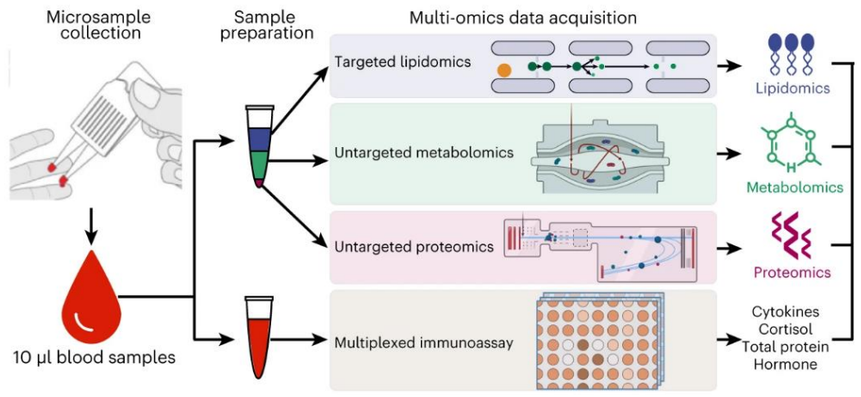
Evaluating the frequency of multi-omics longitudinal data collection in the perspective of chronic disease development.

D.Heylen [1,2], G. Ertaylan [2], J.Aerts [3] and J. Hooyberghs [1]

1 Theory lab, Data Science Institute (DSI), CENSTAT, Hasselt University, Diepenbeek, Belgium.
2 Flemish institute for Technological Research (VITO) Mol, Belgium.
3 Visual Data Analysis Lab, Department of Biosystems, KU Leuven

Incentive of the research

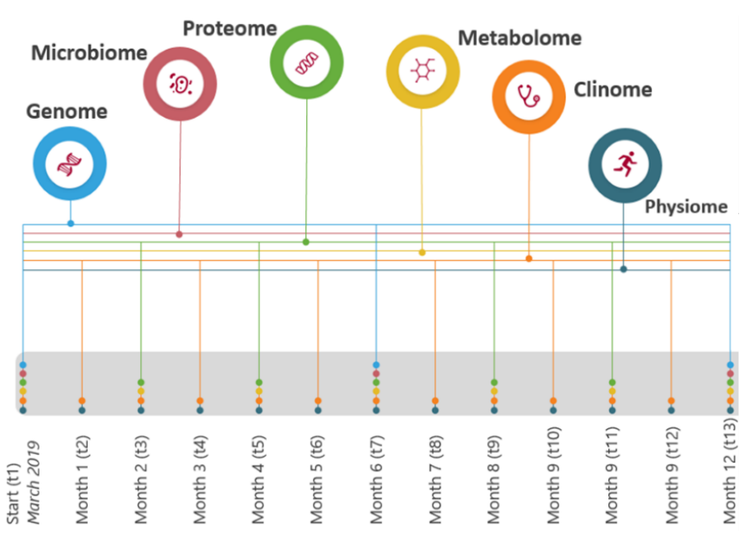
Recently, frequent longitudinal measurements of multiple omics data are emerging as new, technical advancements for the in-depth analyses of molecules in small blood volumes are appearing (1). This makes the sampling process less disruptive for patients (no venepuncture, self-collect possible) and can allow the routine usage of such measurements in clinical practice. **However, we aim to prove the surplus of these highly granular measurements** for assessing disease development since applications that prove this are currently lacking.



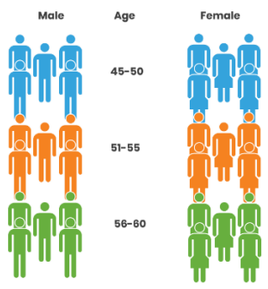
Advancements in data collection From Zimmer et. al 2022 (3)

Data resources

I AM Frontier (IAF) cohort



IAF details: A longitudinal cohort study from VITO for the development of chronic diseases.
- 1000 Proteins
- ~200 clinical parameters
- Whole genome sequencing
- Metabolomics
- Questionnaires



UK Biobank cohort

Data	Size	Data types
UK Biobank	+500 000 individuals	Genome, metabolome, clinome, proteome and anthropometric data available. (All but the protein data is complementary with the IAF data)

Left: I AM Frontier data collection flow diagram: Datatypes were collected at different time points, circle stacks on each time point indicate which datatypes are collected at that moment in time. Right: Gender and age distribution of the cohort. Cohort consists of 30 participants (n=30).

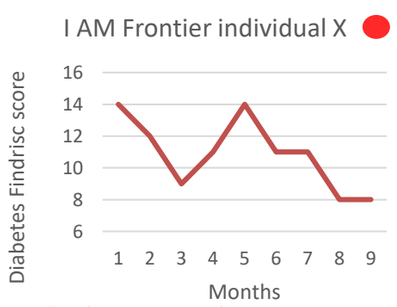
Cohort specific endeavors:

biobank uk — Identification of clear disease profiles.

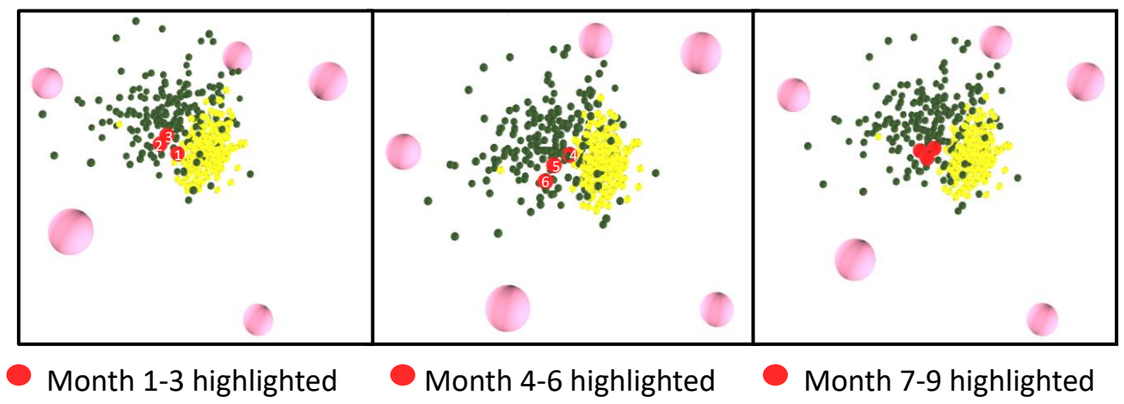
Iam my health — To unravel subtle dynamics as part of **intermediate changes between health and disease.**

For 20 overlapping biochemical features between IAF and UKB it was assessed whether a non-diagnosed IAF individual, with an initial high risk for type 2 diabetes (T2D) that declines throughout the study period, shows **monthly** intermediate changes in a disease spectrum stretching from healthy 'baseline' (2) individuals to T2D diagnosed individuals of the UKB cohort. This was done by applying the Pareto task inference method (ParTI) (3) on the combined dataset Intermediate changes identified with ParTI were contextualized within a customized Clinical knowledge graph database environment (4). The CKG library represents established molecular interaction pathways up to a molecule-specific level for the disease of interest. This enabled us to evaluate the surplus of highly frequent sampling by assessing the time-dependent development of chronic diseases.

Assessment of monthly disease relevant clinical changes



Diabetes Findrisc score: does not incorporate any of the selected biochemical features



Month 1-3 highlighted — Month 4-6 highlighted — Month 7-9 highlighted

Numbers indicate specific months

The Pareto Task Inference (ParTI) is based on an evolutionary theory and its main concept is that if data points of a high-dimensional dataset fall on a simple shape like a line, a triangle, a tetrahedron, it is due to tradeoffs in the biological system, rather than by chance. This method also computes statistical significance for the resulting simplex. If a significant simplex is found, the vertices are denoted archetypes, that specialize in a certain task

- Pareto archetypes
- UK biobank T2D diagnosed individuals
- UK biobank healthy baseline (2) individuals
- I AM Frontier individual X

15/20 of the selected clinical biochemistry features are known to be associated with T2D. Feature specific downstream detailed topological and quantitative pathway differences between phenotype groups can illustrate what pathway region underwent most changes in between months

illustrative subset of the T2D pathway present in CKG

Conclusion

We believe the applied tools can help for the presented combination of multi-omics data to evaluate how time-sensitive each datatype is with regard to intermediate changes between health and disease. In the next step intermediate phenotype profiles will be annotated in disease pathways stored in a CKG graph database. This can help in granularly defining how specific molecules in the human body play a role in disease development.

References.

1) Shen X, Kellogg R, Panyard DJ, Bararpour N, Castillo KE, Lee-McMullen B, et al. Multi-omics microsampling for the profiling of lifestyle-associated changes in health. Nat Biomed Eng. 2023 **2)** Pusparum M, Ertaylan G, Thas O. From population to subject-specific reference intervals. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) [Internet]. **3)** Zimmer A, Korem Y, Rappaport N, Wilmanski T, Baloni P, Jade K, et al. The geometry of clinical labs and wellness states from deeply phenotyped humans. Nat Commun 2021 121 **4)** Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol 2022

D.H. is funded through Hasselt University BOF grant BOF20OWB29 and also receives funding from VITO NV (R-11362).