (see Fig. S1). An additional 16 donors withdrew consent in the course of the study. During visit 1, whole blood, fecal samples and nasal swabs were collected. Punch biopsies of the skin were obtained from 340 of these donors. Half of the subjects were randomly selected (stratified by age and sex) to return for a visit 2, when repeat sampling of whole blood, fecal samples and nasal swabs was performed. Detailed medical histories and questionnaires were completed from all donors, recorded by medical personnel using an electronic case report form.

Biochemistry tests, immunological analysis, and viral serologies were performed on serum-separator tubes using AU 400 Olympus, DXC 660 I, Advanced 2020, DXI (Beckman Coulter), UF 50 Sysmex (Biomérieux) and Modular E170 (Roche) analyzers; Modular E170 (Roche), IRMA (Immunotech), RIA (Labodia), and Hydrasys (Sebia) systems; and DXI immunoassay system (Beckman Coulter), respectively. Hematology analysis were performed on EDTA tubes and coagulation tests were performed on citrate tubes using LH750 (Beckman Coulter) and STA-R (Stago) analyzers respectively.

2.7. Sample collections and storage

Blood, nasal swabs and stool samples were collected from all donors according to established protocols. For 500 individuals, samples were collected at V1 only; and for the remaining 500 donors, samples were collected at V1 and V2 – separated by 14–42 days – thus providing validation samples to be used in phenotypic studies. For donors randomized for two sample collections, biopsies of the skin were performed once at V1, in 340 donors.

From each volunteer, 20 mL of blood was collected into 2 Na Heparin tubes, and 5 mL of blood into 1 EDTA tubes for cytometric studies and banking of DNA, respectively. These tubes were maintained at 18–25 °C, during daily transport to Institut Pasteur (Paris), and processed within 6 h of collection. An additional 50 mL of blood was collected using a pre-heparinized large-bore syringe. This sample was aliquoted into 40 — 1 mL TruCulture® tubes within 15 min of collection. The TruCulture® systems were developed to provide reproducible induction of innate or adaptive immune responses and are described elsewhere [26]. After stimulation, the liquid supernatants from the TruCulture® tubes were aliquoted, and the cell pellet was stabilized in Trizol. Both samples were stored at −80 °C.

Stool samples were collected in a double-lined sealable bag with the outer bag containing a GENbag Anaer atmosphere generator (Aerocult, Biomerieux), used to maintain anaerobic conditions, and an anaerobic indicator strip (Anaerotest®, Merck Millipore) to record maintenance of the anaerobic atmosphere. Subjects were asked to produce the fecal specimen at their home within 24 h before their scheduled visits (V1, V2). Upon reception at the clinical site, the specimen was aliquoted into cryotubes and stored at −80 °C.

Nasal swabs were obtained with sterile, dry flocked swabs (FLOQSwab™). Right and left nostrils were sampled separately. All swabs were stored in stabilization media and frozen at −80 °C.

Skin punch biopsies were performed under local anesthesia. The biopsy was taken using a sterile single use biopsy punch (7 mm * 3 mm round dermal punch). The material collected was shipped the same day of the collection at 4 °C to Genethon (Evry, Ile de France, France) where human fibroblast cell lines were generated and aliquots stored.

The processing of each donor involved the production and registration of more than 180 tubes. To ensure effective traceability of all samples, a customized software system was developed for managing 2D barcoded tubes. A central sample database has been established to aggregate all sample information for each donor, visit, and sample type.

2.8. Case report forms

Detailed medical histories and questionnaires collecting general information about socio-demographic, lifestyle and family health history were recorded in an electronic case report form. For example, the questionnaire collected information concerning family status, income, occupational status and educational level, smoking habits, alcohol intake, sleeping habits, depressive symptoms, family medical history and nutritional behavior and habits (for details, see supplementary material: case report forms).

2.9. Statistical analyses

Statistical analysis was performed using the Open Source R Software, version 3.0.1 [27]. All statistical graphics were generated using the 'ggplot2' package, version 0.97 [28]. The hierarchical clustering of our continuous explanatory variables was based on the Spearman's correlation score (Rs) using the 'hclust' function available from the base functions. Random Forest (RF) models [2] were built using the 'randomForest' package (version 4.6–7). For each RF model built (sex, age categories, smoking status) a forest of 1000 trees was computed, and the 'mtry' parameter was set to be the square root of the number of available explanatory variables. When investigating outliers in our sample, we used a z-score based criterion. For a given metric, we considered a donor as an outlier if its measurement was 3 standard deviations away for the mean of the whole sample. Principal component analysis (PCA) on the outlier cases of our dataset was performed with the 'FactoMineR' package version 1.25. Regression analyses were conducted using the *glm* function in R. Levels of immunoglobulins were log-transformed and standardized, prior to regression analyses. The representative nature of the cohort was assessed by stratified sampling: 500 individuals were sampled 10,000 times among all cohort participants, in order to match the proportions of males and females and of 10-decades age groups observed in the general population. Public data from the *Institut National de la Statistique et des Etudes Economiques* (INSEE; National Institute of Statistics and Economic Studies) were retrieved for the entire *Ille-et-Vilaine* French department and the city of Rennes (http://www.insee.fr/en/default.asp).

3. Results

3.1. Sample and data overview

From September 17th, 2012 to August 8th, 2013, a total of 1238 donors were screened and 1012 healthy donors were enrolled (Fig. 1). The reasons for excluding the 226 pre-screened donors included withdrawal of consent (n = 54), as well as medical history (n = 67), physical exam findings (n = 54) or laboratory test results (n = 51) that were not in accordance with the defined inclusion or non-exclusion criteria (Fig. S1, Table S1). Questionnaires were completed and clinical laboratory testing was performed at visit V0 (Tables 1–2, Tables S1–3). Among those enrolled, 12 donors withdrew consent during the collection phase of the protocol.

This resulted in a final set of 1000 subjects, with all donors having completed an evaluation at visit V1 and 50% of them (500 subjects) returning for evaluation at visit V2. During V1, 340 had a skin biopsy.

Donor recruitment was conducted in the vicinity of the city of Rennes, in the *Ille-et-Vilaine* French region. We first compared the socio-economic characteristics of the *Milieu Intérieur* cohort to those of the general population of this region (Table S4), after adjustment to match regional age and sex stratification (see Methods section). We observed ~10% higher unemployment levels in the *Milieu Intérieur* cohort (16.9% with 95% confidence interval (CI) [14.7%–19.1%]), when compared to the *Ille-et-Vilaine* region or the city of Rennes (6.0% and 8.2%, respectively) (Table S5). The cohort also contained a higher proportion of retired persons (16.6% [15.1%–17.9%] *versus* 8.1% and 4.9%) (Table S5). Among employed people, socio-professional categories of the *Milieu Intérieur* donors were biased towards more employees and fewer laborers. We also observed that the educational level of the *Milieu Intérieur* donors was generally higher. Finally, 42.5% [39.5%–45.4%] of participants were renters, a value that is intermediate between those of the *Ille-et-Vilaine* region and the city of Rennes (33.1% and 59.5%, respectively) (Table S5), consistent with the fact that donors reside in Rennes as well as in surrounding areas.

Table 1 Sample collections obtained from study subjects.

		Visit 0	Visit 1	Visit 2
Whole blood collection		20 mL	87 mL	83 mL
CLT[a]	Complete blood count: RBC count, HCT, HGB, MCV, MCH, WBC count, NEUTRO, MONO, LYMPHO, EOS, BASO, PLT count	X		X
CLT	Blood electrolytes: Na, K, Ca, P, Cl, HCO3	X		
CLT	Liver function tests: HSA, ALP, AST, ALT, GGT, BILI, TPROT	X		
CLT	Inflammation: CRP	X		X
CLT	Renal function tests: BUN, CREAT, UA	X		
CLT	Lipids/metabolism: GLUC, TCHOL, LDL, HDL, TRIGLY	X		
CLT	Serology: HBV (HBs Ag), HCV (anti HCV IgG, viral load if Ab+), HIV (anti-HIV IgM, IgG), CMV (anti-CMV IgG), HTLV-1 (anti-HTLV-1 IgG), influenza (anti-Influenza IgG)	X		
CLT	Immunoglobulin electrophoresis: serum immunoglobulin concentrations (IGM, IGG, IGA, IGE)	X		
R[b]	Immunophenotyping (Na Heparin tube): cytometric analysis for major subsets of immune cells in circulation		X	X
R	Functional immune stimulation (Na Heparin syringe): TruCulture tubes ×40		X	X
R	Genetic tests (EDTA tube): TruCulture tubes ×40		X	
Urine collection		>5 mL		>5 mL
CLT	Biochemistry: proteinuria, glycosuria (dipstick)	X		
CLT	Pregnancy test: bHCG concentration (women only)	X		X
CLT	Toxicology: cannaboids	X		
Fecal sample collection			>100 g	>100 g
R	Enterotyping: bacterial, viral, fungal strains		X	X
Nasal swab			2 swabs	2 swabs
R	Enterotyping: bacterial, viral, fungal strains		X	X
Biopsy		7 mm punch		
R	Punch biopsy of skin	X (n = 340)		

[a] CLT, clinical laboratory test; RBC, red blood cell; HCT, hematocrit; HGB, hemoglobin; MCV, mean corpuscular volume; MCH, mean corpuscular hemoglobin; WBC, white blood cell, NEUTRO, neutrophil; LYMPHO, lymphocyte; EOS, eosinophil; BASO, basophil; PLT, platelet; Na, sodium; K, potassium; Ca, calcium; P, phosphorus; Cl, chloride; HCO_3, bicarbonate; HSA, human serum albumin; ALP, alkaline phosphate; AST, aspartate aminotransferase; ALT, alanine aminotransferase; GGT, gamma-glutamyl transpeptidase; BILI, bilirubin; TPROT, total protein; CRP, C-reactive protein; BUN, blood urea nitrogen; CREAT, creatinine; UA, urinalysis; GLUC, glucose; TCHOL, total cholesterol; LDL, low density lipoprotein; HDL, high density lipoprotein; TRIGLY, triglycerides; HBV, hepatitis B virus; HCV, hepatitis C virus; HIV, human immunodeficiency virus; CMV, cytomegalovirus; HTLV, human T cell lymphotropic virus; βHCG, beta-human chorionic gonadotropin.

[b] R, Research tests.

Table 2 Socio-demographic information for study subjects.

	Total (n = 1000)		Male (n = 500)		Female (n = 500)		20–29 years (n = 200)		30–39 years (n = 200)		40–49 years (n = 200)		50–59 years (n = 200)		60–69 years (n = 200)	
Donor characteristics	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
BMI *																
18 < BMI ≤ 25	635	63.5	283	56.6	352	70.4	160	80	135	67.5	128	64	111	55.5	101	50.5
25 < BMI ≤ 30	300	30	181	36.2	119	23.8	34	17	55	27.5	54	27	72	36	85	42.5
30 < BMI ≤ 32	65	6.5	36	7.2	29	5.8	6	3	10	5	18	9	17	8.5	14	7
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Education §																
No diploma	38	3.8	20	4	18	3.6	6	3	4	2	10	5	10	5	8	4
Primary school certificate only	46	4.6	22	4.4	24	4.8	0	0	1	0.5	1	0.5	11	5.5	33	16.5
CAP, BEP, apprenticeship certificate, BEPC (High school diploma equivalent)	332	33.2	170	34	162	32.4	34	17	47	23.5	84	42	84	42	83	41.5
Baccalaureate or technician's certificate	268	26.8	130	26	138	27.6	64	32	65	32.5	44	22	50	25	45	22.5
Higher education (no professional degree)	156	15.6	75	15	81	16.2	49	24.5	36	18	31	15.5	26	13	14	7
Higher education (Masters, PhD, engineer's diploma, MD, etc.)	160	16	83	16.6	77	15.4	47	23.5	47	23.5	30	15	19	9.5	17	8.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Employment																
Steady job	510	51	247	49.4	263	52.6	75	37.5	148	74.0	155	77.5	118	59	14	7
Unemployed	158	15.8	91	18.2	67	13.4	38	19	43	21.5	37	18.5	34	17	6	3
Student	74	7.4	41	8.2	33	6.6	71	35.5	3	1.5	0	0	0	0	0	0
Looking for first job	16	1.6	5	1.0	11	2.2	14	7	1	0.5	1	0.5	0	0	0	0
Housewife/househusband	21	2.1	2	0.4	19	3.8	2	1	4	2.0	5	2.5	8	4	2	1
Retired	215	21.5	114	22.8	101	20.2	0	0	0	0.0	2	1	40	20	173	86.5
NA	6	0.6	0	0.0	6	1.2	0	0	1	0.5	0	0	0	0	5	2.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Full-time	406	40.6	219	43.8	187	37.4	65	32.5	115	57.5	122	61	96	48	8	4
Part-time	135	13.5	38	7.6	97	19.4	28	14	38	19	36	18	24	12	9	4.5
Not answered	459	45.9	243	48.6	216	43.2	107	53.5	47	23.5	42	21	80	40	183	91.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Exclusively during the day	369	36.9	155	31	214	42.8	59	29.5	102	51	110	55	86	43	12	6
Exclusively during the night	36	3.6	12	2.4	24	4.8	7	3.5	10	5	10	5	8	4	1	0.5
Without fixed hours	135	13.5	90	18	45	9	27	13.5	41	20.5	37	18.5	26	13	4	2
Not answered	460	46	243	48.6	217	43.4	107	53.5	47	23.5	43	21.5	80	40	183	91.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Socio-professional category																
Farmer	10	1	6	1.2	4	0.8	1	0.5	2	1	1	0.5	2	1	4	2
Artisans. tradesman or company director	46	4.6	36	7.2	10	2	4	2	11	5.5	10	5	7	3.5	14	7
Senior executive or independent profession	42	4.2	29	5.8	13	2.6	1	0.5	15	7.5	4	2	7	3.5	15	7.5
Middle management	113	11.3	59	11.8	54	10.8	3	1.5	16	8	27	13.5	28	14	39	19.5
Employee	507	50.7	197	39.4	310	62	82	41	116	58	115	57.5	106	53	88	44
Labourer	100	10	76	15.2	24	4.8	11	5.5	19	9.5	19	9.5	27	13.5	24	12
Other categories (e.g. artist. clergy. soldier. police officer)	60	6	48	9.6	12	2.4	7	3.5	11	5.5	18	9	15	7.5	9	4.5
Not answered	122	12.2	49	9.8	73	14.6	91	45.5	10	5	6	3	8	4	7	3.5

(continued on next page)

Table 2 (*continued*)

Donor characteristics	Total (n = 1000)		Male (n = 500)		Female (n = 500)		20–29 years (n = 200)		30–39 years (n = 200)		40–49 years (n = 200)		50–59 years (n = 200)		60–69 years (n = 200)	
	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Monthly income of the household																
0 < salary ≤1000	158	15.8	75	15	83	16.6	78	39	16	8	21	10.5	23	11.5	20	10
1000 < salary ≤ 2000	340	34	160	32	180	36	73	36.5	75	37.5	70	35	60	30	62	31
2000 < salary ≤ 3000	256	25.6	134	26.8	122	24.4	27	13.5	59	29.5	50	25	59	29.5	61	30.5
3000 < salary ≤ 4000	164	16.4	82	16.4	82	16.4	12	6	41	20.5	43	21.5	34	17	34	17
4000 < salary ≤ 5000	59	5.9	35	7	24	4.8	8	4	4	2	12	6	17	8.5	18	9
5000 < salary	17	1.7	11	2.2	6	1.2	1	0.5	2	1	4	2	6	3	4	2
Not answered	6	0.6	3	0.6	3	0.6	1	0.5	3	1.5	0	0	1	0.5	1	0.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Family status																
Single with no children	320	32	155	31	165	33	108	54	45	22.5	42	21	56	28.0	69	34.5
Single with children	87	8.7	30	6	57	11.4	4	2	23	11.5	29	14.5	25	12.5	6	3
In couple but no children	259	25.9	142	28.4	117	23.4	58	29	22	11	27	13.5	48	24.0	104	52
In couple with children	328	32.8	172	34.4	156	31.2	24	12	110	55	102	51	71	35.5	21	10.5
Not answered	6	0.6	1	0.2	5	1	6	3	0	0	0	0	0	0	0	0
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Housing																
Owner	584	58.4	286	57.2	298	59.6	30	15	116	58	140	70	144	72	154	77
Renter	401	40.1	209	41.8	192	38.4	157	78.5	83	41.5	59	29.5	56	28	46	23
Not answered	15	1.5	5	1	10	2	13	6.5	1	0.5	1	0.5	0	0	0	0
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Sport activities																
Little or no physical activity	274	27.4	124	24.8	150	30	71	35.5	67	33.5	54	27	52	26	30	15
Moderate physical activity	656	65.6	314	62.8	342	68.4	110	55	115	57.5	133	66.5	139	69.5	159	79.5
Intensive physical activity (competitions)	70	7	62	12.4	8	1.6	19	9.5	18	9	13	6.5	9	4.5	11	5.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Smoking (cigarettes)																
Non smoker	526	52.6	233	46.6	293	58.6	112	56	96	48	94	47	109	54.5	115	57.5
Former smoker	265	26.5	151	30.2	114	22.8	20	10	48	24	65	32.5	59	29.5	73	36.5
Smoker	209	20.9	116	23.2	93	18.6	68	34	56	28	41	20.5	32	16	12	6
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Alchool																
Never	91	9.1	43	8.6	48	9.6	12	6	24	12	26	13	16	8	13	6.5
Occasionally	523	52.3	220	44	303	60.6	96	48	104	52	108	54	111	55.5	104	52
Once per week	196	19.6	110	22	86	17.2	58	29	39	19.5	29	14.5	40	20	30	15
Several times per week	154	15.4	100	20	54	10.8	34	17	31	15.5	34	17	24	12	31	15.5
Everyday	36	3.6	27	5.4	9	1.8	0	0	2	1	3	1.5	9	4.5	22	11
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100
Feeding at birth																
Breastfed milk	286	28.6	137	27.4	149	29.8	66	33	47	23.5	49	24.5	61	30.5	63	31.5
Formula milk	465	46.5	199	39.8	266	53.2	82	41	113	56.5	108	54	87	43.5	75	37.5
Not known	246	24.6	163	32.6	83	16.6	52	26	40	20	42	21	51	25.5	61	30.5
Not answered	3	0.3	1	0.2	2	0.4	0	0	0	0	1	0.5	1	0.5	1	0.5
Total	1000	100	500	100	500	100	200	100	200	100	200	100	200	100	200	100

Table 2 (*continued*)

	Total (n = 1000)		Male (n = 500)		Female (n = 500)		20–29 years (n = 200)		30–39 years (n = 200)		40–49 years (n = 200)		50–59 years (n = 200)		60–69 years (n = 200)	
Donor characteristics	n	%	n	%	n	%	n	%	n	%	n	%	n	%	n	%
Contraception †																
Intrauterine device (IUD)					72	14.4	7	7	29	29	29	29	7	7	0	0
Oral Contreception					117	23.4	55	54	37	37	20	20	5	5	0	0
Male or female condomn					94	18.8	29	28	26	26	32	32	7	7	0	0
Tubal ligation					7	1.4	0	0	2	2	5	5	0	0	0	0
Other method of contraception					14	2.8	7	7	3	3	2	2	2	2	0	0
None					13	2.6	4	4	3	3	5	5	1	1	0	0
Not answered (or not asked)					183	36.6	0	0	0	0	7	7	78	78	100	100
Total					500	100	102	100	100	100	100	100	100	100	100	100

* Study inclusion criteria set limits for BMI.
§ The certificat d'études primaires (CEP) was a diploma awarded at the end of elementary primary education in France (from 11 to 13 years inclusive until 1936) and certifying that the student had acquired basic skills in writing, reading, mathematics, history, geography and applied sciences. It was officially discontinued in 1989.
† Questions were posed to pre-menopausal women only. Multiple choice was allowed.

3.2. Analysis of sex-, age-, and smoking habit- associated biological parameters

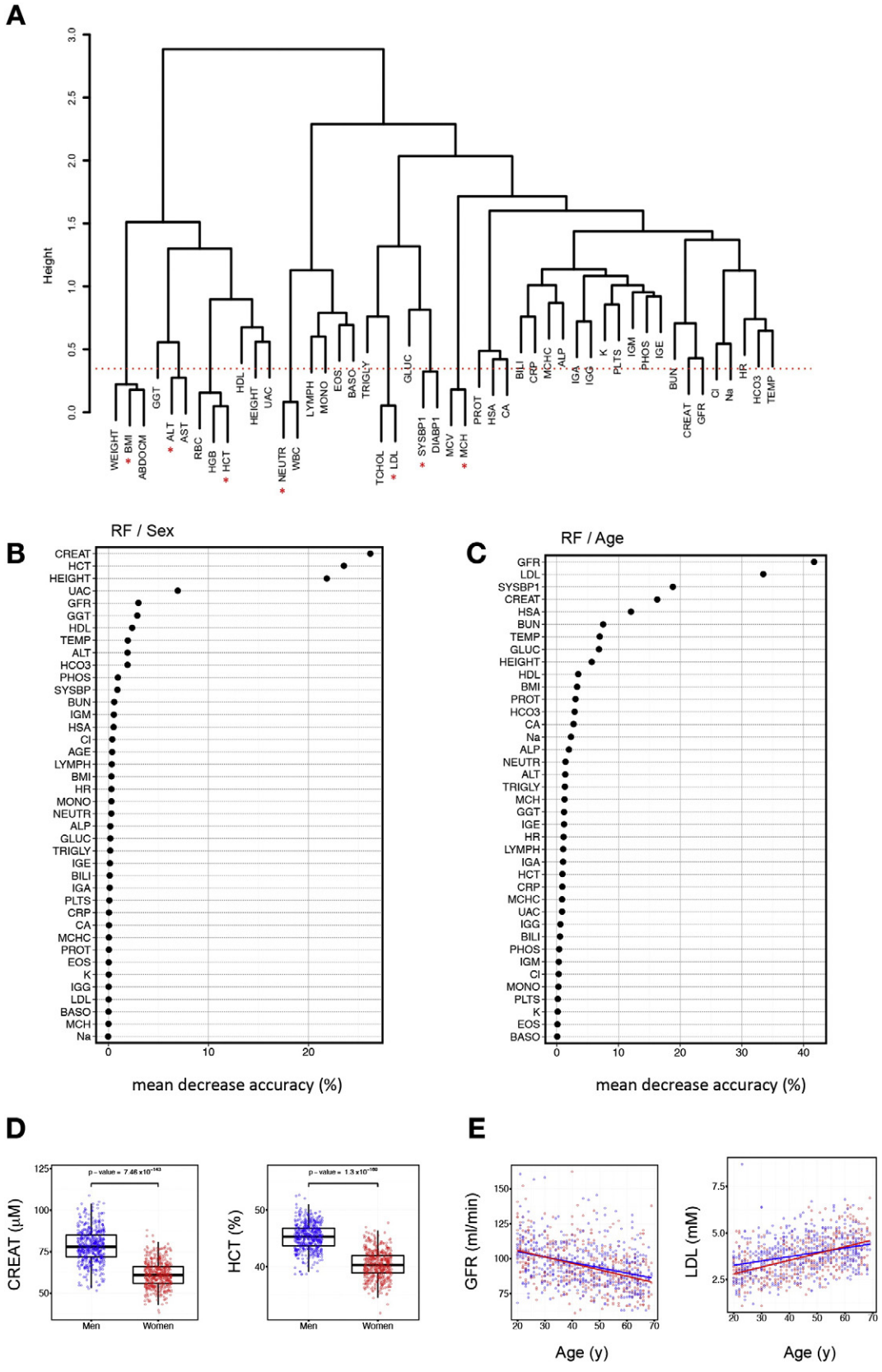
A total of 328 variables were obtained from the study questionnaire (see Case Report Form). The physical examination and clinical laboratory analyses were assembled into a data warehouse using LabKey [1]. To validate the data collected in our study, we first tested our ability to identify known biological correlates of sex, age or smoking-habits. To achieve this, we utilized a discovery-based approach. With the initial aim of reducing the complexity of the biochemical, hematologic and serologic data – thereby increasing the power of our association studies – we correlated all quantitative values from clinical laboratory data for the 1000 donors to each of the other variables using a Spearman's correlation matrix. Results were clustered and plotted using a dendrogram to represent the relationships between variables, with *height* (ordinate axis) being inversely related to the correlation coefficient (Fig. 2A). For pairs or groups of variables that showed high correlation (height < 0.3, equivalent to $r_s > 0.67$), we selected one representative variable (indicated by red star). Next, we utilized the standard machine learning Random Forest (RF) approach [2], applied to the dataset in order to identify the variables that are most important to correctly classify donors based on sex (Fig. 2B) or age (Fig. 2C). Of note, bootstrap aggregation (also referred to as bagging) of data was selected due to its stability and accuracy in statistical classification and regression. This approach, which reduces variance and avoids overfitting, can be applied to a variety of binary data (e.g., male vs. female) and continuous variables (e.g., age). Using this method, we found that serum creatinine (CREAT) concentration, hematocrit (HCT) and height are the features that are most predictive of sex; and lower glomerular filtration rate (GFR), higher plasma low density lipoprotein (LDL) concentration and higher systolic blood pressure (SYSBP1) to be most associated with age. These results were validated using univariate tests (Table S3) and representative box-plots are shown for the most significant variables (Figs. 2D, E).

To assess other determinants, while controlling for sex- and age-associated effects on clinical laboratory data, we used a linear regression model, considering sex and age as independent covariates. This permitted us to examine the features predictive of smoking habits, again employing Random Forest analysis to segregate non-smokers, not exposed to second-hand smoke ($n = 394$) from active smokers ($n = 208$) present in our sample (Fig. 3A). Validating prior findings [3–6], we report that serum IgG and bilirubin concentrations were lower in smokers as compared to non-smokers (Figs. 3A, B); whereas monocyte, neutrophil and lymphocyte numbers were higher in smokers as compared to non-smokers (Figs. 3A, C). These observations may be related to lower antioxidant concentrations [7], and a diminished adherence of leucocytes to blood vessel walls [8]. Interestingly, a comparison of non-smokers and prior smokers present in our sample ($n = 251$) indicates that smoking cessation restores the biochemical and immunological phenotypes associated with non-smokers (Figs. 3B, C). Together, these data highlight that our sample population can be used to study associations in the general French population and can be compared to prior study cohorts.

3.3. Smoking habits confer increased risk for metabolic syndrome among healthy donors

Over the past two decades, there has been increasing concern about the prevalence of obesity and its association with diabetes and cardiovascular disease (CVD), and their link to metabolic syndrome [9]. While several assessment scores have been established, the metabolic syndrome score is now a widely applied measure. Metabolic syndrome is most commonly defined by six variables: increased abdominal circumference

(AbdoCM > 94cm European men, > 80cm European women), elevated systolic blood pressure (SYSBP ≥ 130mmHg), elevated diastolic blood pressure (DYSBP ≥ 85mmHg), elevated triglyceride levels (TG ≥ 1.7mM), diminished levels of high density lipoprotein (HDL < 1 mM men, < 1.3 mM women) and glucose concentration (≥ 6.1 mM) [10]. We thus analyzed donors for these six criteria, using accepted cut-values for European men and women, and for each criterion, data was



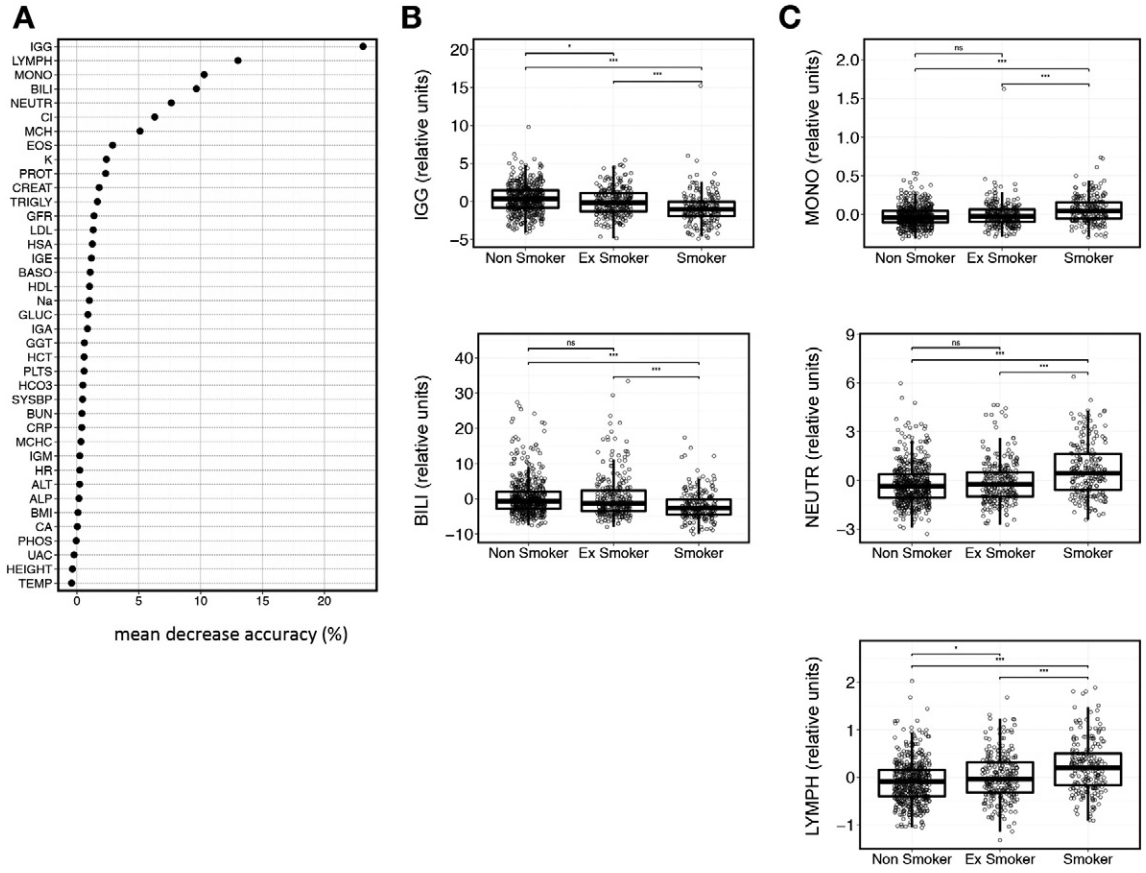
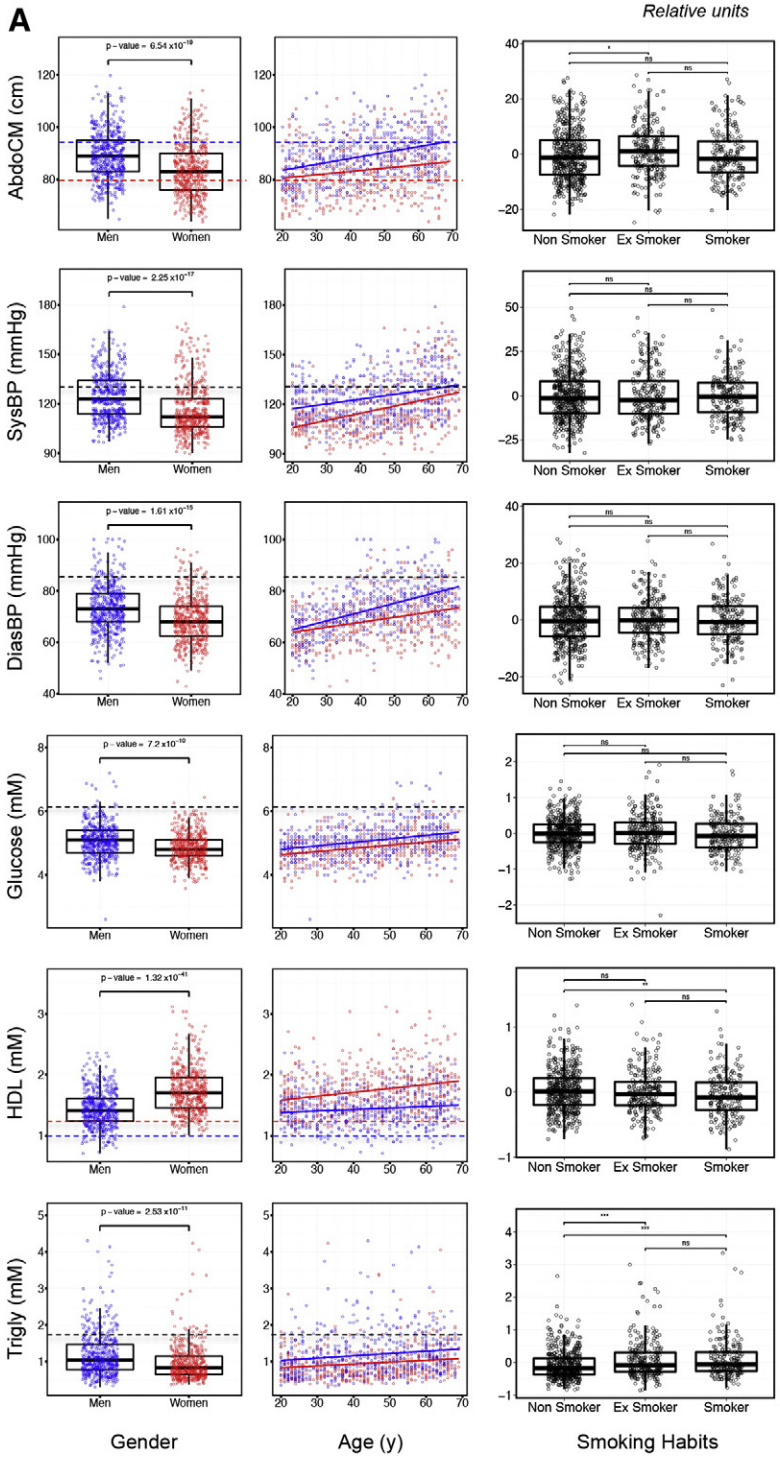


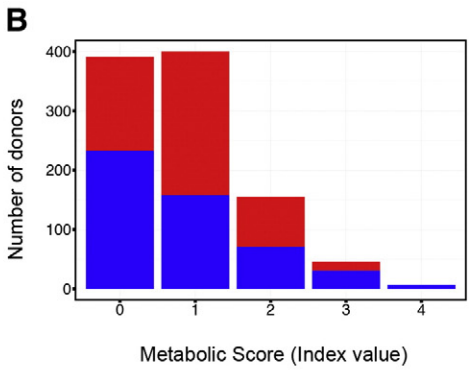
Figure 3 Tobacco use associated with lower IgG, bilirubin concentrations and higher number of circulating monocytes, neutrophils and lymphocytes. (A) The same variables selected from Fig. 2A were regressed for sex and age, then subjected to Random Forest analysis to identify variables that discriminate active smokers (n = 208) from non-smokers with no reported passive smoking (n = 394). The random forest classification had an estimated error rate of 24.5% on the out-of-bag error. Variables are reported according to their impact on the out-of-bag error (percentage of mean decrease accuracy). (B, C) The top variables found to be lower (B) or higher (C) in smokers as compared to non-smokers are shown, with the inclusion of ex-smokers with no reported passive smoking (n = 201) as an additional group. Variables that measured importance through permutation are depicted: serum IgG concentration (IGG, g/L); bilirubin concentration (BILI, μM); the absolute number of monocytes (MONO, $\times 10^3/\mu L$), neutrophils (NEUTR, $\times 10^3/\mu L$) and lymphocytes (LYMPH, $\times 10^3/\mu L$). Individual donors are represented by an open black circle. The data is overlaid by boxplots that represent the set of donors tested; the median value is indicated by the black bar, the lower and upper edges correspond to the first and third quartiles (the 25th and 75th percentiles), respectively, and the whiskers extend to the highest/lowest value that is within 1.5x interquartile range (IQR). A student t-test was used to determine statistical differences between two groups for the given variables (p-value indicated and bracket defining the two groups being compared, *, $p < 0.05$; ***, $p < 0.001$; ns, not significant).

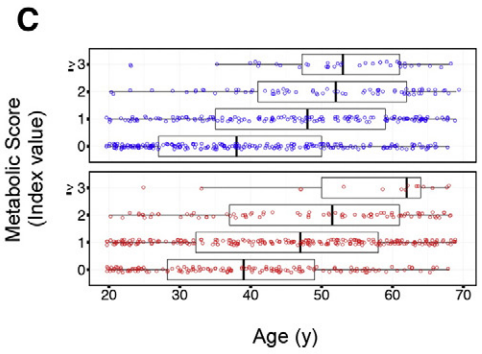
Figure 2 Unbiased assessment of the clinical laboratory data revealed expected sex-, and age-associations. (A) Biological measurements from the electronic case report forms (eCRF) and clinical laboratory data were evaluated using Spearman's correlation matrix and plotted using a dendrogram. For subsequent data mining, representative variables were selected (red star) from pairs or groups of factors showing high correlation (height < 0.3, indicated by dotted red line). (B, C) Random Forest method was employed to identify variables that discriminate men and women (B) or age, used as a continuous variable (C). The random forest classification on sex indicated an estimated error rate of 2.4% based on the out-of-bag error, while the random forest regression on age had a mean of squared residuals of 95.1, with 54% variance explained. Variables are reported according to their impact on the out-of-bag error (percentage of mean decrease accuracy). (D, E) The top two variables for sex (D) or age (E) that measured importance through permutation are depicted: creatinine concentration (Creat, μM); hematocrit (HCT, %); glomerular filtration rate (GFR, mL/min); and low density lipoprotein concentrations (LDL, mM). Individual donors are represented by an open circle (blue, men; red, women). The data is overlaid by boxplots that represent the set of donors tested; the median value is indicated by the black bar, lower and upper edges correspond to the first and third quartiles (the 25th and 75th percentiles), respectively, and the whiskers extend to the highest/lowest value that is within 1.5x interquartile range (IQR); and a student t-test was used to determine statistical differences between men and women for the given variables (p-value indicated at the top of each graph) (D). Regression lines indicate the respective curve for men and women and results of univariate statistical analyses can be found in Table S2 (E).

reported in relation to sex and age (Fig. 4A). Notable differences in the diastolic, systolic blood pressure, triglyceride and glucose levels were observed, with men having significantly higher levels than women (n.b., comparisons were made for those criteria in which reference values were similar between men and women). In all instances, biologic measures showed a significant increase with advancing age. Quantitative laboratory data were again regressed out for sex and age effects, and the component variables were evaluated among smokers, ex-smokers and non-smokers (Fig. 4A).

To generate a composite metabolic score, 1 point was assigned for each of the assessed variables, taking blood pressure elevation as a single value (i.e., elevated SYSBP and/or DIASBP = 1 point) [11]. The index value for the metabolic score indicated that 400 individuals (40%) had at least one positive criterion, 155 donors (15.5%) had a score of 2, and 53 donors (5.3%) had a score of ≥ 3, despite meeting all criteria for being a healthy donor (Fig. 4B). Notably, women had a higher probability of scoring ≥ 1 due to the low threshold for abdominal circumference for







D

Metabolic Score	0	1	2	≥3
Men	233	158	71	38
Women	158	242	84	15
χ^2 (df)				43.1 (3)
p-value				2.3 x 10^{-9}

Metabolic Score	0	1	2	≥3
20-29yr	111	68	17	3
30-39yr	97	74	27	2
40-49yr	81	80	28	11
50-59yr	63	86	35	16
60-69yr	39	92	48	21
χ^2 (df)				87.9 (12)
p-value				1.3 x 10^{-13}

Metabolic Score	0	1	2	≥3
Non Smokers	211	229	73	13
Smokers	95	66	34	13
χ^2 (df)				13.1 (3)
p-value				4.4 x 10^{-3}

European women (299 women vs. 135 men being above the respective abdominal circumference cut value) (Table S6). Indeed all variables that constitute the metabolic score, with the exception of HDL, showed a sex bias (Fig. S2). As a result, we observed a significant association between sex and the metabolic index ($\chi^2 = 43.1$, degrees of freedom (df) = 3, $p = 2.3 \times 10^{-9}$; Fig. 4B); and there was a significant increase in median age when donors were stratified based on metabolic index ($\chi^2 = 87.9$, df = 12, $p = 2.3 \times 10^{-13}$, Fig. 4C). While smoking habits did not impact each of the individual variables, there was a significant relative risk increase associated with smoking as compared to non-smokers, after regressing out sex and age (Fig. 4D, $\chi^2 = 13.1$, df = 3, $p < 0.005$; Table S6). Thus, smoking habits are an independent risk factor for the metabolic syndrome, distinct from its known association with CVD.

3.4. Sex, age and relationship status are risk factors for altered immunological status

Common infections and abnormal levels of immunoglobulins are conditions that may alter the immunological state of individuals. In our study population, circulating levels of immunoglobulins (i.e., IgM, IgG, IgE and IgA) were quantified, as well as influenza- and cytomegalovirus (CMV)- specific IgG antibodies (Table S3). We investigated association of demographic, socio-economic variables and/or lifestyle habits with these serological parameters (Table 2). Regression analyses were used to identify independent predictors among the 73 available variables.

Using univariate regression analysis, we found that positive detection of anti-influenza virus IgG was significantly associated with higher stature ($p = 5.7 \times 10^{-4}$), sex (incidence of 86.1% and 77.8% in men and women; $p = 1.0 \times 10^{-3}$), a higher weight ($p = 2.1 \times 10^{-3}$) and a younger age ($p = 2.3 \times 10^{-2}$). However, only sex and age remained significantly associated with influenza specific IgG when multiple regression analyses were performed, including sex, age, height and weight predictors ($p = 1.1 \times 10^{-3}$, 2.4×10^{-2}, 0.85 and 0.17, respectively). Similarly, when stratifying by sex, parameters such as height and weight were no longer associated with anti-influenza virus IgG ($p > 0.05$). More interestingly, a younger age was found to be associated with infection only in women ($p_{women} = 7.8 \times 10^{-3}$ vs. $p_{men} = 0.91$). Indeed, the sex ratio in IgG-samples was 0.90 in individuals between 20 and 39 years, while it dropped to 0.53 in people between 40 and 69 years. Together, and in accordance with previous findings [12], our analyses support the notion that men are at higher risk of being positive for anti-influenza IgG, and suggests a female-specific influence of age on influenza infection.

Conversely, factors associated with positive detection of anti-CMV IgG were an older age ($p = 7.6 \times 10^{-7}$), the consumption of raw fruits and vegetables ($p = 4.0 \times 10^{-4}$ and $p = 1.6 \times 10^{-4}$, respectively), being female ($p = 1.7 \times 10^{-3}$), a shorter sleep duration ($p = 1.5 \times 10^{-3}$), single status ($p = 5.8 \times 10^{-3}$) and a lower stature ($p = 1.0 \times 10^{-2}$), by univariate regression analysis. All these factors remained significantly associated in a multiple regression analysis, with the exception of height and the consumption of raw fruits (Table S7). While age was consistently associated with CMV infection in both men and women ($p_{men} = 2.1 \times 10^{-2}$ and $p_{women} = 3.8 \times 10^{-3}$), consumption of raw vegetables ($p_{men} = 0.59$ and $p_{women} = 8.2 \times 10^{-3}$), relationship status ($p_{men} = 0.48$ and $p_{women} = 4.6 \times 10^{-3}$) and hours of sleep ($p_{men} = 0.13$ and $p_{women} = 7.7 \times 10^{-2}$) were significant (or trended towards significance) in women only. By contrast, the association of CMV infection with being single in men was restored when restricting the analysis to men who have children ($p_{men} = 1.9 \times 10^{-2}$; Fig. S2).

Next, the different classes of immunoglobulins were evaluated for their association with available demographic data, using multiple regression of the most significant univariate predictors. Elevated IgG levels were associated with smoking ($p = 1.2 \times 10^{-13}$; Fig. 3), influenza virus infection ($p = 5.2 \times 10^{-4}$), multivitamins consumption ($p = 8.2 \times 10^{-3}$) and being a woman ($p = 3.1 \times 10^{-2}$). The three former factors remained significant (or trended towards significance) in males and females, when considered separately. Elevated IgM levels were associated with being a woman ($p = 5.0 \times 10^{-4}$) and with lower BMI ($p = 1.8 \times 10^{-2}$). Elevated IgE levels were associated with being a man ($p = 1.4 \times 10^{-6}$), younger age ($p = 1.4 \times 10^{-3}$), exposure to silica

Figure 4 Among healthy donors, being a male, increasing age and tobacco use are independent risk factors for higher metabolic score index value. (A) Variables that are used for determining an individual's metabolic score are plotted individually, representing differences between men and women; across age, as a continuous variable; or among non-smokers, ex-smokers and smokers. The variables included abdominal circumference (AbdCM, cm), systolic blood pressure (SysBP, mm Hg), diastolic blood pressure (DysBP, mm Hg), glucose levels (mM), high density lipoprotein concentration (HDL, mM), and triglyceride concentrations (Trigly, mM). For depiction of sex and age associations, individual donors are represented by an open circle (blue, men; red, women). For smoking habit associations, data was regressed for sex and age, data is plotted as relative units for respective plots, and individual donors are represented by an open black circle. Dotted black lines indicate reference values for European population; and where relevant sex-dependent reference indicators are used (blue dotted line, men; red dotted line, women). (B) The metabolic score was calculated for each donor and plotted to represent number of donors having indicated index values. Bar graphs indicate men (blue) and women (red). (C) Age association with metabolic score index values is shown, indicating men (blue circles) in the top plot, and women (red circles) in the bottom plot. (D) Contingency tables are shown for indicated comparisons and results from χ^2 testing are reported. The data is overlaid by boxplots that represent the set of donors tested; the median value is indicated by the black bar, the lower and upper edges correspond to the first and third quartiles (the 25th and 75th percentiles), respectively, and the whiskers extend to the highest/lowest value that is within 1.5x interquartile range (IQR) (A, C). Where indicated a student t-test was used to determine statistical differences between two groups for the given variables (*p*-value indicated and bracket defining the two groups being compared, *, $p < 0.05$; **, $p < 0.01$; ns, not significant).

($p = 1.0 \times 10^{-2}$), smoking ($p = 4.7 \times 10^{-2}$), and a familial history of atopy ($p = 5.0 \times 10^{-2}$). When stratifying these analyses by sex, exposure to silica was significant in men only ($p_{men} = 4.6 \times 10^{-3}$ and $p_{women} = 0.48$). Finally, elevated IgA levels were associated with an older age ($p = 5.3 \times 10^{-5}$), being a man ($p = 5.5 \times 10^{-3}$) and non-smoking ($p = 2.1 \times 10^{-2}$).

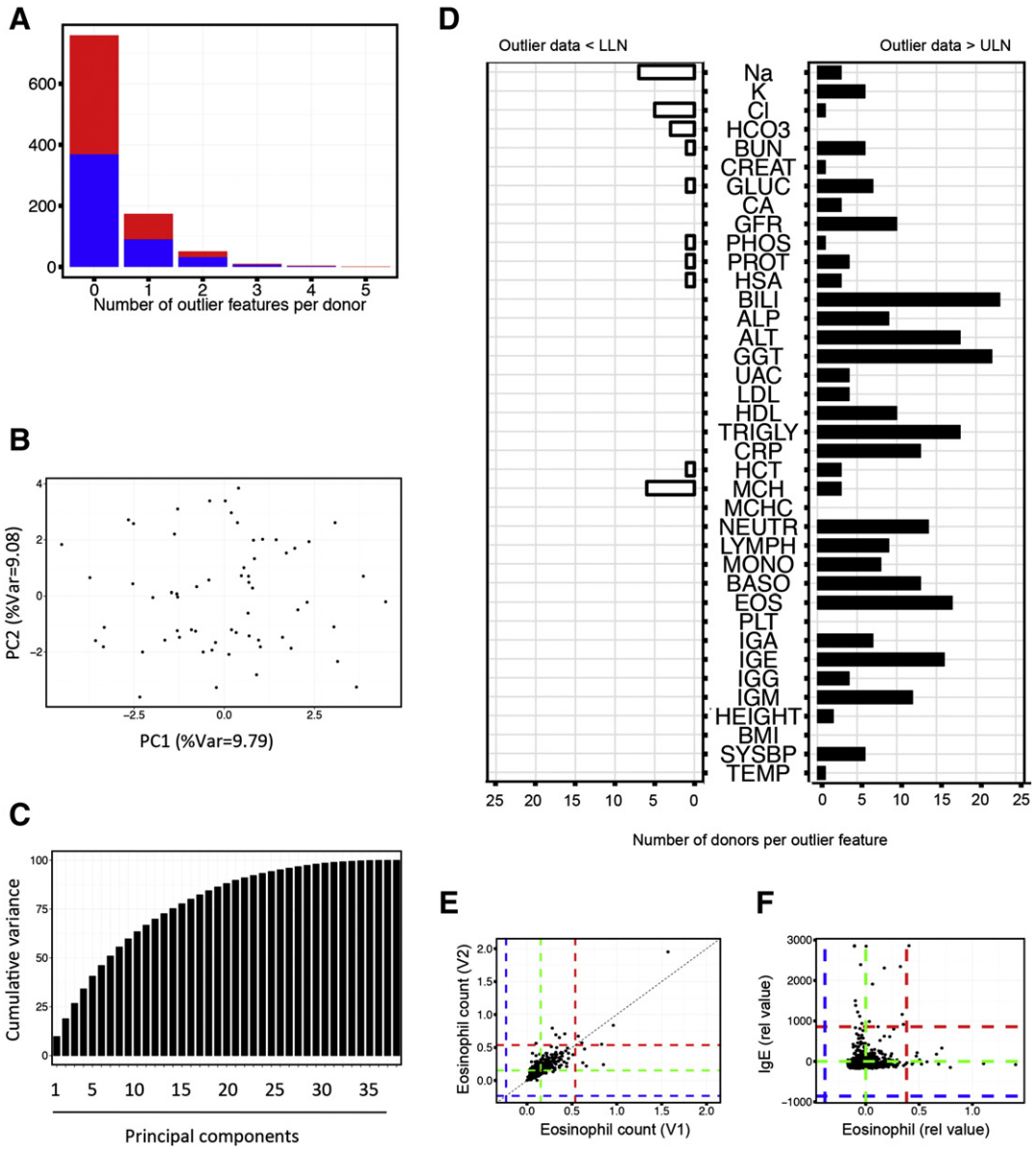


Fig. 5 Outlier data maps primarily to liver function tests and complete blood count measurements, yet shows now underlying structure based on measured variables. (A) For each donor, and for each measurement, an outlier status was assessed based on a z-score criteria. An aggregated score was computed for each donor to represent the number of times a given donor had been flagged as an outlier. The distribution of aggregated outlier cases among our cohort is shown. Colors indicate sex (men, blue; women, red). (B, C) A Principal Component Analysis (PCA) on 62 donors presenting >2 outlier cases has been performed. The scatterplot shows the projection of the donors onto the plane composed of the first 2 principal components, capturing 18.8% of the variance (B). The cumulative variance from the principal components is represented (C). (D) For each measurement considered in our analysis, we represented the number of donors that had been flagged as an above-the-range (on the right), or below-the-range outlier (on the left). (E) The absolute number of eosinophils is represented for V1 and V2, shown as representative data for the measured variables. Values have been regressed-out to take into consideration age and sex effects. Individual donors are represented by a black dot. A dotted line depicts the theoretical ideal correlation between the 2 visits. Dotted blue and red lines show the lower and upper threshold, respectively, as defined by our z-score based outlier detection. The green line shows the mean measurement. (F) The relationship between the eosinophil count and IgE level is shown. Both values have been regressed-out to take into consideration age and sex effects. Individual donors are represented by a black dot. Blue and red dotted lines show the lower and upper threshold, respectively, as defined by our z-score based outlier detection. The green line shows the mean measurement.

3.5. Outlier phenotypes showed independence among the measured variables

Despite the stringent criteria used for the recruitment of healthy donors (Table S1), we observed donors presenting extreme values within the observed range of biological measures. We identified 241 donors (24.1%) with clinical laboratory values that were outliers with respect to at least one variable, as defined by z-score based criteria (Fig. 5A). In only 66 individuals (6.6%), we observed two or more outlier events (Fig. 5A). To assess possible structure among the outlier events, we analyzed the data from those 66 individuals with outlier values for one or more laboratory tests, and projected the data using principal component analysis (PCA) (Fig. 5B). The dataset showed a lack of structure, which could also be observed by the broad distribution of variance across the top 35 component axes (Fig. 5C).

To interrogate the variables for which donors had outlier events, we plotted the number of donors per feature (Fig. 5C). Interestingly, liver function tests (e.g., ALT, GGT, BILI) and circulating immune cell counts (e.g., EOS, NEUTR, BASO) were highly represented among the feature space. We also observed higher numbers of donors (>15) with outlier TRIGLY and IGE levels. Selected variables were re-tested in the 500 donors sampled at V2, allowing the evaluation of repeatability. As shown, 4 of 8 (50%) of the donors with elevated numbers of EOS during V1 also showed higher levels at V2 (Fig. 5E). These data reinforce the added value of repeat testing for spurious outlier clinical laboratory data, but may also indicate the impact of environmental determinants on transient biochemical or cell number elevations. Finally, we investigated a possible association between EOS and IgE concentrations, as both are associated with allergic phenotypes. In support of the conclusions of the PCA, EOS count and IgE concentrations showed no correlation among healthy donors.

Our findings collectively help to define and validate the constitution of a healthy reference population, which will serve as a foundation for understanding and quantify the extent to which phenotypic variation in immune responses is under genetic or environmental control.

4. Discussion

The immune system is responsible for maintaining a healthy state, preventing infection and maintaining homeostasis. For some individuals, however, immune dysfunction can occur and results in increased susceptibility to infections, inflammation, autoimmunity, allergy or even cancer. Moreover, such individual heterogeneity in the immune response may have a major impact on the likelihood to respond to therapy or the development of side effects secondary to vaccine administration. Most prior studies aiming to understand the extent to which variation in immune responses is associated with immunopathology *sensu lato* have taken a disease-based approach, from which considerable insight into immune mechanisms have been obtained. Nonetheless, to utilize this information in diagnosis and disease management, the definition of the baseline parameters for immune function across the human "healthy" population is required. To achieve this goal, the *Milieu Intérieur* Project aims to provide a foundation for defining perturbations in an individual's immune system responses.

The *Milieu Intérieur* clinical study was designed and performed in healthy volunteers to develop a diverse sample collection with wide ranging associated meta-data. Ultimately, the generation of genetic data (based on genome-wide genotyping and whole exome sequencing) and multiple phenotypes (molecular, cellular and organismal) in available samples of the study cohort will produce a rich data warehouse. This will allow data mining studies for associations and consequently increase our knowledge of the different factors involved in the regulation of immune responses. During the design of the clinical study, we encountered the challenge of defining the genuine meaning of being "healthy" according to rational and measurable parameters. As such, strict criteria for enrolling donors were established, taking into consideration both recruitment feasibility, and the statistical power provided by a 1000-persons study, covering 10 strata (segregated across sex and age, by decade). While some exclusion factors were easy to apply, such as chronic infections (e.g., HCV) or severe disease (e.g., cancer, autoimmunity, etc.), others were more challenging, such as the boundary for allergic individuals, those that are exposed to known toxins (e.g., cigarettes), and persons with presyndromic signs (e.g., hypertension). Although the use of reference values for hematological, biochemical and serology parameters, commonly accepted in the clinic to define the healthiness of an individual, was considered as inclusion/exclusion criteria, there was a concern about the potential loss of extreme phenotypes. We thus chose cut off values that might indicate the requirement for medical follow-up (e.g., liver enzyme concentrations > 3 × ULN). Factors affecting the immune system and/or the composition of microbiota were also considered, including pre-term birth, current and prior exposure to medical treatments (e.g., aspirin), or the use of homeopathic medicaments (e.g., essential oils). Ultimately, we settled on the allowance of parameters expected to be present in >5% of the sampled individuals, and excluded any condition that necessitated past or current medical treatment. Detailed personal and family medical histories were systematically recorded, and associated meta-data will be used to define genetic, immunologic, and enterotype associations; and/or to regress out potential confounding factors. We hope that this set of criteria will help the international community taking steps towards a consensus definition of a healthy status for immunologic studies.

In considering selection biases linked to cross-sectional population-based studies [13], we consider several potential sources. The primary sources of selection bias are selective survival with fixed exposure in time (e.g., older donor survival effect); and non-fixed exposure in time (e.g., smoking, diet, alcoholic intake, professional exposure are variable in time). With respect to the survival effect bias, we in fact see this as an opportunity, as evidence for an age-associated narrowing of immunologic, genetic and enterotype variation may point towards a core signature of healthy status. To address non-fixed exposure, we highlight that our complete questionnaires provide an overview of both current and past exposures/habits. An additional caveat is that the healthy volunteers were selected from a pre-existing donor database, curated by a Clinical Research Organization. These volunteers

may be more "health-conscious" than *one-off* volunteers. This recruitment strategy may also explain the higher percentage of out-of-work persons, and the higher level of education as compared to the local *Ille-et-Vilaine* population.

Following our initial validation of known associations of health and clinical laboratory/immunological parameters, we investigated correlates with the metabolic syndrome index score. Among the general adult population, it is estimated that 20–25% meet the criteria for having metabolic syndrome (index score ≥ 3), so the identification of risk factors is central to establishing public health initiatives. While metabolic syndrome has been carefully evaluated in the context of disease settings, few studies have investigated healthy donors for risk factors. Interestingly, in a "healthy" setting, our study revealed that sex (i.e., being a man), aging and active cigarette smoking are each independent risk factors for an elevated metabolic score index. While epidemiological data support our findings for men and age as associated risk factors, the evidence for smoking as an independently associated variable (i.e., measured after regressing out sex and age) has been so far controversial. Our data indicate that four of the six individual component biologic variables are not statistically different for smokers as compared to non-smokers, however the global score supported its association with metabolic syndrome. Previous published studies, focusing primarily on individuals presenting overweight and obesity, showed an additive effect for smoking as an associated risk factor [14]. Conversely, other studies have failed to detect such associations and at least one study conducted among Turkish women found a protective effect of smoking on metabolic syndrome [15]. This has been attributed in part to the use of different definitions of metabolic syndrome. A recent meta-analysis evaluated data from 13 prospective studies for which primary data was available (n = 56,691 participants overall), and in a dose-response analysis, active smoking habits was positively associated with risk of metabolic syndrome (pooled relative risk [RR] = 1.26, 95% CI: 1.1–1.44) [16]. Our results, which differ from previous studies in that they are based on healthy donors, provide additional support for their findings and are consistent with experimental data indicating that cigarette smoking modifies hormone levels (e.g., cortisol), which in turn may result in the establishment of a more "insulin-resistant" state or the increase in waist circumference, a result of deposition visceral fat mass [17].

Our investigations of the metabolic syndrome score in healthy donors also revealed a troubling sex-bias. With the sole exception of HDL levels, the reference values for women and men have differing cut-values. For abdominal circumference, the effects are dramatic with nearly 60% women being considered above the threshold value as compared to 27% for men. Given the wide application of the metabolic syndrome score since 2001 [18], we suggest that the threshold values be adjusted for sex-associated differences. Notably, this has been done in USA populations, however it must be considered whether the adjusted thresholds have been set based on a shared definition of health, or instead due to the epidemic of obesity that is currently raging in first-world countries. Indeed these indicators impact public health initiatives and treatment endpoints, and therefore must be properly calibrated and correlated to real endpoints of health and disease. It is our hope that the *Milieu Intérieur* project will contribute to the identification of genetic, enterotype and immune response associations to metabolic score and other health indicators, possibly leading to the innovation of personalized algorithms.

Cytomegalovirus (CMV) infection is one of the most common infections of the general population [19], with a seroprevalence of 43% in Europe [20] and 50% in the US [21]. CMV is known to be transmitted through direct contact with infected bodily fluids, including urine, saliva, tears, but also blood and semen. A large number of studies have evaluated risk factors for CMV infection among pregnant women, but few have studied CMV incidence in a well-defined healthy donor population. We confirm that being single is an important risk factor for men and women, due to an increased number of partners [22], while having children has no direct impact. This challenges the notion that CMV transmission often occurs from children to adults and we suggest instead that adult-to-adult transmission is more common. Longitudinal studies will be required to confirm these observations, and to confirm that exposure to children varies with marital status. We identified another factor of interest: the consumption of raw vegetables. To our knowledge, this habit has not been previously described as a risk factor for CMV infection, and challenges the view that CMV is mainly transmitted by direct contact. Interestingly, recommendations to prevent infection usually include the avoidance of food sharing with young children [23]. While this might suggest that indirect contact is a risk factor, there had been little evidence to support this public health measure. One study has evaluated the duration of CMV viability on environmental surfaces and found that the virus could remain viable for 6 h on wet surfaces, including crackers [24]. Together, our observation supports that CMV transmission from food sharing has been underestimated and should be thoroughly evaluated in order to adapt preventive behaviors.

Interestingly elderly donors in our cohort showed a higher percentage (44%) of CMV negative individuals than previously reported studies (20–30% CMV- for >60 years old) [20]. This likely reflects the "healthy" status of our donors, as defined by stringent inclusion criteria, in contrast with previous studies that were performed on the general population. This is also reflected across the entire cohort, which was 64.8% CMV-, higher than other reported studies in both Europe (43%) [20] and the US (50%) [21]. Identifying host and environmental factors that may lead to increased resistance to CMV infection throughout life could have major implications for cardiovascular disease, sepsis and healthy aging.

To summarize, our study outlined herein provides an initial overview of the *Milieu Intérieur* cohort, which we believe constitutes a rich source of information and materials that will ultimately help to characterize and define topics relating to immunity, genetics, environment and lifestyle behaviors.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.clim.2014.12.004.

Conflict of interest statement

The authors declare that there are no conflicts of interest.

Acknowledgments

This work benefited from the support of the French government's "Invest in the Future program", managed by the Agence Nationale de la Recherche (ANR, reference 10-LABX-69-01). We acknowledge the advice of the "healthy donor" working group that helped us to define the inclusion/exclusion criteria: Laurent Abel, François Aubin, Raphaëlle Bourdet-Sicard, Ingrid Callies, Miguel Fenoy, Bridget Holmes, Laurence Mathivet, Cédric Ménager, Stéphane Reynier, Manfred Schmoltz, Muriel Vray). Scientific advisory board members that supported the trial design discussions included: Hyam Levitsky, Philip Greenberg, John Marioni, Jonathan Braun, Adrian Hayday and Thomas Joos. We also thank Dusko Ehrlich and Joël Doré for advice on stool sample collection procedures. We would like to acknowledge all the Biotrial team of doctors, nurses, technicians and project managers for their expertise and professionalism. We also thank the donors for their contribution to the study.

References

[1] E.K. Nelson, B. Piehler, J. Eckels, A. Rauch, M. Bellew, P. Hussey, S. Ramsay, C. Nathe, K. Lum, K. Krouse, D. Stearns, B. Connolly, T. Skillman, M. Igra, LabKey Server: an open source platform for scientific data integration, analysis and collaboration, BMC Bioinformatics 12 (2011) 71.

[2] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[3] A. Gulsvik, M.K. Fagerhoi, Smoking and immunoglobulin levels, Lancet 1 (1979) 449.

[4] S.A. McMillan, J.P. Douglas, G.P. Archbold, E.E. McCrum, A.E. Evans, Effect of low to moderate levels of smoking and alcohol consumption on serum immunoglobulin concentrations, J. Clin. Pathol. 50 (1997) 819–822.

[5] H.A. Schwertner, Association of smoking and low serum bilirubin antioxidant concentrations, Atherosclerosis 136 (1998) 383–387.

[6] M.R. Smith, A.L. Kinmonth, R.N. Luben, S. Bingham, N.E. Day, N.J. Wareham, A. Welch, K.T. Khaw, Smoking status and differential white cell count in men and women in the EPIC-Norfolk population, Atherosclerosis 169 (2003) 331–337.

[7] M. Tsuchiya, A. Asada, E. Kasahara, E.F. Sato, M. Shindo, M. Inoue, Smoking a single cigarette rapidly reduces combined concentrations of nitrate and nitrite and concentrations of antioxidants in plasma, Circulation 105 (2002) 1155–1157.

[8] M.M. Rahman, I. Laher, Structural and functional alteration of blood vessels caused by cigarette smoking: an overview of molecular mechanisms, Curr. Vasc. Pharmacol. 5 (2007) 276–292.

[9] R.S. Padwal, A.M. Sharma, Prevention of cardiovascular disease: obesity, diabetes and the metabolic syndrome, Can. J. Cardiol. 26 (Suppl. C) (2010) 18C–20C.

[10] K.G. Alberti, R.H. Eckel, S.M. Grundy, P.Z. Zimmet, J.I. Cleeman, K.A. Donato, J.C. Fruchart, W.P. James, C.M. Loria, S.C. Smith Jr., Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity, Circulation 120 (2009) 1640–1645.

[11] P.W. Wilson, R.B. D'Agostino, H. Parise, L. Sullivan, J.B. Meigs, Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus, Circulation 112 (2005) 3066–3072.

[12] S.L. Klein, A. Hodgson, D.P. Robinson, Mechanisms of sex disparities in influenza pathogenesis, J. Leukoc. Biol. 92 (2012) 67–73.

[13] M. Delgado-Rodriguez, J. Llorca, Bias, J. Epidemiol. Community Health 58 (2004) 635–641.

[14] H. Cena, M.L. Fonte, G. Turconi, Relationship between smoking and metabolic syndrome, Nutr. Rev. 69 (2011) 745–753.

[15] A. Onat, H. Ozhan, A.M. Esen, S. Albayrak, A. Karabulut, G. Can, G. Hergenc, Prospective epidemiologic evidence of a "protective" effect of smoking on metabolic syndrome and diabetes among Turkish women–without associated overall health benefit, Atherosclerosis 193 (2007) 380–388.

[16] K. Sun, J. Liu, G. Ning, Active smoking and risk of metabolic syndrome: a meta-analysis of prospective studies, PLoS One 7 (2012) e47791.

[17] K. Linder, F. Springer, J. Machann, F. Schick, A. Fritsche, H.U. Haring, G. Blumenstock, M.B. Ranke, N. Stefan, G. Binder, S. Ehehalt, Relationships of body composition and liver fat content with insulin resistance in obesity-matched adolescents and adults, Obesity (Silver Spring) 22 (2014) 1325–1331.

[18] S.M. Grundy, H.B. Brewer Jr., J.I. Cleeman, S.C. Smith Jr., C. Lenfant, Definition of metabolic syndrome: report of the National Heart, Lung, and Blood Institute/American Heart Association conference on scientific issues related to definition, Circulation 109 (2004) 433–438.

[19] J. Bodurtha, S.P. Adler, W.E. Nance, Seroepidemiology of cytomegalovirus and herpes simplex virus in twins and their families, Am. J. Epidemiol. 128 (1988) 268–276.

[20] P.R. Lubeck, H.W. Doerr, H.F. Rabenau, Epidemiology of human cytomegalovirus (HCMV) in an urban region of Germany: what has changed? Med. Microbiol. Immunol. 199 (2010) 53–60.

[21] S.L. Bate, S.C. Dollard, M.J. Cannon, Cytomegalovirus seroprevalence in the United States: the national health and nutrition examination surveys, 1988–2004, Clin. Infect. Dis. 50 (2010) 1439–1447.

[22] K.B. Fowler, R.F. Pass, Sexually transmitted diseases in mothers of neonates with congenital cytomegalovirus infection, J. Infect. Dis. 164 (1991) 259–264.

[23] G.J. Demmler-Harrison, Congenital cytomegalovirus: public health action towards awareness, prevention, and treatment, J. Clin. Virol. 46 (Suppl. 4) (2009) S1–S5.

[24] J.D. Stowell, D. Forlin-Passoni, E. Din, K. Radford, D. Brown, A. White, S.L. Bate, S.C. Dollard, S.R. Bialek, M.J. Cannon, D.S. Schmid, Cytomegalovirus survival on common environmental surfaces: opportunities for viral transmission, J. Infect. Dis. 205 (2012) 211–214.

[25] E. Kesse-Guyot, V. Andreeva, K. Castetbon, M. Vernay, M. Touvier, C. Mejean, C. Julia, P. Galan, S. Hercberg, Participant profiles according to recruitment source in a large Web-based prospective study: experience from the Nutrinet-Sante study, J. Med. Internet Res. 15 (2013) e205.

[26] D. Duffy, V. Rouilly, V. Libri, M. Hasan, B. Beitz, M. David, A. Urrutia, A. Bisiaux, S.T. Labrie, A. Dubois, I.G. Boneca, C. Delval, S. Thomas, L. Rogge, M. Schmolz, L. Quintana-Murci, M.L. Albert, Functional analysis via standardized whole-blood stimulation systems defines the boundaries of a healthy immune response to complex stimuli, Immunity 40 (2014) 436–450.

[27] RC Team, R: A Language and Environment for Statistical Computing, in, R Foundation for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[28] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer, New York, 2009.