Blood biomarker data analysis guide

The purpose of this document is to provide guidance on the process from biomarker data to publication. It explains the results files, addresses common questions about the metabolomics panel and individual biomarkers, and points to resources that facilitate the biomarker data analyses. For details, please see the key publications on the Nightingale Health CoreMetabolomics blood platform listed at the end of the document.

Overview of the results files

ProjectID ProjectName-Date-Results.xlsx

- The "Results" sheet reports biomarker concentrations in mmol/L and other units.
- The "Quality Control Tags and Notes" sheet contains quality information related to the whole sample, such as blood specimen type and excluded samples along with the reason for their exclusion. It also contains observations relating to sample procedures, such as low glucose and low protein.
- The "Tags per Biomarker" sheet provides information related to specific biomarkers, such as which biomarker levels are below the limit of quantification in a given sample.
- The "Biomarker Annotations" sheet includes a data dictionary with biomarker annotation names, units, and suggested biological groupings.
- The "Tag Annotations" sheet explains sample quality-related tags.

ProjectID ProjectName-Date-Results.csv

- Contains the same biomarker results as the xlsx file but in a machine-readable format for direct loading into data analysis software.
- The biomarker concentrations and the quality control tag variables related to whole sample are concatenated.
- The CSV column names match the xlsx file row labelled “Shorthand name in CSV file”.
- Biomarker units are given in the xlsx file row labelled “Unit”.

ProjectID ProjectName-Date-Quality-control-report.pdf

- Describes sample quality observations specific to your sample set.
- Lists the success rate of measurement for each biomarker and the rate of observed quality control tags.
- Explains quality control tags such as high ethanol and low glucose.

ProjectID ProjectName-Date-Results-overview.pdf

- Shows histogram distributions of each biomarker in your sample set (black line) in comparison to reference distributions based on general population studies (grey background).
- Note there can be substantial deviations between the distributions in your samples compared to the reference data without this indicating poor sample quality. Reasons for deviations can be differences in sample donor characteristics as well as sample collection protocols and storage aspects.
- Most of the biomarkers are normally distributed. Some biomarkers will have right skewed distributions, including triglycerides, ketone bodies and occasionally creatinine. Lipid concentrations in XXL and XL sized lipoprotein particles are commonly severely skewed – this is not a sign of poor sample quality.
- The overview also provides information on lipid composition in each lipoprotein subclass in your study. These results are generally consistent across cohorts. The plots may help to interpret the information from the lipoprotein lipid ratio measures.

Biomarker annotation and tags in the results files

The biomarker annotations in the results sheet are the following:

Zero (0)	Biomarker has a very low concentration.
(value)	Concentration of biomarker; units are indicated in the “Biomarker annotations” sheet.
NA	Value rejected by automatic quality control. This may be due to overlapping signals preventing quantification. Ratio measures are also marked NA if the denominator is below the Limit of Quantification. This is common for lipid composition measures of the largest VLDL particles, e.g. ‘XL-VLDL-TG %’.
TAG	Tag(s) or note associated with the biomarker concentration; see “Tags per biomarker” worksheet. As an example, glycerol may not be quantifiable in a given sample due to overlapping ethanol signal, which is noted by the tag.
Empty	Biomarker is not applicable for the sample type or was not measured.

Zero values and limit of quantification

Zero values indicate very low concentration; for statistical analyses, it is possible to set the zero-values at, or slightly below, the lowest observed value of the given biomarker.

The lowest nonzero value observed corresponds well to the limit of quantification (LOQ). There is no hard LOQ for each biomarker since the quantification is a holistic assessment relating to the entire spectrum.

The “below LOQ” tag is marked when the concentration is smaller than the range for accurate quantification of the given biomarker. For lipoprotein lipid measures the LOQ is < 0.01 mmol/L. The “below LOQ” is common for measures of the largest VLDL particles, such as ‘XL-VLDL-TG’, and this does not reflect poor sample quality. Biomarker concentration tagged “below LOQ” can still be used for the values for epidemiological analyses, but you may consider sensitivity analyses with those biomarker concentrations excluded.

Overview of the biomarker panel

Biomarker coverage

The Nightingale CoreMetabolomics service is designed to measure biomarkers from multiple key metabolic pathways, whilst ensuring consistent and rapid profiling of large sample collections. As such, the emphasis is on accurate quantification rather than widest possible biomarker coverage. Although most measures are lipoprotein lipids, the biomarker coverage is not based on biological prioritisation of certain metabolic pathways. The physical principles of of NMR mean that the biomarkers measured will be present in high abundance in the bloodstream.

The biomarkers can be grouped into molecular categories such as 'Cholesterol' and 'Amino acids'. This kind of grouping is provided in the "Biomarker annotations" sheet in the xlsx results file. Some biomarkers have a subgroup, such as 'Branched-chain amino acids'. The grouping suggested is intended to facilitate visualization and biological understanding.

Correlation structure and multiple testing considerations

The correlation structure of the biomarkers is illustrated on the last page of this document. Due to the highly correlated nature of lipoprotein lipids, Bonferroni correction for 250 independent variables can be overly conservative. Many papers have used principal components to estimate the number of independent tests in the biomarker data and typically reached 30–50 (Würtz et al, Am J Epidemiol 2017). Alternative statistical methods have come to similar numbers of independent tests (see Barrios et al, Sci Rep 2018 and Zheng et al, Gigascience 2018).

Biomarker distributions and CVs of blind duplicates in the UK Biobank

Distributions of the Nightingale biomarkers observed in UK Biobank are reported in the UK Biobank showcase: https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/nmrm_app1.pdf. These may be helpful to estimate commonly observed levels of the biomarkers in general population settings.

Please, however, note that the UK Biobank plasma samples have suffered from unintended dilution during the initial sample storage process at the UK Biobank. A ~5–10% dilution is estimated from the aliquot used for the Nightingale biomarker measurements. See Julkunen et al, Nat Commun 2023 for details.

Coefficients of variation (CV) for each biomarker are listed in a table that can be downloaded as a PDF here. The CVs are computed from blind duplicates in the UK Biobank and reported both within spectrometer and between spectrometers.

Metabolite pathways analyses

Data-driven pathway analyses with the Nightingale biomarker data are challenging since the lipid biomarkers are not annotated in metabolism databases. Many metabolic pathways are also only sparsely captured by the Nightingale biomarker panel. An example of pathway analyses on the polar metabolites can be found in Smith et al, eLife 2022.

Biomarkers with metabolite identifiers in PubChem, CHEBI, or CAS databases are listed in a table that can be downloaded as a PDF here.

Notes on specific blood biomarkers

Fatty acids

The fatty acid levels reported are esterified fatty acids stemming from within the lipoprotein particles. These are bound to triglycerides, phospholipids and cholesteryl esters inside the lipoprotein particles, and therefore not free fatty acids. The measures can also be labelled total circulating fatty acids.

Glycoprotein acetyls: a biomarker of chronic inflammation

Glycoprotein acetyls (GlycA) is a biomarker for chronic inflammation (Ritchie et al, Cell Syst 2015). The molecular origin is the NMR signal reflecting the abundance of several acute-phase proteins. The NMR signal arises from the acetyl groups of glycoprotein glycans of at least these circulating glycoproteins: alpha-1 antitrypsin, alpha-1-acid glycoprotein, haptoglobin, transferrin, and alpha-1-antichymotrypsin.

GlycA has been shown to predict the risk for numerous cardiometabolic diseases and indicate susceptibility to severe infectious disease (Kettunen et al, Circ Genom Precis Med 2018). Studies show that GlycA has greater long-term stability than high-sensitivity C-reactive protein (Crick et al, Immonology 2024).

Apolipoprotein B

The apolipoprotein B (apoB) measured is apoB-100 isoform. Studies have shown good consistency for apolipoproteins measured by clinical chemistry and Nightingale NMR (Tikkanen et al, JAHA 2021; Figure S1).

Clinical LDL cholesterol

The LDL-C measure matching routine clinical chemistry is labelled ‘clinical LDL C’ in the Nightingale biomarker panel. This is calibrated against direct LDL-C assays. It also matches broadly with levels from the Friedewald equation for LDL-C.

The variable simply labelled ‘LDL-C’ is a measure of so-called ‘size-specific LDL-C’, which adds up to the sum of cholesterol in LDL subclasses. This is always lower than ‘clinical LDL-C’. In other words, ‘Clinical LDL-C’ and ‘size-specific LDL-C’ refer to different methods for defining LDL as described in detail in Holmes and Ala-Korpela, Nat Rev Cardiol 2019.

Lipoprotein subclass and their lipid composition

The biomarker panel features 14 subclasses of lipoproteins: 6 for VLDL, 1 for IDL, 3 for LDL and 4 for HDL. The 14 lipoprotein subclasses have been calibrated via high-performance liquid chromatography (Würtz and Soininen, Atherosclerosis 2020). For each lipoprotein subclass, we measure the particle concentration, and the concentration of individual lipid types (esterified and free cholesterol, triglycerides and phospholipids). The total lipid measurement is the sum of phospholipids, cholesterol esters, free cholesterol and triglycerides. Cholesterol describes the sum of free and esterified cholesterol. The ratios are calculated by dividing the lipoprotein measures by total lipid concentration in each subclass. A key publication on the interpretation of these lipoprotein subclass measures is Ala-Korpela et al, Intl J Epidemiol 2022.

Consistency with clinical chemistry, gas chromatography and mass spectrometry

- Clinical chemistry: several of the biomarkers in the Nightingale panel are commonly measured in cohorts and patient settings by clinical biochemistry analyses. Scatter plots comparing the biomarker concentrations for routine lipids, apolipoproteins, glucose, creatinine and albumin can be found in the Supplementary material of Tikkanen et al, J Am Heart Assoc 2021 (Figure S1) and Julkunen et al, Nat Commun 2023 (Figure S6).
- Fatty acids: scatter plots of the fatty acids measured by Nightingale in comparison to gas chromatography are illustrated in Supplementary Figure 7 of Julkunen et al, Nat Commun 2023. Further comparison of gas chromatography to Nightingale is reported in Rosqvist et al, Clin Nutr 2022;41:2637.
- Polar metabolites by mass spectrometry: correlations of the small number of polar metabolites overlapping between the Nightingale panel and commercial mass spectrometry platforms are reported in the Supplementary Material (p 4) of Julkunen et al, Nat Commun 2023.
- Polar metabolites by NMR: scatter plots of polar metabolites measured by Nightingale compared to an in-house NMR system at Leiden University Medical Center are reported to be R=0.83–0.97 in Deelen et al, Nat Commun 2019, Supplementary Figure 17.

Getting started with the biomarker data analyses

The Nightingale biomarkers are suited for the same statistical toolsets as used for clinical biomarkers. To become familiar with the biomarker data, we recommend starting with the main lipid measures, glucose and creatinine. Next step can be all the biomarkers except the lipoprotein subclass measures. For the detailed lipoprotein measures, you may consider first including one type of lipid metrics (e.g. total lipid or total particle concentration in each of the 14 subclasses).

Although all the Nightingale biomarkers are quantified simultaneously, there is no need to analyse all of them in a given research application. For an example of biomarker selection without including all lipoprotein subclass measures, see e.g. Tikkanen et al, J Am Heart Assoc 2021.

The following analyses may serve as helpful positive controls:

- Compare correlations of Nightingale biomarkers with clinical chemistry for lipids, glucose and creatinine if those measures are available in your cohort.
- Plot associations of the biomarkers with BMI since these are consistent across cohorts. For an overview of cross-sectional associations of BMI as continuous variable per SD scaled biomarker concentration, please see e.g. Würtz et al, PLoS Med 2014.
- Plot biomarker associations with incident diabetes since these are also consistent across studies. For examples of odds ratios per SD, see e.g. Ahola-Olli et al, Diabetologia 2019.
- Compare your results to those reported in the UK Biobank using the online atlas of biomarker associations across the spectrum of diseases, see Julkunen et al, Nat Commun 2023.
- For genetic analyses, positive control SNPs can be found in Kettunen et al, Nat Genet 2012, e.g. rs174547 in the FADS gene. Examples of genetic variants in HMGCR and PCSK9 on detailed lipid measures are described in e.g. Sliz et al, Circulation 2018.

We recommend plotting individual biomarker distributions stratified by characteristics of your cohort, e.g. per time-point, intervention group, or collection site, as relevant to your study. This may also reveal if there are apparent issues with subsets of the samples.

Examples of statistical approaches to derive multi-biomarker scores for risk prediction can be found in Nightingale Health Biobank Collaborative Group, Metabolomic and genomic prediction of common diseases in 700,217 participants in three national biobanks, Nat Commun 2024.

Free data analysis tools

For visualizing association testing with the biomarkers, we recommend an R-package called ggforestplot. This package is developed by Nightingale Health and has been used for many publications. It is freely available at nightingalehealth.github.io/ggforestplot/articles/ggforestplot.html.

Researchers at University of Cambridge have developed an R-package called ukbnmr for technical artefact correction of Nightingale biomarker data in the UK Biobank. In general, it is not necessary to consider such corrections to the biomarker data from other cohorts. The biomarker-disease association magnitudes in the UK Biobank are identical whether or not this type of correction is done; however a slight increase in power may be gained for genetic analyses. The comprehensive quality control assessment is described in Ritchie et al, Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. Sci Data 2023;10:64.

Atlas of biomarker-disease associations in the UK Biobank

A free webtool provides a comprehensive atlas of associations for biomarker with prevalence, incidence and mortality of >700 disease endpoints in the UK Biobank: https://research.nightingalehealth.com/atlas.

The atlas provides an easy way to examine the disease specificity of each biomarker. This tool can also be helpful for replicating results from other cohorts. The biomarker-disease atlas is described in Julkunen et al, Nat Commun 2023 and has been updated to include results for all ~500,000 UK Biobank participants since the original publication.

Genome-wide association summary statistics

GWAS summary statistics of the Nightingale biomarkers are available as part of the following papers:

- Tambets et al, Genome-wide association study for circulating metabolites in 619,372 individuals. MedRxiv 2024.
- Zoodsma et al, A genetic map of human metabolism across the allele frequency spectrum. MedRxiv 2025.
- Karjalainen et al, Genome-wide characterization of circulating metabolic biomarkers, Nature 2024.
- GWAS summary statistics from UK Biobank data are also available through MR-Base database, developed by the University of Bristol.

Publications describing the Nightingale Health CoreMetabolomics blood platform

- Julkunen et al, Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. Nat Commun 2023;14:604.
  The Methods section describes the experimental setup and the biomarker coverage. The Supplement contains details about the quality control, sample tags, and technical repeatability. The Methods section also includes scatter plots comparing the Nightingale biomarkers to clinical biochemistry, gas chromatography and mass spectrometry.

- Würtz et al, Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. Am J Epidemiol 2017;186:1084-1096.
  This paper illustrates statistical methods that have been useful in applications of the Nightingale biomarker data in population cohorts.

- Soininen et al, Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. Circ Cardiovasc Genet 2015;8:192-206.
  This paper illustrates part of the experimental setup and and outlines a roadmap from research tool towards clinical use.

- Documentation on the Nightingale biomarker data in UK Biobank Showcase, available at https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220.
  This online documentation describes the biomarker platform and recommended approaches for data preprocessing and epidemiological analyses. This resource also includes distributions of all the biomarkers, coefficients of variation from blinded duplicates, and assessment of biological repeatability for samples measured in same individuals 5 years apart.

- A full list of publications on Nightingale biomarker data in research studies is available at https://research.nightingalehealth.com/success-stories/publications.

Paragraph on Nightingale Health CoreMetabolomics for your publications

Feel free to use the following paragraph:

"Metabolic biomarkers were quantified from serum/plasma samples of XXXX individuals using Nightingale Health NMR CoreMetabolomics (Nightingale Health Plc, Finland). This assay provides simultaneous quantification of routine lipids, apolipoproteins, lipoprotein subclass profiling with lipid concentrations within 14 subclasses, fatty acid composition, and various low-molecular metabolites including amino acids, ketone bodies and glycolysis-related metabolites in molar concentration units. Details of the experimentation and epidemiological applications of the Nightingale Health metabolomics platform for research have been described previously (Julkunen et al, Nat Commun 2023;14:604)."

If you have a focus on lipoprotein subclass measures in your paper, you may wish to also include:

"The 14 lipoprotein subclass sizes are defined as follows: extremely large VLDL with particle diameters from 75 nm upwards and a possible contribution of chylomicrons, five VLDL subclasses (average particle diameters of 64.0 nm, 53.6 nm, 44.5 nm, 36.8 nm, and 31.3 nm), IDL (28.6 nm), three LDL subclasses (25.5 nm, 23.0 nm, and 18.7 nm), and four HDL subclasses (14.3 nm, 12.1 nm, 10.9 nm, and 8.7 nm). The following lipid components within the lipoprotein subclasses are quantified: phospholipids (PL), triglycerides (TG), cholesterol (C), free cholesterol (FC), and cholesteryl esters (CE). The mean size for VLDL, LDL and HDL particles is calculated by weighting the corresponding subclass diameters with their particle concentrations (Soininen et al, Circ Cardiovasc Genet 2015;8:192)."

Correlation structure of the Nightingale blood biomarkers