A Comprehensive Prospective Cohort in Preventive Medicine: Protocol and Profile of the First 1,000 Participants in a Health Screening Program

Pierre Bauvin 1, Alaedine Benani 1/2/3, Maryne Lepoittevin 1, Marguerite Sentilhes 1, Marie Bringer 4, Djihane Ahmed Lecheheb 1, Nina De Luca 1, Stéphane Ohayon 1, Claude Dalle 1/†, Xavier Tannier 3, Eric Vibert 5, Philippe Gabriel Steg 6, Sylvain Bodard 7/8/9/10

1. Preventive Medicine, Data Science and AI Lab, Zoī, F-75010, Paris, France,
2. Service de médecine vasculaire, hôpital européen Georges Pompidou (HEGP), AP-HP, Université Paris-Cité, Paris, France.
3. Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Limics, 75006 Paris, France.
4. Groupe Hospitalier Universitaire (GHU) Paris - Psychiatrie et Neurosciences, Hôpital Sainte-Anne 1 rue Cabanis, 75014 Paris, France
5. Université Paris-Saclay, INSERM 1193 et AP-HP, Hôpital Paul Brousse, Paris, F
6. Université Paris-Cité, INSERM 1148 et AP-HP, Hôpital Bichat, Paris, F
7. Memorial Sloan Kettering Cancer Center (MSK), Department of Radiology, Cornell University, 1275 York Avenue, New York, NY 10065, USA
8. Transplantation Sciences, Massachusetts General Hospital, Harvard Medical School, Boston, USA
9. Université de Paris Cité, Hôpital Universitaire Necker Enfants Malades, Service d'Imagerie Adulte, AP-HP, F-75015, Paris, France sylvain.bodard@aphp.fr
10. CNRS UMR 7371, INSERM U 1146, Laboratoire d'Imagerie Biomédicale (LIB), Sorbonne Université, F-75006, Paris, France

† Deceased

Acknowledgments

The Zoī Preventive Medicine, Data Science and AI Lab is funded by Zoī SAS, Paris, France. The authors also express their thanks to Kévin David-Girard, Marion De Meslon, Damien Grosgeorge, Nassim Hmidou, Alexandre Hollard, Paul Roujansky, Léon-Marie de Souza, and Élodie Vorkaufer.

We pay tribute to Dr. Claude Dalle, whose vision and commitment were foundational to this protocol. A pioneer in the field of preventive medicine, he profoundly shaped the scientific thinking that led to this work.

Abstract

The Zoī cohort is a prospective longitudinal cohort study, designed to advance evidence-based personalized prevention, by systematically screening for undiagnosed or asymptomatic conditions, identifying early risk markers, and predicting future disease risks. Conducted within a dedicated prevention-focused health center, data collection takes place in a standardized environment and combines over 500 self-reported items, clinical examinations, extensive biomarker profiling (196 biomarkers), and multimodal imaging (vascular, breast, abdominal, and pelvic ultrasound, full-body composition, retinal scan). For several major diseases, risk is further estimated through established clinical prediction models. Longitudinal follow-up is collected via yearly re-evaluations and through a dedicated application.

This manuscript presents the cohort design and the characteristics of the first 1,000 participants. Participants (67.5% male, mean age 51.1 years, high education levels) exhibited a high level of health awareness, lower obesity and smoking rates than the general population, yet almost half (45.6%) of those who reported no known ongoing diseases had at least one undiagnosed chronic condition (i.e., either disease or risk factor), with hypertension and hypercholesterolemia being the most frequent. Male sex and older age were significantly associated with disease unawareness ($p<0.05$). These findings highlight a discrepancy between self-reported and objectively measured health status, even among a well-educated and health-conscious cohort.

This deeply phenotyped, longitudinal cohort will serve as a platform that supports interdisciplinary research collaborations. It will enable development and validation of early risk stratification models, essential for predictive medicine, as well as evaluation of preventive interventions, to advance evidence-based precision prevention in public health settings.

Keywords: Cohort, Preventive medicine, Personalized, Prevention

Why was the cohort set up?

Chronic diseases are a leading cause of morbidity and mortality in high-income countries, prompting a progressive shift from reactive to preventive healthcare models [1]. This evolution is exemplified by the Predictive, Preventive, Personalized, and Participatory (4P) medicine framework [2], which leverages advanced data analytics and individualized care strategies to anticipate health risks before the onset of disease and to support proactive health management [3,4].

In France, investment in preventive healthcare has historically been below the European average (2.8% vs. 3.2% of total health expenditure in 2020) [5,6]. This limited investment may have contributed to a sustained burden of preventable conditions such as diabetes, hypertension, obesity and lifestyle-related cancers [7]. In recent years, national health strategies have sought to reverse this trend by expanding preventive infrastructures, including digital tools and dedicated screening centers supported by both public and private initiatives, within the context of France's universal health coverage system, which combines a single public payer with complementary private insurance schemes [8–10]. Yet, important challenges remain in embedding prevention into routine care, including inconsistent follow-up, low adherence, and variability in the quality of risk assessment and screening [11].

Several large-scale epidemiological cohorts have generated valuable insights, such as CONSTANCES in France [12]. These cohorts, while foundational, often include participants who are more educated and in better health than the general population. The Zoī cohort shares this limitation but complements existing resources by implementing a highly standardized protocol and by associating self-reported questionnaires and annual comprehensive health check-ups: all participants undergo the same assessments under controlled conditions, with minimal missing data [13,14].

Moreover, follow-up intervals are often irregular or infrequent, or based only on electronic health records, limiting the ability to detect subclinical changes or generating biased effects [15,16]. Continuous or at least yearly follow-up—ideally supported by digital tools—remains uncommon, but greatly valuable, especially when integrating multimodal data [17].

Finally, while many cohort studies have focused on primary or secondary prevention among rural or socioeconomically vulnerable populations, less attention has been paid to structured preventive health assessments among urban, health-literate individuals, especially in France [18–22]. Despite good healthcare access, these populations may still be exposed to under-recognized risk factors—such as chronic stress, air pollution, and lifestyle imbalance [23–25].

To address these gaps, the Zoī cohort was established as a prospective, open-ended study embedded in a high-throughput preventive health center in Paris. Zoī emphasizes an individualized and proactive approach. Participants undergo standardized, multimodal assessments rooted in 4P medicine—integrating clinical evaluation, imaging, advanced biomarker profiling, and continuous digital follow-up—enabling precise, actionable health insights over time.

This article describes the design, methods, and initial findings from the first 1,000 participants enrolled in the cohort. It provides insight into the population's baseline health characteristics and highlights the potential of structured preventive assessments to detect early, undiagnosed risk conditions even in a health-aware, urban demographic.

Who is in the cohort?

The Zoī cohort is a cohort of voluntary participants, primarily French adult residents, undergoing extensive analysis with longitudinal follow-up. Participants included in the cohort were adults aged 18 years or older, enrolled in the Zoī program either as paying customers or as beneficiaries of a company-sponsored program, since November 2023. All participants were informed prior to inclusion, with the possibility to opt out of this analysis, in accordance with national regulations. At baseline, they were invited to complete a comprehensive health and behavioral questionnaire and to attend a health screening center (HSC) in Paris, France, where they underwent a morphological, clinical, radiographic and biological assessment, performed routinely as part of the check-up. Recruitment is ongoing with no predefined end date. The Paris center is projected to have the capacity to conduct up to 2,000 check-ups per month.

A sample of the first 1,000 participants of the cohort is included in the present cross-sectional analysis. 50.4% of them were paying customers, and 49.6% are beneficiaries of a company-sponsored program. **Table 1** provides an overview of the demographic, anthropometric, and health-related characteristics of the sample. To contextualize the cohort within the broader population, comparative analyses were conducted using publicly available data from French national epidemiological cohorts [26–29], including CONSTANCES, a nation-wide health cohort [12]. We also used recent studies from the EPIC cohort for European context [30,31].

Table 1. Demographic, anthropometric, and health-related characteristics of the sample with comparisons to the French general population, the CONSTANCES cohort and the European EPIC cohort.

	Zoī cohort (N=1000)	**French general population**	**French cohort (CONSTANCES) (N=57,922)**	**European cohort (EPIC) (N=428,728)**
Sex				
Female	325 (32.5%)	52.3%*	53.9%	70.8%
Male	675 (67.5%)	47.7%*	46.1%	29.2%
Age				
Mean (SD)	51.1 (11.6)	49.1*	-	51.1 (9.8)
Median [Min, Max]	51.3 [18.0, 91.4]			51.5 [18, 99]
Education level				

Primary education (before high school)	37 (3.7%)	-	27.4%	27.4%
Secondary education (high school degree)	59 (5.9%)	-	16.6%	20.9%
Undergraduate degree	133 (13.3%)	-	23.5%	23.1%
Higher education (graduate or above)	578 (57.8%)	-	30.5%	24.2%
Missing or don't want to answer	193 (19.3%)	-	2.0%	3.7%
Self-rated low health				
Yes	156 (15.6%)	31.5%‡	20.1%	-
Self-declard alcohol abstinent				
Abstinent	94 (9.4%)	-	18.3%	12.5%
Smoking				
Active	115 (11.5%)	31.9%**	-	28.4%
Former	229 (22.9%)		28.6%	25.7%
Never	656 (65.6%)	-	-	42.8%
Body Mass Index (kg/m²)				
Mean (SD)	24.4 (4.0)	25.5*	-	25.3 (4.2)
Median [Min, Max]	24.0 [15.2, 48.9]	-	-	24.7 [10.2, 77.9]
Body Mass Index status ¶				
Normal weight or underweight	607 (60.7%)	53.7%*	-	54.5%
Overweight	314 (31.4%)	30.3%*	-	33.8%

Obesity	79 (7.9%)	17.0%*	10.1%	11.6%
Obesity class 1	60 (6.0%)	11.9%*	-	-
Obesity class 2	13 (1.3%)	3.1%*	-	-
Obesity class 3	6 (0.6%)	2.0%*	-	-
Diagnosed diabetes				
Yes	26 (2.6%)	5.7%†	2.2%	2.4%

Abbreviations: SD, standard deviation; EPIC, European Prospective Investigation Into Cancer and Nutrition;

** From [29]*
*** From [27]*
† From [26]
‡ From [28]
¶ *According to ethnicity-specific thresholds*

The cohort was predominantly male, with 325 (32.5%) women, contrasting with national data where women constitute a majority of the population (52.3% in the French general population and 53.9% in CONSTANCES). The mean age was 51.1 years (SD = 11.6, range: 18–91 years), slightly higher than that of the French general population (49.1 years) but similar to that of the EPIC cohort (51.1 years).

Regarding educational status, the cohort was highly educated, with 578 participants (57.8%) holding a graduate degree or above, compared to 30.5% in CONSTANCES. Only 3.7% reported having no education beyond primary school, substantially lower than the 27.4% observed in CONSTANCES.

Health-related factors revealed notable differences with national data. The mean BMI (Body Mass Index) was 24.4 kg/m² (SD = 4.0), slightly lower than the French general population (25.5 kg/m²). Obesity prevalence was markedly lower in the cohort (7.9%) compared to the other cohorts (French general population, 17.0%; CONSTANCES, 10.1%; EPIC, 11.6%). Diagnosed diabetes was also less frequent (2.6%) than in the general population (5.7%) but similar to the CONSTANCES and EPIC cohort (2.2% and 2.4%, respectively).

Active smoking was reported in only 11.5% of participants, considerably lower than the 31.9% prevalence reported in the French general population. Meanwhile, 22.9% were former smokers.

Self-reported health perception was relatively positive, with only 15.6% of participants rating their health as "low" or "fair," which is notably lower than the 31.5% reported in the French general population and 20.1% in CONSTANCES. Additionally, 9.4% of participants self-declared as alcohol abstinent, which is lower than the 18.3% reported in CONSTANCES.

No differences were observed between paying customers and beneficiaries of a company-sponsored program, regarding any of the previous demographic, anthropometric, and health-related characteristics (Supplementary Table

1). Notably, no differences were observed regarding proportion of alcohol abstinent, smokers, participants with obesity or with diabetes (respectively, p=0.621, p=0.670, p=0.50 p=0.786).

How often are participants followed up?

Following each clinical assessment, participants could receive individualized recommendations of interventions, encompassing pharmacological treatments, dietary modifications, physical activity plans, and other health-promoting behaviors, as well as therapeutic education aimed at fostering long-term self-management skills. When appropriate, referrals to medical specialists were also provided for further evaluations. Annual health reassessments are made available to participants in this cohort to track health trajectories and intervention efficacy. In between annual reassessments, participants were followed longitudinally via a dedicated smartphone application, which facilitated self-reporting at regular intervals of lifestyle factors, medication and supplement usage, and engagement in physical activities such as walking and meditation.

The application also served as a behavioral intervention tool to enhance adherence to preventive strategies. To promote sustained adherence, the application provided reminders, prompts and actionable lifestyle modifications, and employed dedicated algorithms intended to integrate the preventive measures into participants' daily routines.

The health reassessments and follow-up data will be addressed in a subsequent study, as only the first participants have begun their follow-up visits to date.

What is measured?

Extensive data were systematically collected from multiple sources: self-reported questionnaires, clinical examination, medical devices (electrocardiogram, spirometry), laboratory tests (urine, blood, saliva), and medical imaging, ensuring a broad assessment of health status. The protocol was designed to accommodate future scientific developments, ensuring its continued relevance and applicability.

A structured health and behavioral questionnaire was administered to all participants, encompassing sociodemographic characteristics, medical and family history, environmental exposures, dietary habits, physical activity, mental health, sleep patterns, sexual activity and other relevant domains. Standardized scales, such as the Epworth Sleepiness Scale or the STOP-BANG questionnaire, were incorporated for sleep assessment [32,33]. A detailed description of the current version of the questionnaire is provided in **Supplementary Table 2**.

During clinical examination, physical and functional parameters were recorded, including blood pressure, heart rate, heart rate variability, waist and neck circumference, and visual acuity. Systematic assessments included spirometry (Spirobank II Smart), 12-lead electrocardiography (PCA 500, QT Medical), hand grip strength testing (K-Grip, Kinvent), audiometry, and non-invasive evaluation of advanced glycation end products (AGEs) using skin autofluorescence (AGE Reader, Diagnoptics) [34].

Biological sampling included saliva, urine, and blood specimens. Saliva samples were self-collected upon awakening, 30 minutes post-awakening, and at nighttime. Urine samples included first morning void and 24-hour

urine collections, while blood samples were obtained by a nurse at the HSC visit. A total of 196 biomarkers were analyzed, including standard metabolic markers and more advanced measures such as oxidative stress and neurotransmitter evaluation [35]. A full list of biomarkers is available in **Supplementary Table 3**.

Medical imaging assessments included non-invasive modalities such as vascular, breast, abdominal, and pelvic ultrasounds (ACUSON Juniper, Siemens), dual-energy X-ray absorptiometry (DEXA) for bone density and body composition (Lunar iDXA, GE Healthcare), cone beam computed tomography (CBCT) for lung, sinus, and dental assessments (NewTom 7G, NewTom), and optical coherence tomography (OCT) for retinal and optic nerve evaluation (Mirante, Nidek France) [36]. All imaging results were systematically reviewed by expert radiologists and ophthalmologists.

Diagnoses

Diagnosed conditions encompassed both clinically diagnosed diseases and early risk markers. The latter were defined as measurable deviations from optimal health that do not meet clinical disease thresholds but may indicate an increased risk of future disease. These included: prediabetes, overweight, specific micronutrient deficiencies, fatty acid dysregulation, elevated inflammatory markers, and early sleep disturbances.

Most diagnoses were established based on clinical and biological measurements using internationally recognized criteria; e.g. diabetes and pre-diabetes were defined according to the American Diabetes Association (ADA) criteria [37], hypertension was classified based on the 2024 European Society of Cardiology (ESC) guidelines [38], hypercholesterolemia was determined using the American College of Cardiology (ACC)/American Heart Association (AHA) guidelines [35] based on cardiovascular risk profiles, obesity was derived from the recent The Lancet Diabetes & Endocrinology Commission via ethnicity-specific BMI cut-offs and excess adiposity direct measurement [40], etc. The complete list can be found in Supplementary Data. More complex diagnoses were established by MDs and extracted from medical records.

Clinical prediction models

Several validated clinical prediction models were retrospectively applied to the cohort data, focusing on disease prevention. Selection criteria prioritized externally validated models with low risk of bias and strong applicability to the study population and measurements [41]. Risk models included cardiovascular risk assessment tools (SCORE2, SCORE2-OP, SCORE2-Diab) [42–44], and predictive models for lung (LLPv3) [45], colorectal [46], liver [47], breast cancers [48], as well as type 2 diabetes [49].

Quality control

Standardized protocols were implemented across all clinical assessments to ensure consistency and reliability. Measurements were performed using the same equipment, with the same team of HCPs (Healthcare Professionals), under controlled environmental conditions, including lighting, sound, temperature and humidity of each room. Data collection was almost entirely automated, minimizing the risk of data entry errors and missing values. Standardized

tools and classifications, including SNOMED CT, LOINC, and FHIR-like data structures, facilitated data interoperability and broader applicability of the findings [50].

For research purposes, all participant data undergo pseudonymization [51]. For free-text data, pseudonymization is performed using a machine learning approach based on a standard token classification model for general-purpose named entity recognition (NER). This model consists of a Transformer followed by multiple constrained Conditional Random Fields (CRF) classification heads [51]. To ensure confidentiality, we use a publicly available model trained on synthetic, fictitious data[1]. Identifying information is replaced with plausible alternatives. Data access is strictly restricted, and confidentiality measures adhere to the highest ethical and security standards.

The current study followed the STANDING Together consensus recommendations to tackle algorithmic bias and promote transparency, specifically the Recommendations for the Documentation of Health Datasets (**Supplementary Table 3**) [52].

Can I get hold of the data?

The Zoī cohort is a novel resource for advancing research in predictive and preventive health, given its depth, standardization and structuration, following an open-ended design with an annual maximum capacity of tens of thousands. In alignment with the mission to democratize personalized medicine, the Zoī cohort welcomes scientific collaborations with research projects that align with the commitment to understand, develop and implement preventive and predictive healthcare strategies. Access to the Zoī cohort can only be provided at no cost to investigators, ensuring that financial barriers do not hinder scientific progress.

Hence, collaboration with Zoī is guided by strict adherence to legal and ethical standards, prioritizing data confidentiality and security, with rigorously validated tools that preclude the possibility of individual re-identification [53]. Proposals for research projects involving the Zoī cohort are to be reviewed by the Zoī Scientific and Medical Committee that evaluates missions based on scientific merit, and an independent Ethical Committee for ethical considerations.

What has it found? Insights from the first 1,000 participants

Biological measurements

Key biological parameters were assessed in the cohort, providing insights into renal function, lipid metabolism, and metabolic health.

As examples, the LDL cholesterol, the Glomerular Filtration Rate (GFR), uric acid levels, and apolipoprotein B are displayed in Figure 1.

[1] *https://github.com/aphp/eds-pseudo*

The mean LDL-cholesterol was 122.3 (±37) mg/dL, with a majority of participants, especially males, exceeding the European Society of Cardiology 2019 thresholds: 98.5% are above 55 mg/dL, 93.3% above 70 mg/dL and 68.8% are above 100 mg/dL (**Fig. 1A**) [54].

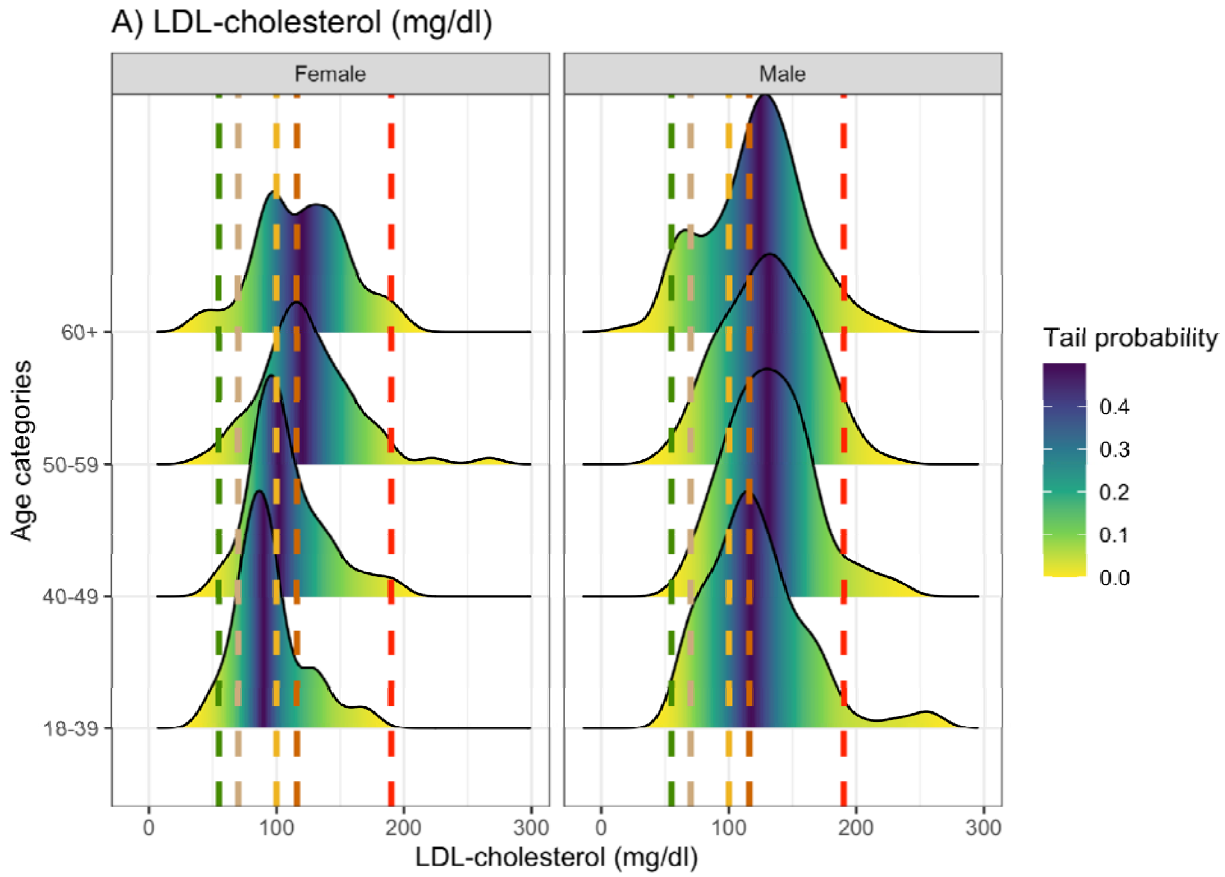
The mean GFR was 84.8 ± 9 mL/min, with 44.8% of participants below 90 mL/min, 2.2% below 60 mL/min and 0.7% below 45 mL/min, the KDIGO 2024 guidelines thresholds [55]. Age-stratified distributions indicated a progressive decline with age (**Fig. 1B**).

Uric acid levels showed a mean of 5.7 ± 1 mg/dL, with a substantial proportion (52.4%) of participants exceeding the URRAH Project threshold of 5.6 mg/dL, particularly in males and older age groups (**Fig. 1C**) [56].

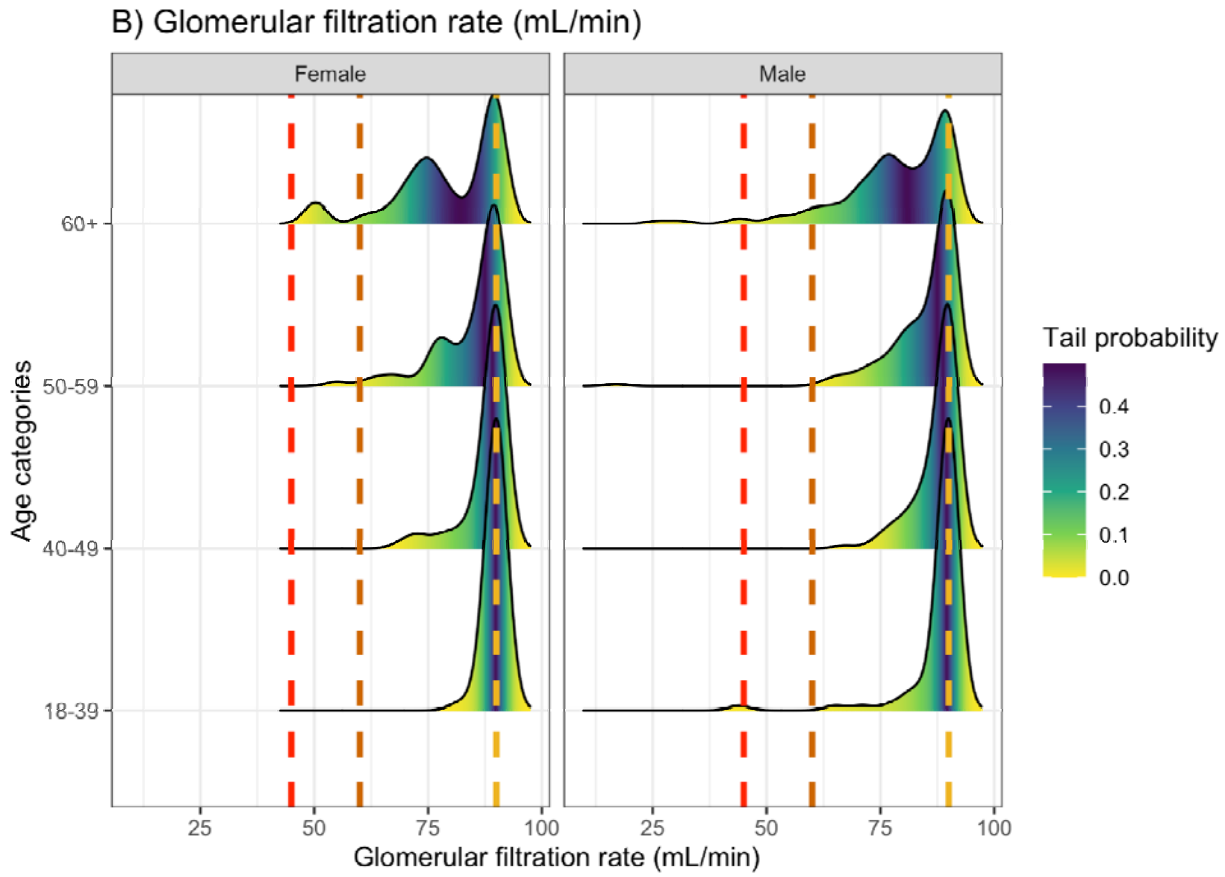
Apolipoprotein B (ApoB) had a mean concentration of 0.9 ± 0.2 g/L, with distributions indicating a shift towards higher values in men compared to women. Reference thresholds of 0.65, 0.85, and 1.0 g/L from the SEPHAR study were used for comparison (**Fig. 1D**) [57].

Additional measurements of interest, associated with cardiovascular and metabolic health, include Apolipoprotein A1 (mean 1.5 ± 0.2 g/L), fasting blood sugar (mean 87.5 ± 14.7 mg/dL), HbA1c (mean 5.3 ± 0.5 %), HOMA index (mean 1.2 ± 0.9), lipoprotein A (mean 25.3 ± 23.2, median 13.0 mg/dL). The complete list of all biomarkers is available in **Supplementary Table 2**.

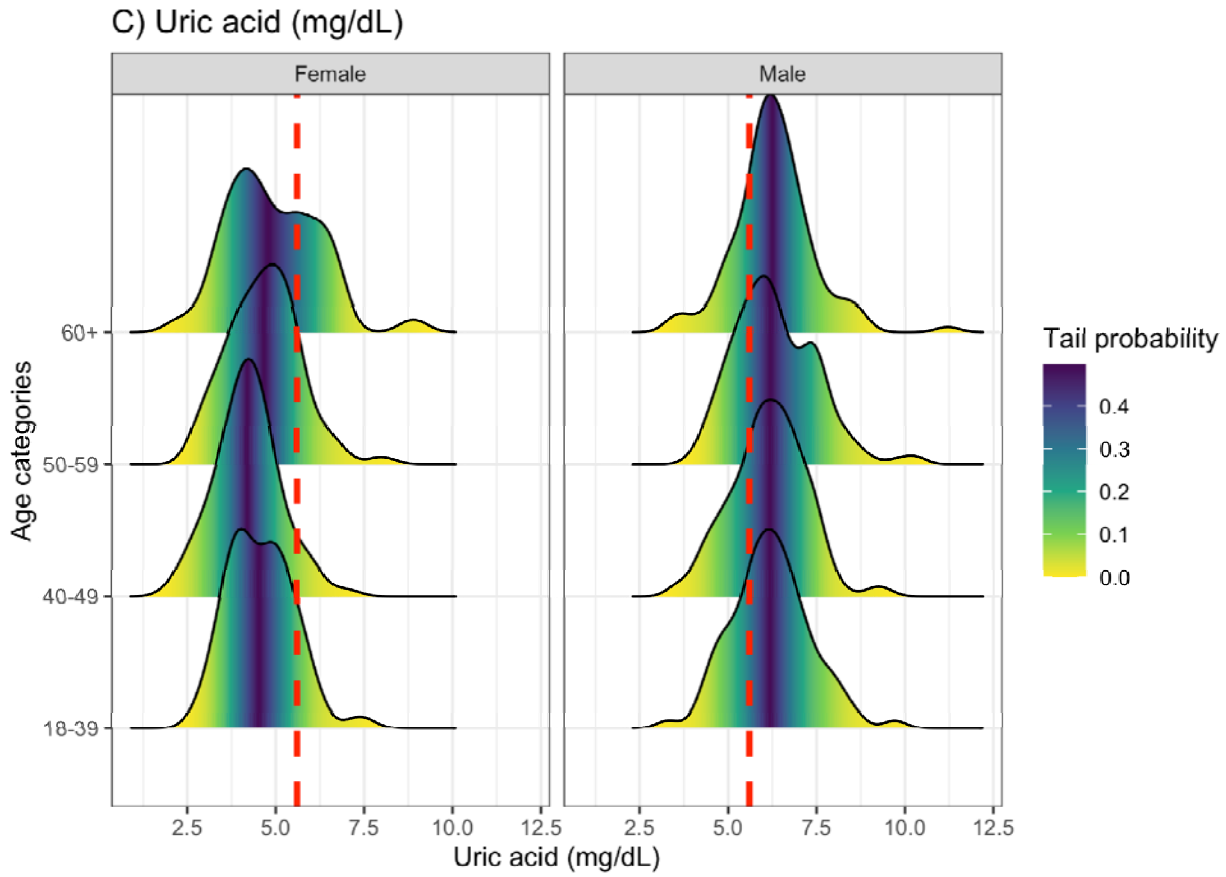
Fig. 1: Distribution of some specific health parameters measured in the sample, by age and sex: (A) LDL-cholesterol, (B) glomerular filtration rate, (C) uric acid, (D) Apolipoprotein B.



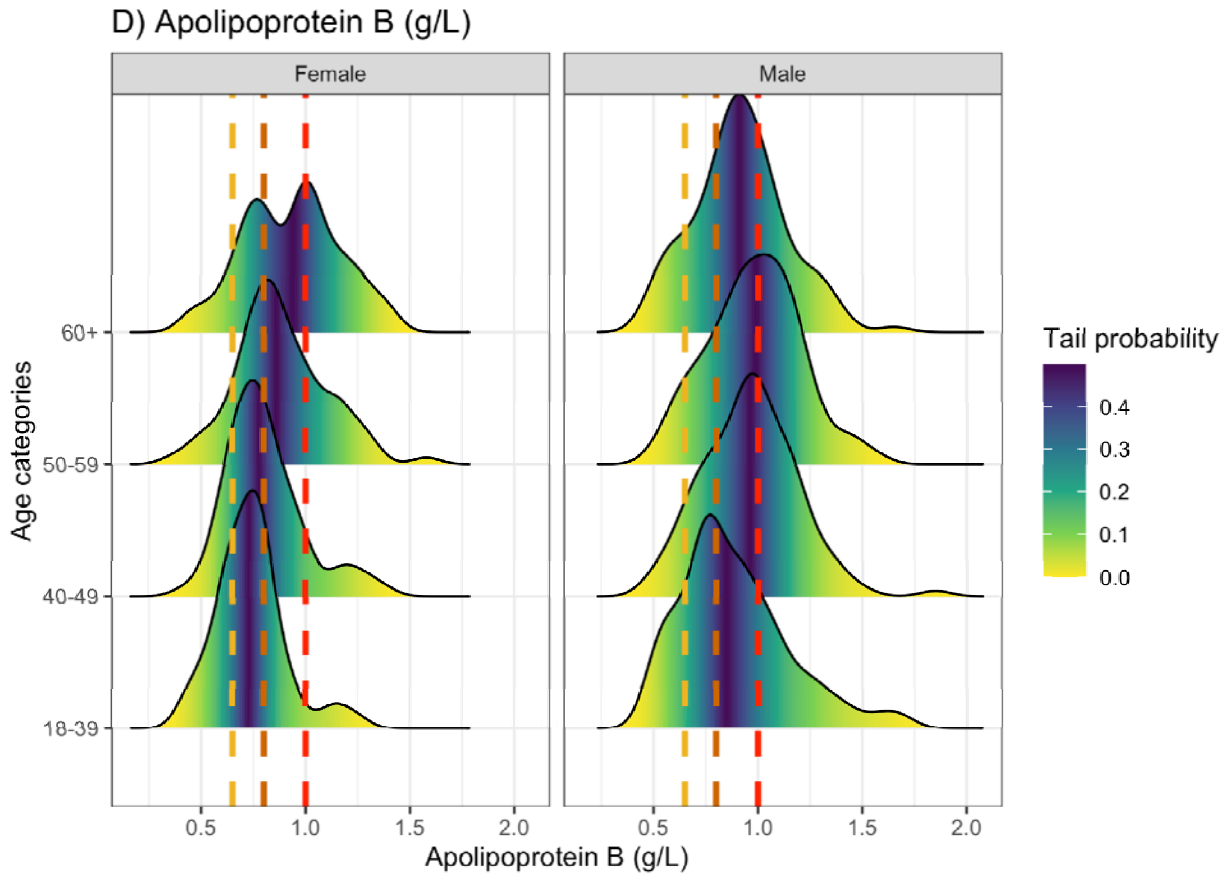
With thresholds corresponding to European Society of Cardiology 2019 guidelines: 55, 70, 100, 116, and 190 mg/dL.



With thresholds corresponding to KDIGO 2024 Clinical Practice Guideline: 45, 60 and 90 mL/min



With threshold corresponding to the URRAH Project: 5.6 mg/dL



With thresholds from the SEPHAR study: 0.65, 0.85, 1.0 g/L

Prevalence of conditions and comparison with self-reported health status

Almost all participants (n=905, 90.5%) were found to have at least one disease or early risk marker during the comprehensive health check-up, known beforehand or not: 56.9% with at least one disease, and 88.2% with at least one early risk marker. The most common conditions were two early risk markers: fatty acid dysregulation (69.8%) and iodine deficiency (61.6%). For comparison, hypertension was found in 22.6% of the participants, known beforehand or not.

Participants were asked if they had a known ongoing diagnosis of some chronic conditions: diabetes, obesity, hypertension, hypercholesterolemia, osteoporosis, renal deficiency, hyperthyroidism, sleep apnea, or MASLD (metabolic dysfunction-associated steatotic liver disease). Among participants who self-reported having no known such condition (n = 539), only 54.4% were confirmed as having none of these diagnoses upon clinical evaluation. The remaining 45.6% were found to have at least one previously undiagnosed chronic condition: 164 (30.4%) had one condition, 70 (13.0%) had two conditions, 11 (2.0%) had three conditions, and 1 (0.2%) had four conditions (**Table 2A**).

Table 2. Prevalence of chronic conditions in the cohort, (A) presence or absence of diagnoses among members specifically declaring the absence of said chronic condition, (B) number of chronic conditions measured among members self-reporting as healthy.

A) Presence or absence of diagnoses among members declaring the absence of said chronic condition

Chronic condition declared not diagnosed by participants*	**Confirmed absent**	**Identified during check-up**
Hypercholesterolemia (from ACC/AHA guidelines based on cardiovascular risk profiles)	220 (48.6%)	233 (51.4%)
Hypertension	491 (78.7%)	133 (21.3%)
Prediabetes	487 (86.4%)	78 (13.6%)
MASLD	707 (94.0%)	45 (6.0%)
Sleep apnea syndrome	306 (95.0%)	16 (5.0%)
Osteoporosis	495 (98.0%)	10 (2.0%)
Obesity	681 (98.1%)	13 (1.9%)
Diabetes	565 (98.4%)	9 (1.6%)
Renal deficiency	505 (98.6%)	7 (1.4%)
Biological hyperthyroidism	539 (99.6%)	2 (0.4%)
Declaring none of the above	293 (54.4%)	246 (45.6%)

Abbreviation: MASLD, Metabolic dysfunction-Associated Steatotic Liver Disease

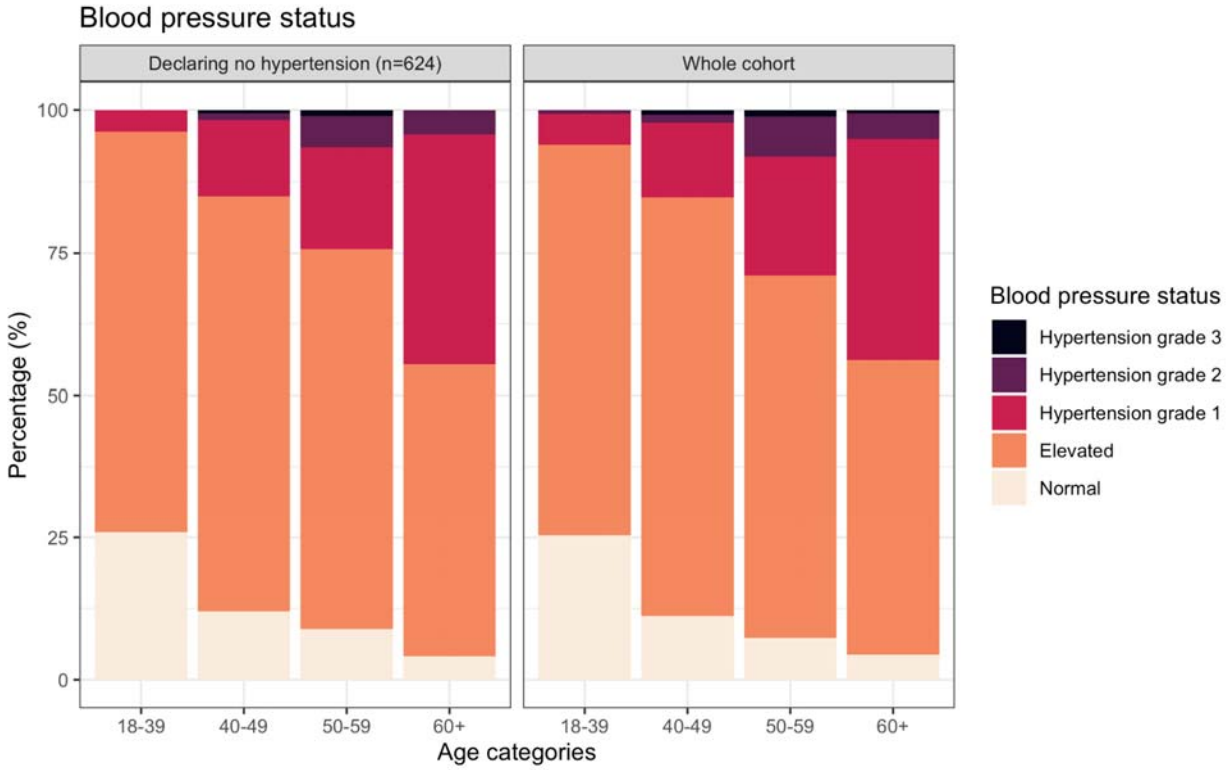
** Patients not responding to a question about the previous diagnosis of a disease were excluded from the corresponding analysis, to ensure that we only consider patients actively declaring no disease.*

B) Number of chronic conditions identified among participants self reporting as healthy

Number of actual conditions	Count
0	293 (54.4%)
1	164 (30.4%)
2	70 (13.0%)
3	11 (2.0%)
4	1 (0.2%)

Hypercholesterolemia (defined from ACC/AHA guidelines based on cardiovascular risk profiles) was frequently underdiagnosed, with 51.4% of participants who reported no prior diagnosis of hypercholesterolemia being newly identified with the condition (**Table 2B**). Hypertension was newly diagnosed in 21.3% of participants who had previously reported no history of the condition, distributed as follows: 17.9% in stage 1, 2.9% in stage 2, and 0.5% in stage 3 (**Fig. 2**).

Fig.2: Distribution of blood pressure status in the total population, and in the subgroup self-declaring no hypertension



Blood pressure categories correspond to the 2024 European Society of Cardiology (ESC) guidelines: normal is below 120/70 mmHg, elevated blood pressure is 120-139/70-89 mmHg, Grade 1 is 140-159/90-99 mmHg, Grade 2 is 160-179/100-109 mmHg and grade 3 is above 180/110 mmHg.

In a multivariable regression analysis assessing factors associated with unawareness of a current condition, while adjusting for potential confounders (age, sex, income level, and self-perceived health status), higher education levels were not significantly associated with unawareness of a current diagnosis ($p = 0.605$). However, male sex ($p < 1e-5$) and older age (1 year increment, $p < 1e-5$) were strongly associated with unawareness of condition.

Identification of suspicious lesions

A total of 136 patients (13.6%) were prescribed additional investigations to further characterize or monitor a potentially tumoral lesion. These complementary explorations were recommended based on initial imaging or biology findings or clinical suspicion. Follow-up data and final diagnostic outcomes will be the subject of a subsequent study.

What are the main strengths and weaknesses?

The Zoī cohort study presents several key strengths. Its comprehensive and standardized health assessment protocol enables detailed characterization of participants' health status. This includes an extensive array of data sources—self-reported questionnaires, clinical examinations, medical devices, laboratory analyses, imaging techniques, and follow-up measures—collected under controlled conditions. The analysis of a first sample has revealed a substantial mismatch between perceived and actual health status: among participants reporting no prior diagnoses, 45.6% were found to have at least one common condition during the check-up. This underscores the potential of structured assessments to uncover undiagnosed disease, even in populations with high access to healthcare.

Moreover, it offers valuable opportunities for machine learning applications in the health domain. While datasets derived directly from routine care are often larger and may better represent the general population, they are typically collected for clinical purposes. This difference in intent introduces distinct types of bias, such as selection bias due to care pathways or missing data patterns influenced by clinical practices. In contrast, the dataset described here is comprehensive in both its collection and processing, ensuring higher data quality. This completeness significantly enhances its utility for machine learning tasks, as models can be trained without requiring complex imputation strategies or introducing additional noise. Furthermore, the dataset's well-defined structure and consistent annotations facilitate the development of robust algorithms, making it particularly suited for benchmarking and model evaluation, beyond the simple bigger-is-better paradigm [58].

Another strength lies in the early detection of subclinical risk states. The assessment of early risk markers has proven to be a key component in identifying individuals at heightened risk for future disease onset [59], suggesting that even in the absence of overt disease, early deviations from optimal health may be detectable. Additionally, integrating a wide range of validated predictive scores could further refine risk stratification and inform predictive medicine strategies [60,61].

Additional strengths of the study include: dataset's completeness and quality control, longitudinal data collection, enabling the evaluation of predictive tools or preventive strategies, and adaptability of the protocol to incorporate future scientific advances.

However, the cohort has several limitations. The study population is predominantly middle-aged and highly educated, with lower smoking and obesity prevalence and better self-rated health than the general French population. While this may limit generalizability, it provides a valuable opportunity to examine health trajectories in a health-conscious subpopulation presumed to be well-informed and with high access to healthcare services.

Another limitation concerns the current state of the protocol regarding diseases and early risk markers assessment. Most diagnoses were derived from clinical and biological measurements using internationally recognized criteria. While this approach enables standardisation and ensures up-to-date diagnostic classification, it isn't suitable for more complex diseases that require access to detailed medical records. As a result, some diagnoses may be missed, or may necessitate manual, retrospective assessments by trained clinicians. Efforts are currently underway to improve the protocol by incorporating systematic disease reporting during check-ups by the team of HCPs, using the ICD-10 classification. This enhancement aims to reduce the number of missed diagnoses and lessen the reliance on retrospective evaluations.

As a longitudinal, deeply phenotyped cohort with ongoing data collection, the Zoī cohort is well-positioned to contribute to the evolving landscape of predictive medicine. Its design supports future collaborations, risk model development and evaluation, and policy-relevant insights into early intervention effectiveness.

References

1. De la Torre K, Min S, Lee H, Kang D. The Application of Preventive Medicine in the Future Digital Health Era. J Med Internet Res. 2025;27:e59165.

2. Vogt H, Hofmann B, Getz L. The new holism: P4 systems medicine and the medicalization of health and life itself. Med Health Care Philos. 2016;19:307–23.

3. Tuena C, Semonella M, Fernández-Álvarez J, Colombo D, Cipresso P. Predictive Precision Medicine: Towards the Computational Challenge. In: Pravettoni G, Triberti S, editors. P5 EHealth Agenda Health Technol Future [Internet]. Cham: Springer International Publishing; 2020 [cited 2025 Mar 10]. p. 71–86. Available from: https://link.springer.com/10.1007/978-3-030-27994-3_5

4. Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. Nat Med. 2022;28:666–77.

5. OECD, European Union. Health at a Glance: Europe 2022: State of Health in the EU Cycle [Internet]. OECD; 2022 [cited 2025 Mar 10]. Available from: https://www.oecd.org/en/publications/health-at-a-glance-europe-2022_507433b0-en.html

6. Raynaud D. Les dépenses de prévention, complexes à mesurer, très difficiles à comparer…: Regards. 2023;N° 61:55–66.

7. Elmadani M, Mokaya PO, Omer AAA, Kiptulon EK, Klara S, Orsolya M. Cancer burden in Europe: a systematic analysis of the GLOBOCAN database (2022). BMC Cancer. 2025;25:447.

8. Thacharodi A, Singh P, Meenatchi R, Tawfeeq Ahmed ZH, Kumar RRS, V N, et al. Revolutionizing healthcare and medicine: The impact of modern technologies for a healthier future—A comprehensive review. Health Care Sci [Internet]. 2024 [cited 2025 Mar 11];3:329–49. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11520245/

9. Policies EO on HS and, Or Z, Gandré C, Seppänen A-V, Hernández-Quevedo C, Webb E, et al. France: health system review [Internet]. World Health Organization. Regional Office for Europe; 2023 [cited 2025 Mar 11]. Available from: https://iris.who.int/handle/10665/371027

10. France's new Global Health Strategy [Internet]. [cited 2025 Mar 11]. Available from: https://www.who.int/news/item/12-10-2023-france-s-new-global-health-strategy

11. Mosadeghrad AM, Afshari M, Isfahani P, Ezzati F, Abbasi M, Farahani SA, et al. Strategies to strengthen the resilience of primary health care in the COVID-19 pandemic: a scoping review. BMC Health Serv Res. 2024;24:841.

12. CONSTANCES team, Zins M, Goldberg M. The French CONSTANCES population-based cohort: design, inclusion and follow-up. Eur J Epidemiol. 2015;30:1317–28.

13. Santoro C, Donal E, Magne J, Sade LE, Penicka M, Katbeh A, et al. Inter-center reproducibility of standard and advanced echocardiographic parameters in the EACVI-AFib echo registry. Echocardiogr Mt Kisco N. 2023;40:775–83.

14. Shaw PA, Deffner V, Keogh RH, Tooze JA, Dodd KW, Küchenhoff H, et al. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. Ann Epidemiol. 2018;28:821–8.

15. Kristman V, Manno M, Côté P. Loss to follow-up in cohort studies: how much is too much? Eur J Epidemiol. 2004;19:751–60.

16. Farmer R, Mathur R, Bhaskaran K, Eastwood SV, Chaturvedi N, Smeeth L. Promises and pitfalls of electronic health record analysis. Diabetologia. 2018;61:1241–8.

17. Toledano MB, Smith RB, Brook JP, Douglass M, Elliott P. How to Establish and Follow up a Large Prospective Cohort Study in the 21st Century--Lessons from UK COSMOS. PloS One. 2015;10:e0131521.

18. Ring KL, Duska LR. How far is too far? Cancer prevention and clinical trial enrollment in geographically underserved patient populations. Gynecol Oncol. 2024;184:8–15.

19. Gabet A, Lamarche-Vadel A, Chin F, Olié V. Disparités régionales de la mortalité prématurée par maladie cardiovasculaire en France (2008-2010) et évolutions depuis 2000-2002. Bull Épidémiologique Hebd [Internet]. 2014 [cited 2025 Jun 10]; Available from: http://www.invs.sante.fr/beh/2014/26/2014_26_1.html

20. Wang H, Roy S, Kim J, Farazi PA, Siahpush M, Su D. Barriers of colorectal cancer screening in rural USA: a systematic review. Rural Remote Health. 2019;19:5181.

21. Douthit N, Kiv S, Dwolatzky T, Biswas S. Exposing some important barriers to health care access in the rural USA. Public Health. 2015;129:611–20.

22. Wercholuk AN, Parikh AA, Snyder RA. The Road Less Traveled: Transportation Barriers to Cancer Care Delivery in the Rural Patient Population. JCO Oncol Pract. 2022;18:652–62.

23. Teo KK, Rafiq T. Cardiovascular Risk Factors and Prevention: A Perspective From Developing Countries. Can J Cardiol. 2021;37:733–43.

24. Tokgozoglu L, Hekimsoy V, Costabile G, Calabrese I, Riccardi G. Diet, Lifestyle, Smoking. Handb Exp Pharmacol. 2022;270:3–24.

25. Münzel T, Hahad O, Sørensen M, Lelieveld J, Duerr GD, Nieuwenhuijsen M, et al. Environmental risk factors and cardiovascular diseases: a comprehensive expert review. Cardiovasc Res. 2022;118:2880–902.

26. Lailler G, Piffaretti C, Fuentes S, Nabe HD, Oleko A, Cosson E, et al. Prevalence of prediabetes and undiagnosed type 2 diabetes in France: Results from the national survey ESTEBAN, 2014–2016. Diabetes Res Clin Pract. 2020;165:108252.

27. Obradovic I, Pasquereau A, Spilka S, Andler R, Douchet M-A, Guignard R, et al. État des lieux de la situation du tabagisme en France : tendances observées et nouveaux enjeux: Santé Publique. 2023;Vol. 35:19–33.

28. Lahbib H, Carcaillon-Bentata L, Beltzer N, Delpierre C, Coste J. Trends and social inequalities in self-reported health and activity limitations in France between 2017 and 2021:

results from four nationwide representative surveys. BMC Public Health. 2024;24:1916.

29. Fontbonne A, Currie A, Tounian P, Picot M-C, Foulatier O, Nedelcu M, et al. Prevalence of Overweight and Obesity in France: The 2020 Obepi-Roche Study by the “Ligue Contre l’Obésité.” J Clin Med. 2023;12:925.

30. González-Gil EM, Matta M, Morales Berstein F, Cairat M, Nicolas G, Blanco J, et al. Associations between degree of food processing and all-cause and cause-specific mortality: a multicentre prospective cohort analysis in 9 European countries. Lancet Reg Health - Eur. 2025;50:101208.

31. Aguilera-Buenosvinos I, Morales Berstein F, González-Gil EM, Dossus L, Gunter MJ, Biessy C, et al. Adherence to the Mediterranean Diet and Obesity-Linked Cancer Risk in EPIC. JAMA Netw Open. 2025;8:e2461031.

32. Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep. 1991;14:540–5.

33. Chung F, Yegneswaran B, Liao P, Chung SA, Vairavanathan S, Islam S, et al. STOP Questionnaire: A Tool to Screen Patients for Obstructive Sleep Apnea. Anesthesiology. 2008;108:812–21.

34. Meerwaldt R, Graaff R, Oomen PHN, Links TP, Jager JJ, Alderson NL, et al. Simple non-invasive assessment of advanced glycation endproduct accumulation. Diabetologia. 2004;47:1324–30.

35. Marc DT, Ailts JW, Campeau DCA, Bull MJ, Olson KL. Neurotransmitters excreted in the urine as biomarkers of nervous system activity: Validity and clinical applicability. Neurosci Biobehav Rev. 2011;35:635–44.

36. Alfonso F, Sandoval J, Cárdenas A, Medina M, Cuevas C, Gonzalo N. Optical coherence tomography: from research to clinical application. Minerva Med. 2012;103:441–64.

37. The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Diabetes Care. 2003;26:s5–20.

38. McEvoy JW, McCarthy CP, Bruno RM, Brouwers S, Canavan MD, Ceconi C, et al. 2024 ESC Guidelines for the management of elevated blood pressure and hypertension. Eur Heart J. 2024;45:3912–4018.

39. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation [Internet]. 2019 [cited 2025 Feb 25];140. Available from: https://www.ahajournals.org/doi/10.1161/CIR.0000000000000678

40. Rubino F, Cummings DE, Eckel RH, Cohen RV, Wilding JPH, Brown WA, et al. Definition and diagnostic criteria of clinical obesity. Lancet Diabetes Endocrinol. 2025;13:221–62.

41. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A

Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. 2019;170:51–8.

42. SCORE2 working group and ESC Cardiovascular risk collaboration, Hageman S, Pennells L, Ojeda F, Kaptoge S, Kuulasmaa K, et al. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. Eur Heart J. 2021;42:2439–54.

43. SCORE2-Diabetes Working Group and the ESC Cardiovascular Risk Collaboration, Pennells L, Kaptoge S, Østergaard HB, Read SH, Carinci F, et al. SCORE2-Diabetes: 10-year cardiovascular risk estimation in type 2 diabetes in Europe. Eur Heart J. 2023;44:2544–56.

44. SCORE2-OP working group and ESC Cardiovascular risk collaboration, De Vries TI, Cooney MT, Selmer RM, Hageman SHJ, Pennells LA, et al. SCORE2-OP risk prediction algorithms: estimating incident cardiovascular event risk in older persons in four geographical risk regions. Eur Heart J. 2021;42:2455–67.

45. Field JK, Vulkan D, Davies MPA, Duffy SW, Gabe R. Liverpool Lung Project lung cancer risk stratification model: calibration and prospective validation. Thorax. 2021;76:161–8.

46. Steffen A, MacInnis RJ, Joshy G, Giles GG, Banks E, Roder D. Development and Validation of a Risk Score Predicting Risk of Colorectal Cancer. Cancer Epidemiol Biomarkers Prev. 2014;23:2543–52.

47. Liu Y, Zhang J, Wang W, Li G. Development and validation of a risk prediction model for incident liver cancer. Front Public Health. 2022;10:955287.

48. Tyrer J, Duffy SW, Cuzick J. A breast cancer prediction model incorporating familial and personal risk factors. Stat Med. 2004;23:1111–30.

49. Wilson PWF. Prediction of Incident Diabetes Mellitus in Middle-aged Adults: The Framingham Offspring Study. Arch Intern Med. 2007;167:1068.

50. Bodenreider O, Cornet R, Vreeman D. Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018;27:129–39.

51. Tannier X, Wajsbürt P, Calliger A, Dura B, Mouchet A, Hilka M, et al. Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse. Methods Inf Med. 2024;63:021–34.

52. Alderman JE, Palmer J, Laws E, McCradden MD, Ordish J, Ghassemi M, et al. Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations. Lancet Digit Health. 2025;7:e64–88.

53. Guillaudeux M, Rousseau O, Petot J, Bennis Z, Dein C-A, Goronflot T, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. NPJ Digit Med. 2023;6:37.

54. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, et al. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J. 2020;41:111–88.

55. Stevens PE, Ahmed SB, Carrero JJ, Foster B, Francis A, Hall RK, et al. KDIGO 2024 Clinical Practice Guideline for the Evaluation and Management of Chronic Kidney Disease. Kidney Int. 2024;105:S117–314.

56. Maloberti A, Mengozzi A, Russo E, Cicero AFG, Angeli F, Agabiti Rosei E, et al. The Results of the URRAH (Uric Acid Right for Heart Health) Project: A Focus on Hyperuricemia in Relation to Cardiovascular and Kidney Disease and its Role in Metabolic Dysregulation. High Blood Press Cardiovasc Prev. 2023;30:411–25.

57. Doroban□u M, Hala□iu V-B, Gheorghe-Fronea O, Bala C-G, Moldovan H, Irinel-Parepa R, et al. The Association between Apolipoprotein B, Cardiovascular Risk Factors and Subclinical Atherosclerosis—Findings from the SEPHAR National Registry on Hypertension in Romania. Int J Mol Sci. 2023;24:2813.

58. Varoquaux G, Luccioni S, Whittaker M. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI. Proc 2025 ACM Conf Fairness Account Transpar [Internet]. Athens Greece: ACM; 2025 [cited 2025 Jun 30]. p. 61–75. Available from: https://dl.acm.org/doi/10.1145/3715275.3732006

59. Facciolà A, Visalli G, D'Andrea G, Varvarà M, Santoro G, Cuffari R, et al. Prevention of cardiovascular diseases and diabetes: importance of a screening program for the early detection of risk conditions in a target population. J Prev Med Hyg. 2022;E934 Pages.

60. Field J, Vulkan D, Davies M, Duffy S, Gabe R. FP09.02 Calibration of the LLP Lung Cancer Risk Stratification Model: Prospective Validation in the UKLS Cohort of 76,000 People. J Thorac Oncol. 2021;16:S212.

61. Van Trier TJ, Snaterse M, Boekholdt SM, Scholte Op Reimer WJM, Hageman SHJ, Visseren FLJ, et al. Validation of Systematic Coronary Risk Evaluation 2 (SCORE2) and SCORE2-Older Persons in the EPIC-Norfolk prospective population cohort. Eur J Prev Cardiol. 2024;31:182–9.

Statements & Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. The research was supported by Zoī SAS's internal projects, without external financial support.

Competing Interests

Financial interests: Claude Dalle, Eric Vibert and Philippe Gabriel Steg received compensation for their collaboration with Zoī Medical Committee. Additionally, independent writer (Djihane Ahmed Lecheheb) and MDs (Marie Bringer, Nina De Luca, Stéphane Ohayon) were compensated for their assistance in manuscript preparation and medical data review. Pierre Bauvin, Alaedine Benani, Maryne Lepoittevin, Marguerite Sentilhes, and Sylvain Bodard are Zoī employees.
Non-financial interests: the authors declare no other relevant non-financial interests to disclose.

Author Contributions

Pierre Bauvin, Alaedine Benani, Marguerite Sentilhes, Xavier Tannier, Claude Dalle, Eric Vibert, Philippe Gabriel Steg, Sylvain Bodard contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Pierre Bauvin, Alaedine Benani, Marie Bringer, Nina De Luca, Stéphane Ohayon. The first draft of the manuscript was written by Pierre Bauvin, Djihane Ahmed Lecheheb and Sylvain Bodard, and all authors reviewed and approved the final manuscript, except Claude Dalle who passed away before the completion of the final draft.

Ethics approval

The study was approved by the Ethics Committee of the Adène Research Foundation (IRB accreditation no.0990-0279)

Consent to participate

The study involved only retrospective, pseudonymized data. All participants were informed generally and individually prior to inclusion, with the possibility to object, in accordance with national regulations.

Consent to publish

No individual-identifiable data or images are included in this manuscript.