Acute Effects of Endurance Exercise with Moderate and High Intensity on Breast Milk Composition Among Women with Overweight/Obesity

Acronym: YT

NCT number: NCT05745922

Document date: 07. September 2025

Statistical Analysis Plan

Study design: Randomized crossover study

Aim: Determine the effect of moderate-intensity continuous training and high-intensity interval training on human milk concentrations of metabolites, lipids, cytokines, hormones and miRNAs in lactating women with overweight/obesity.

Overview:

- Up to N=30 women with BMI > 25 kg/m^2
- Each participant goes through three conditions in random order (six possible sequences) and has milk sampled at four time points per condition (one pre-condition and three post-condition).
- Washout period: 1 week between conditions
- The samples are analysed for a range of outcome variables within different molecular data ('omics') categories.

Conditions:
C1: No exercise, resting in the lab
C2: Moderate-intensity continuous training
C3: High-intensity interval training

Time points:
T1: 7 am (pre-condition)
T2: 11 am (immediately after condition)
T3: 12 am (1-hour post-condition)
T4: 15 am (4 hours post-condition)

Pre-planned outcome groups measured in human milk:
Targeted NMR metabolomics (GCF)
Metabolomics + lipidomics profile panels (Metabolon)
Cytokines (Olink)
Hormones (insulin, ghrelin, leptin, adiponectin, IGF-1)
miRNA (GCF)

Statistical analysis plan

For each outcome category, the primary objective is to evaluate the effect of conditions C2 and C3 on the outcome variables at time points T2, T3, and T4, using condition C1 and time point T1 as references. Outcomes measured in all participants will be analysed according to the intention-to-treat principle, whereas outcomes measured only in participants who followed the protocol perfectly will be analysed according to the per-protocol principle. The decision whether to include all samples or only those from participants who complied with the protocol will depend on the costs of including the extra (non-compliant) samples for each outcome group.

We will perform the following six analysis steps for all outcome groups:

1) Data preparation (batch adjustment, log transformation, imputation)

- If needed, the outcome data will be batch adjusted, imputed (e.g. using the R package zCompositions), and log transformed appropriately before analysis. Adjustments and transformations will depend on whether the data is already normally distributed, and if it is required for the method in question.

- The metabolomics data from Metabolon already comes in a log-transformed, batch-normalized version where missing values (below the detection limit) are imputed to the observed minimum for that metabolite after batch normalization. These will be used in all analyses.

2) Principal component analysis (PCA)

- We will use PCA to detect outliers and identify global structures/important features by capturing the maximum variance in the dataset, and test if the data separates according to the conditions in an unsupervised setting.

- We will identify the number of PCs that explain 95% of the variation in the data at baseline, for use in multiple testing adjustment.

3) Linear mixed model (LMM) analyses:

- We will use LMM analyses to evaluate the effects of each condition and time point on all outcome variables:

- We will analyse the effects of time and time-condition interaction as fixed effects, and participant ID as a random intercept (Eq 1). The analysis will be reference coded, with C1 (no exercise) and T1 (morning) as reference points. We will not adjust for or subtract baseline values in the model.

  Eq. 1: value ~ time + time:condition + (1|id)

- We will run the LMM analyses using e.g. the lme4 R package, and estimate p-values using the Satterthwaite method, e.g., as implemented in the lmerTests R package.

- Multiple testing correction: We will use a multiple testing corrected p-value threshold determined by dividing the alpha level (0.05) by the number of PCs explaining 95% of the variation in the outcome data at baseline.

- For variables passing the multiple testing corrected p-value cut-off, we will test for potential carry-over effects by adjusting for sequence (Eq. 2) in sensitivity analysis:

  Eq. 2: response matrix ~ time + time:condition + sequence + sequence:condition + time:sequencec:condition + (1|id)

4) Repeated Measures (RM)-ASCA+

- We will use RM-ASCA+ to evaluate the overall effect of each condition and time point on each outcome data category. We will model all variables within the same data category together in a multiple-response variable analysis.

- We will analyse both the combined and separate effects of time and time-condition interaction as fixed effects, and participant ID as a random intercept (Eq 3). The analysis will be reference coded, with C1 (No exercise) and T1 (morning) as reference points. We will not adjust for or subtract baseline values in the model.

  Eq. 3: response matrix ~ time + time:condition + (1|id)

- Each variable will be scaled by the standard deviation of the baseline (morning) samples across conditions.

- We will use cluster bootstrapping with 1000 iterations to validate the findings and generate confidence intervals.

The results from the LMM and RM-ASCA+ analyses will guide follow-up analyses using additional machine learning models (PLS-DA, t-SNE).

5) Partial Least Squares Discriminant Analysis (PLS-DA)

- Supervised machine learning approach to analyse differences between the conditions. We will provide the model with a response matrix specifying the conditions, and the model will maximize separation between the conditions. The

method works well with small datasets and multiclass scenarios (here: multiple conditions).

- We will use PLS-DA for feature selection/to explore global structures in the data or find variables that are especially important for one time/condition compared to another. The LMM and RM-ASCA+ analyses will guide the selection of conditions/time points for comparison.

- We will perform cross-validation and permutation testing with 1000 repetitions.

6) T-Distributed Stochastic Neighbour Embedding (t-SNE)

- We will use t-SNE in an exploratory analysis to look for clusters of variables by preserving local relationships (pairwise similarities between data points, clusters).