Graduate Metabolomics Course GSBC 724

Design of Experiment in Metabolomics

Hemant K. Tiwari, Ph.D.
Professor
Department of Biostatistics
School of Public Health

Design of Experiments/ Experimental Design

- A controlled experiment to either test a hypothesis or generate hypotheses.

- In the design of experiments, the experimenter is usually interested in the effect of some process or intervention on some subjects.

History of Experimental Design

- In 1747, James Lind (a Scottish Physician) developed the theory that citrus fruits cured scurvy (Symptoms include bleeding sores, tooth loss, anemia, and a reduced rate of healing for injuries) while serving as surgeon on HMS Salisbury Ship of the Royal Navy. This was the first-ever clinical trial conducted.
- It can be fatal if left untreated.

Lind's Experiment

- Lind selected 12 men from the ship suffering from scurvy. He divided them into six pairs, giving each pair different supplements to their basic diet for two weeks. The treatments were all remedies that had been proposed:
  - A quart of cider every day
  - Twenty-five drops of *elixir vitriol* (Sulfuric acid) three times a day upon an empty stomach
  - One half-pint of seawater every day
  - A mixture of garlic, mustard, and horseradish in a lump the size of a nutmeg
  - Two spoonful of vinegar three times a day
  - Two oranges and one lemon every day

Result: The men given citrus fruits recovered dramatically within a week.

It is known that Scurvy is a disease resulting from a deficiency of Vitamin C.

Source: Wikipedia

Lady Tasting Tea Experiment

Design of experiments was born as a result of an unlikely, but true anecdote: A lady claimed before R.A. Fisher that she was able to ascertain whether milk was poured before or after tea in her cup of tea. Fisher devised a study to verify her claim, and in turn, this gave birth to Experimental Design.

Milk poured first (4 cups)
Tea poured first (4 cups)

There are 70 different possible orderings: $\binom{8}{4} = \frac{8!}{4!4!} = 70$

Her answers:

		True order		Total
		Tea First	Milk First	
Lady's Guesses	Tea First	a=3	b=1	a+b=4
	Milk First	c=1	d=3	c+d=4
Total		a+c=4	b+d=4	N=8

Solution Using Fisher’s Exact Test

- $P_{Fisher} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$

$$P_{Fisher} = 0.24286$$

- P-value is non-significant at the significance level of 0.05

- Thus, the lady is not able to discriminate whether tea or milk was poured first

Experimental Design: Define the Problem

- What is the topic?
- What is the good question for an experiment?
- Is your question testable with the materials in your hand?
- Need to know hypothesis to guide your experiment?
- <u>Design your experiment </u>that will test your hypothesis.

Main Aims of the Experimental Design

- Maximize the Systematic/ experimental variance of the variable(s) of the research hypothesis (i.e., maximize the difference in the dependent variable (outcome) caused by maximizing the differences in the independent variable (treatment).
- Control the variance of extraneous (unwanted) variables that may affect the outcome other than treatment that could be causing differences in the outcome.
- Minimize the random variance/error due to unreliable measurement instruments that have high error of measurement.

Control for Extraneous Variable

- Eliminate the variable (for example if sex specific effect exists, then include only males or females, i.e., stratify the sample).
- Randomization
- Build it into design
- Match subjects

General Statistical Principles of Experimental Design

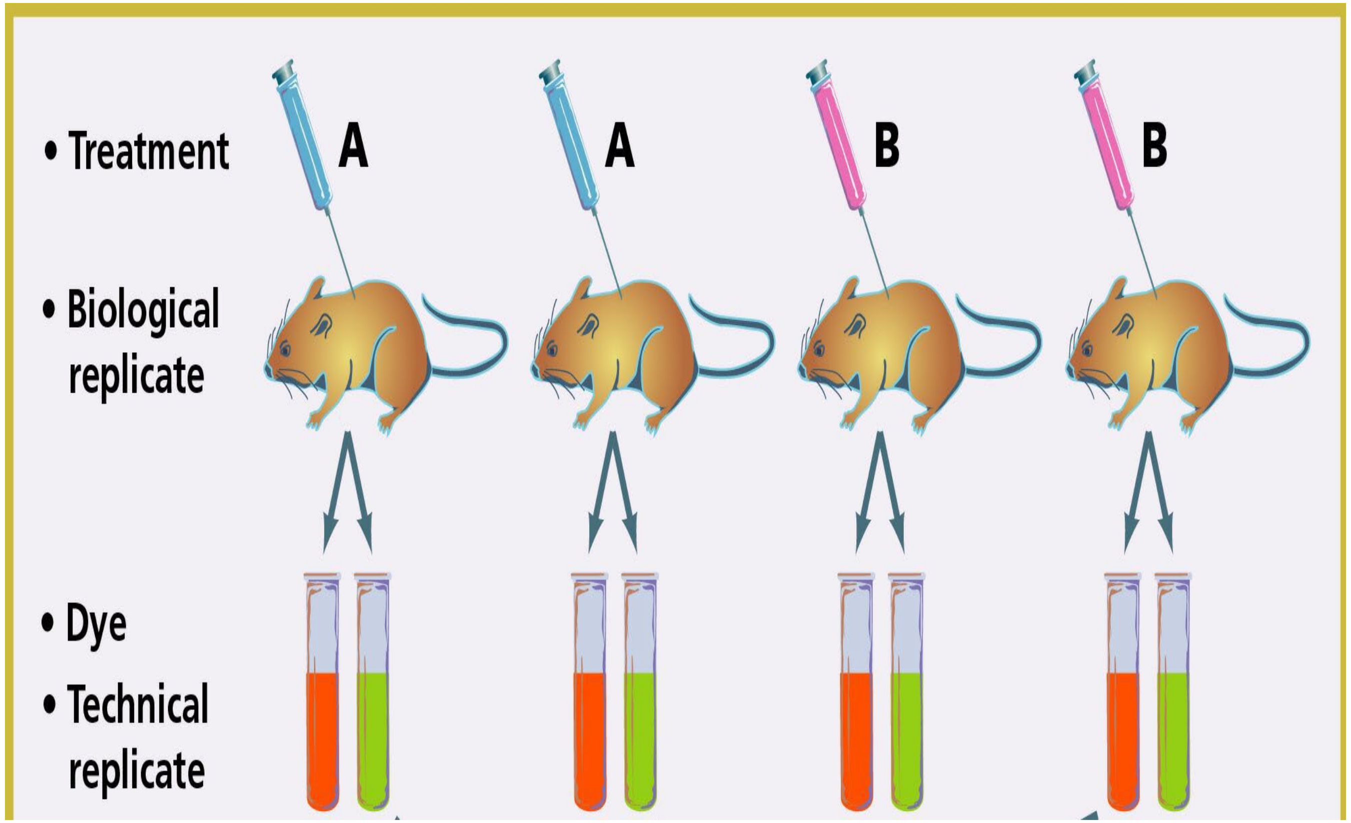
- Replication
- Randomization
- Blocking (Stratification)
- Use of factorial experiments instead of the one-factor-at-a-time methods

Replication

- **Replication** – repetition of a basic experiment without changing any factor settings, allows the experimenter to estimate the experimental error in the system used to determine whether observed differences in the data are “real” or “just noise”, allows the experimenter to obtain more statistical power (ability to identify small effects).

- Replications should not be confused with repeated measurements which refer to taking several measurements of a single occurrence of a phenomenon (single experiment).

Replications should not be confused with repeated measurements.



Replicates

Repeated measure

Replicates

- Number of biological replicates matter in the power of the analysis
- **Experiment:** one mouse per group (treatment group vs. untreated group)- you can only measure the difference in metabolites, but no variance
- 5 or 10 mouse per group- you can measure both the difference in metabolites and the variance (very important for statistical testing)

More terms saying the same things

- What to replicate?
  - Biological replicates (replicates at the experimental unit level, e.g., mouse, plant, pot of plants…)
    - Experimental unit is the unit that the experiment treatment or condition is directly applied to, e.g., a plant if hormone is sprayed to individual plants; a pot of seedlings if different fertilizers are applied to different pots.
  - Technical replicates
    - Any replicates below the experimental unit, e.g., different leaves from the same plant sprayed with one hormone level; different seedlings from the same pot; Different aliquots of the same RNA extraction; multiple arrays hybridized to the same RNA; multiple spots on the same array.

Randomization

- **Randomization** – a statistical tool used to minimize potential uncontrollable factors “lurking variables” (which might vary over the length of the experiment) in the experiment by randomly assigning treatment to the experimental units.

- Results in “averaging out” the effects of the extraneous factors that may be present to minimize the risk of these factors affecting the experimental results.

- Randomization is essential for making causal inferences, e.g., using Randomized Controlled Trials (RCTs).

Randomization

- Experimental units (people, mouse, plant etc.) should be assigned to treatment groups **at random**.
- Can be done by using
  - Computer
  - Coins

Example

- Number the objects to be randomized and then randomly draw the numbers.

Example: Assign treatment/Special Diet and no treatment (control) to 6 mice (3 each)

1

2

3

4

5

6

Special Diet/treatment : 1, 3, 4
Control : 2, 5, 6

Blocking/ Stratification

- **Blocking** – the technique used to increase the precision of an experiment by breaking the experiment into homogeneous segments (blocks) to control any potential block to block variability (e.g., measurement of metabolites in different days or shifts, by different technicians, by different machines).
- Any effects on the experimental results of the blocking factor will be identified and minimized.

Blocking

- Some of these identified uninteresting but varying factors can be controlled through blocking.

  - COMPLETELY RANDOMIZED DESIGN
  - COMPLETE BLOCK DESIGN
  - INCOMPLETELY BLOCK DESIGN

Completely Randomized Design

There is no blocking

➡ Example

- Compare hormone treatment (Trt) and control (no treatment) using 6 Arabidopsis (rockcress) plants (or mice or humans).

1

2

3

4

5

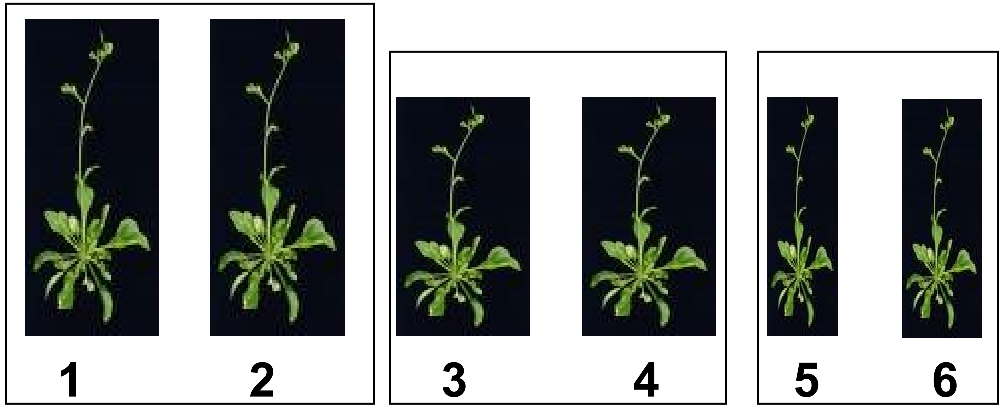
6

	Example 1	Example 2
Hormone Trt:	**(1,3,4);**	**(1,2,6)**
Control :	**(2,5,6);**	**(3,4,5)**

Complete Block Design

- There is blocking, and the block size equals the number of treatments.

Example: Compare hormone treatment and control using 6 Arabidopsis (rockcress) plants. For some reason, plants 1 and 2 are taller, and plants 5 and 6 are thinner.



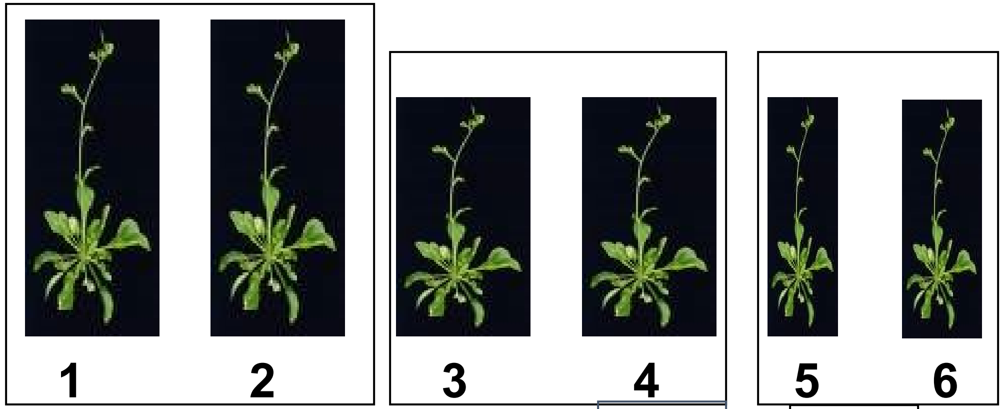
Hormone treatment: (1,4,5) ; (1,3,6)
Control : (2,3,6) ; (2,4,5)

⇨ Randomization within blocks

Incomplete Block Design

➡ There is blocking, and the block size is smaller than the number of treatments. You can assign all treatments in each block.

Example: Compare three hormone treatments (hormone level 1, hormone level 2, and control) using 6 Arabidopsis plants. For some reason plant 1 and 2 are taller, plant 5 and 6 are thinner.



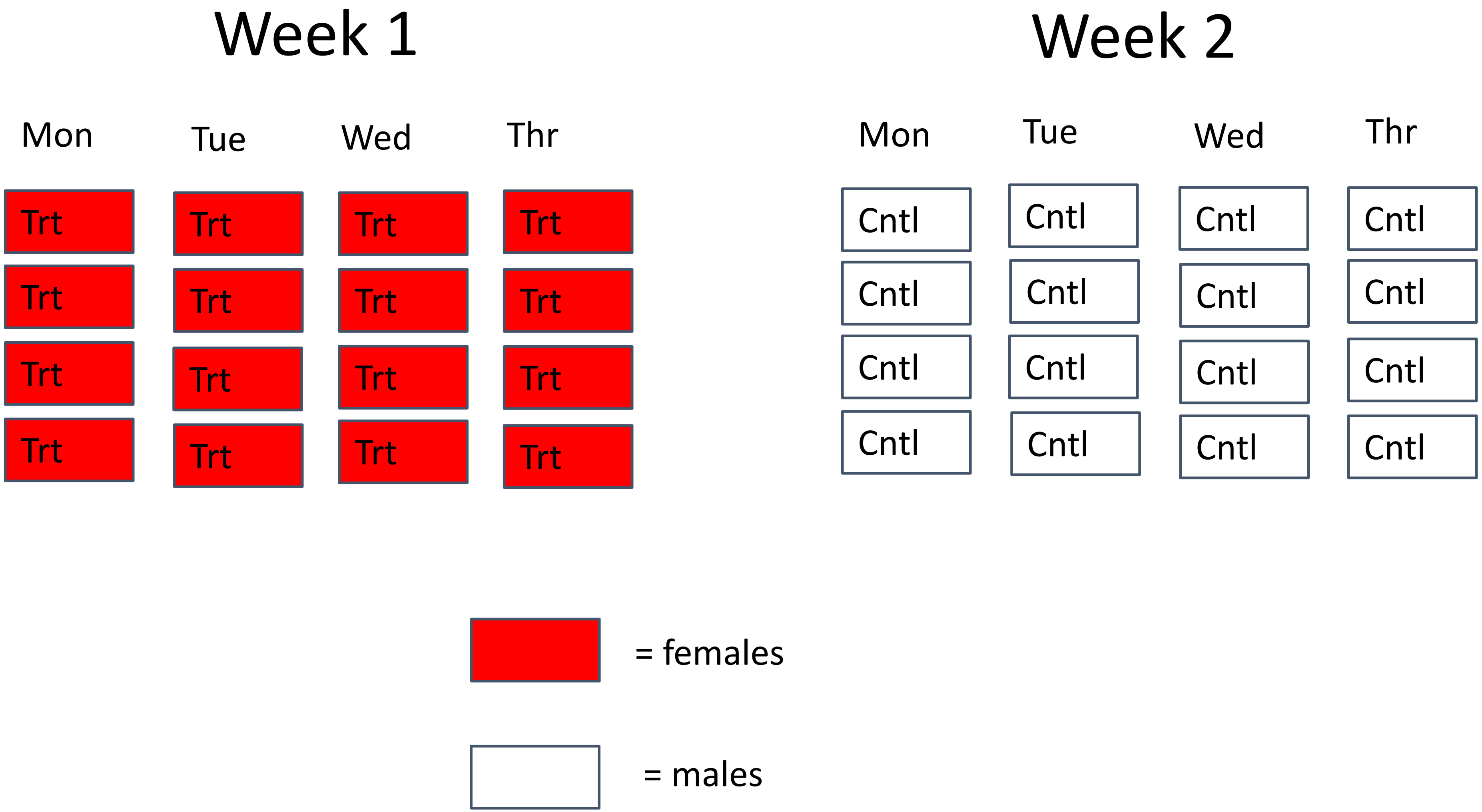
Hormone level1:	**(1,4)**	**(2,4)**
Hormone level2:	**(2,5)**	**(1,6)**
Control :	**(3,6)**	**(3.5)**

⇨ Randomization within blocks

Example 2

- 32 mice (16 males and 16 females)
- Half to be treated and another half left untreated
- A technician can work only 4 mice per day and only on Monday through Thursday

Very Bad design!



Week 1				Week 2			
Mon	Tue	Wed	Thr	Mon	Tue	Wed	Thr
Trt	Trt	Trt	Trt	Cntl	Cntl	Cntl	Cntl
Trt	Trt	Trt	Trt	Cntl	Cntl	Cntl	Cntl
Trt	Trt	Trt	Trt	Cntl	Cntl	Cntl	Cntl
Trt	Trt	Trt	Trt	Cntl	Cntl	Cntl	Cntl

🟥 = females

□ = males

Randomization

Week 1

Mon	Tue	Wed	Thr
🟥 Trt	🟥 Trt	⬜ Trt	🟥 Trt
⬜ Cntl	⬜ Trt	⬜ Trt	🟥 Cntl
🟥 Trt	🟥 Cntl	⬜ Trt	⬜ Cntl
⬜ Cntl	⬜ Cntl	🟥 Cntl	🟥 Trt

Week 2

Mon	Tue	Wed	Thr
⬜ Cntl	🟥 Trt	⬜ Trt	⬜ Cntl
🟥 Cntl	🟥 Cntl	⬜ Cntl	🟥 Trt
⬜ Cntl	⬜ Trt	⬜ Cntl	🟥 Trt
🟥 Trt	⬜ Cntl	🟥 Trt	🟥 Cntl

🟥 = females

⬜ = males

Blocking

Week 1

Mon	Tue	Wed	Thr
□ Trt	🟥 Trt	🟥 Trt	□ Cntl
🟥 Cntl	□ Cntl	🟥 Cntl	□ Trt
□ Cntl	🟥 Cntl	□ Trt	🟥 Cntl
🟥 Trt	□ Trt	□ Cntl	🟥 Trt

Week 2

Mon	Tue	Wed	Thr
□ Trt	🟥 Cntl	□ Trt	🟥 Cntl
□ Cntl	□ Trt	🟥 Trt	□ Cntl
🟥 Trt	□ Cntl	□ Cntl	🟥 Trt
🟥 Cntl	🟥 Trt	🟥 Cntl	□ Trt

🟥 = females

□ = males

Confounding

- Confounding - A concept that basically means that multiple effects are tied together into one parent effect and cannot be separated. For example,
  - For example, a study looking at the association between obesity and heart disease might be confounded by age, diet, smoking status, and other risk factors that might be unevenly distributed between the groups being compared.
  - As experiments get large, higher-order interactions are confounded with lower-order interactions or main effects.

Sample Size and Power

- Purpose
  - Planning a study: number of individuals to recruit or number of mice to test a research hypothesis
  - Understand sample size implications of alternative study designs
  - Sample was already collected and wants study using new technology
    - Genome-Wide Association Study (GWAS) was done, but wants to do metabolomics or RNA-Seq on the same data set

Sample Size and Power Calculation

- Often the number of samples to be used for the experiments dictated by the reality of resources available, not science.
  - How much money is available for the experiment
  - What is the cost per sample
  - Thus, sample size = $ available / cost per sample

Hypothesis Testing

- Power calculations are based on the principles of hypothesis testing
- A hypothesis is a statement about the population parameter
- The two complementary hypotheses in a hypothesis testing problem are called the null hypothesis (H_0) and alternate hypothesis (H_1)
- A statistically significant result does not imply that a research hypothesis is correct (as this implies 100% certainty). Because a *P*-value is based on probabilities, there is always a chance of making an incorrect conclusion regarding accepting or rejecting the null hypothesis (H_0).

Two types of errors in hypothesis testing

H_0: θ=0 versus H_1: θ$\neq$ 0, where θ is a parameter to be tested

		Decision	
		Accept H_0 (Null)	Reject H_0 (Alternate)
Truth	H_0	Correct Decision	Type I error (α)
	H_1	Type II error (β)	Correct decision

Reject a true null hypothesis

Accept a false null hypothesis

- **Type I Error:** Probability of finding a statistically significant effect when the truth is that there is no effect. A type I error is also known as a false positive
- **Type II Error:** A type II error is also known as a false negative, and the researcher concludes there is not a significant effect when there really is.

Significance Level and Power

- The probability of making a type I error is represented by your alpha level (α), the *P*-value below which you reject the null hypothesis. The alpha level is also called the significance level. If alpha=0.05, a *P*-value of 0.05 indicates that you are willing to accept a 5% chance that you are wrong when you reject the null hypothesis.
- The probability of making a type II error is called Beta (β), and this is related to the power of the statistical test (**Power** = 1- β).
- You can decrease your risk of committing a type II error by ensuring your test has enough power.
- Goal is to minimize both types of errors.

Power depends on …

- Design
- The method of analyzing the data
- The effect size
- Variance of the effect of interest
- The chosen significance level (α)
- The sample size

We usually use significance levels of 5% and 80% power to estimate the sample size for a hypothesis

Factors Affecting the sample size

Effect size ↑	Required sample size ↓
Variation of data ↑	Required sample size ↑
Type I error rate ↓	Required sample size ↑
Power ↑	Required sample size ↑

- Type I error rate (α) is kept fixed and becomes smaller as number of tests increase
- Effect size and variation of the data (σ^2) is either obtained through pilot study or vary to calculate different sample sizes.

To calculate Sample Size

- Need to know level of significance (α)
- Statistical power (1- β)
- Effect size (expected difference)
- Standard deviation
- What statistical test we are going to use

Sample Size Formula for the difference in means in Case-Control Design

- A sample size formula to test difference of means between two groups (two-tailed test)
- $n = \frac{r+1}{r} * \frac{\sigma^2(z_{1-\beta}+z_{1-\alpha/2})^2}{\Delta^2}$, *where*
- n= Number of samples which we need to find out
- r = Control to case ratio
- $z_{1-\beta}$ = standard normal deviate corresponds to 1-β
- $z_{1-\alpha/2}$ = standard normal deviate corresponds to two-tailed significance level
- Δ = difference in means of the outcome
- σ^2 =Variance of the difference of the means

Simple Example

- How many people would you need in each group (assuming both groups of equal size) to achieve 80% power if SD =σ=10, difference in mean is 5 with fixed α=0.05. So $z_{1-\alpha/2}$ =1.96, $z_{1-\beta}$ =0.84, for 1-β=0.80, so β=0.20 and r=1, then
  - $n = \frac{r+1}{r} * \frac{\sigma^2(z_{1-\beta}+z_{1-\alpha/2})^2}{\Delta^2}$ = 2 (100) $(.84+1.96)^2$/ (5^2) = 62.72 ~63

  63 per group implies 126 altogether.

Real Example

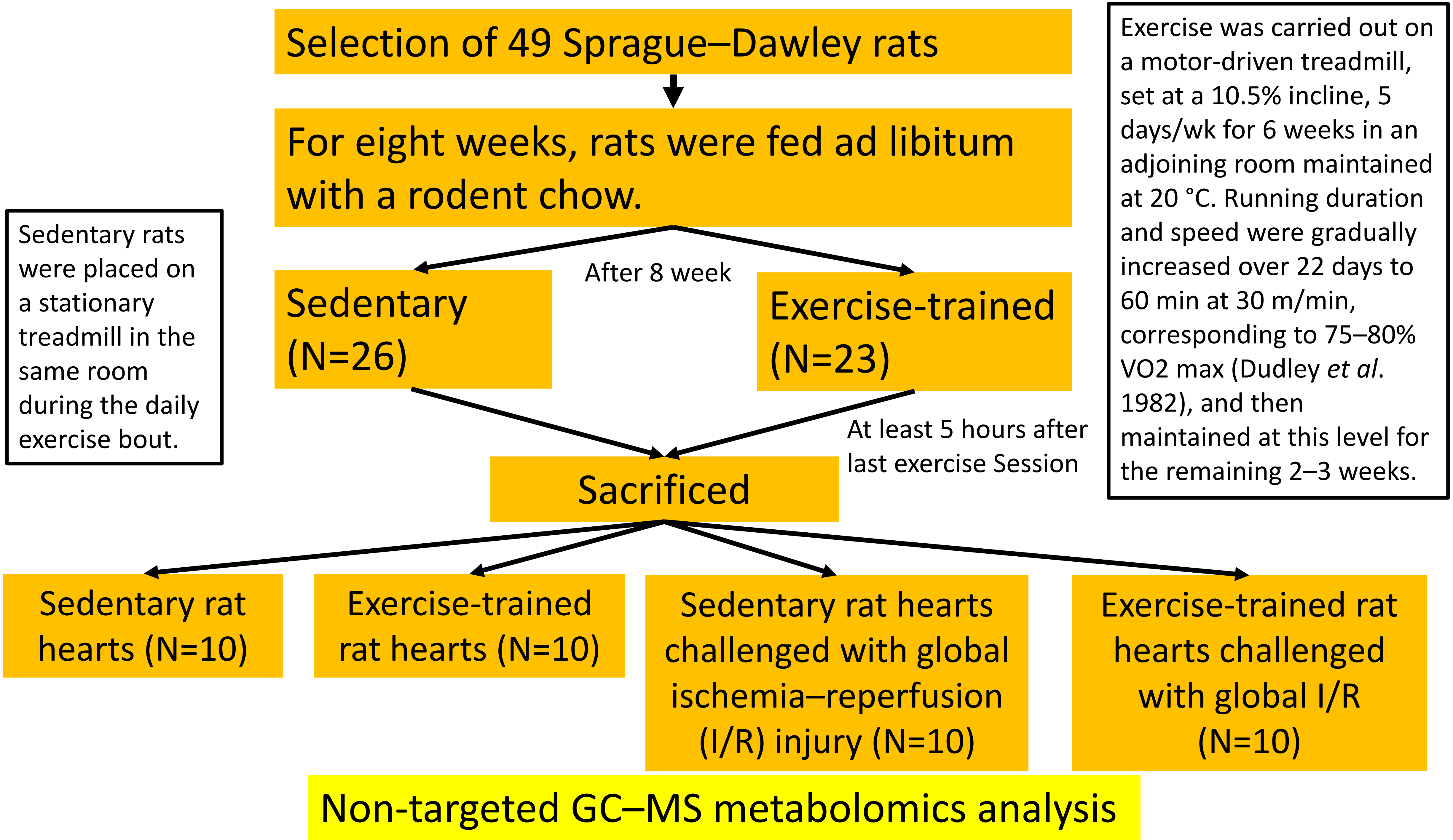
- Parry *et al*. **Untargeted metabolomics analysis of ischemia-reperfusion-injured hearts ex vivo from sedentary and exercise-trained rats.** *Metabolomics*. 2018 Jan;14(1). pii: 8. doi: 10.1007/s11306-017-1303-y. Epub 2017 Dec 4.

Ischemia/**Reperfusion** (I/**R**) **injury** is defined as the cellular damage that results from a period of ischemia that is followed by the re-establishment of the blood supply to the infarcted tissue.

Ischemia is a condition in which blood flow (and thus oxygen) is restricted or reduced in a part of the body. Cardiac ischemia is decreased blood flow and oxygen to the heart muscle. (AHA)

Parry *et al*. (2018)

- **Scientific Premise—**The effects of exercise on the heart and its resistance to disease are well-documented. Recent studies have identified that exercise-induced resistance to arrhythmia is due to the preservation of mitochondrial membrane potential.
- **Objectives—**To identify novel metabolic changes that occur parallel to these mitochondrial alterations, they performed non-targeted metabolomics analysis on hearts from sedentary and exercise-trained rats challenged with isolated heart ischemia–reperfusion injury (I/R).



Selection of 49 Sprague–Dawley rats

For eight weeks, rats were fed ad libitum with a rodent chow.

After 8 week

Sedentary (N=26)

Exercise-trained (N=23)

Sedentary rats were placed on a stationary treadmill in the same room during the daily exercise bout.

Exercise was carried out on a motor-driven treadmill, set at a 10.5% incline, 5 days/wk for 6 weeks in an adjoining room maintained at 20 °C. Running duration and speed were gradually increased over 22 days to 60 min at 30 m/min, corresponding to 75–80% VO2 max (Dudley *et al*. 1982), and then maintained at this level for the remaining 2–3 weeks.

At least 5 hours after last exercise Session

Sacrificed

Sedentary rat hearts (N=10)

Exercise-trained rat hearts (N=10)

Sedentary rat hearts challenged with global ischemia–reperfusion (I/R) injury (N=10)

Exercise-trained rat hearts challenged with global I/R (N=10)

Non-targeted GC–MS metabolomics analysis

Non-targeted GC–MS metabolomics analysis

- **Sample Preparation:** Left ventricular tissue was flash frozen in liquid nitrogen, weighed (25–50 mg wet wt.), then placed in a buffer (50% acetonitrile, 50% water, 0.3% formic acid) at a standard concentration of 25 mg/475 µl buffer and fully homogenized on ice for 20–25 s. Tissues were then placed on dry ice and stored at −80 °C.
- Samples were analyzed by GC/MS.
- Four groups with ten biological replicate samples were analyzed (40 total). If more than three individuals did not have a metabolite detected in a group (of 10 total), they were excluded from further analysis for that metabolite. **In groups missing values, the lowest value of that group was used to impute those values**.

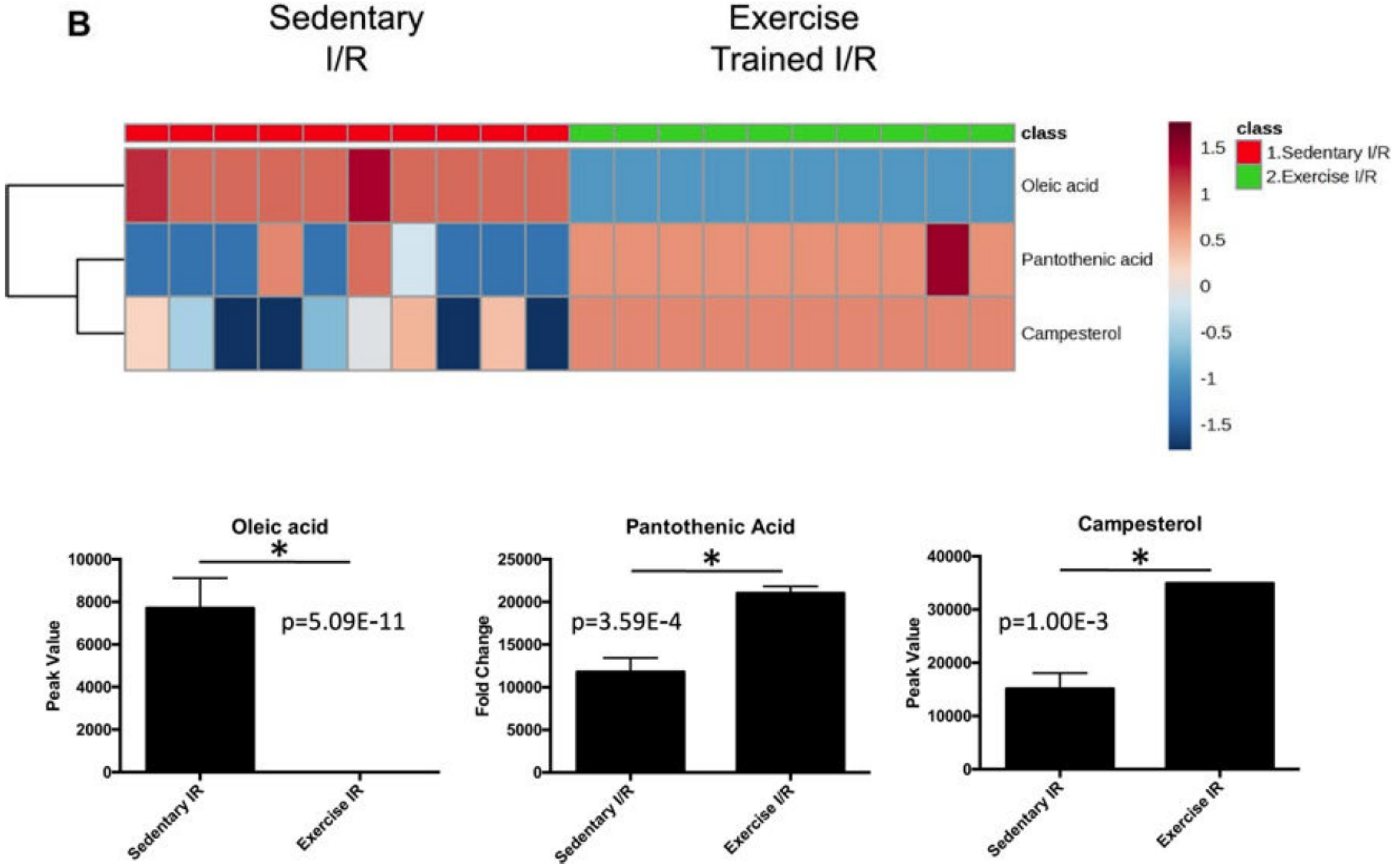
Statistical Analysis

- MetaboAnalyst (v3.0), an R package (v2.14.0), was used to detect metabolite peak areas (as representative of concentration).
- These data were scaled using the Pareto scaling feature.
- A one-way analysis of variance (**ANOVA**) and Fisher's Least Significance Difference (**LSD) posthoc test** across the groups (hearts from sedentary animals, sedentary hearts challenged with I/R, hearts from exercise-trained animals, and exercise-trained hearts challenged with I/R) were performed.
- ANOVA-significant metabolites (FDR < 0.05) were matched to metabolomics pathways using the Pathway Analysis, and Enrichment Analysis features in Metaboanalyst 3.0.
- Only metabolites identified and detected in all groups were included in the one-way ANOVA.
- Differences between sedentary and exercise-trained groups were compared using an independent t-test (2-tailed).
- Comparisons of increases after exercise training between muscle types were analyzed using a 2-tailed t-test followed by a one-way ANOVA.

Results

- Non-targeted GC–MS metabolomics analysis of 4 groups revealed 15 statistically significant metabolites between groups by ANOVA using MetaboAnalyst (p-value < 0.001).
- Enrichment analysis of these metabolites for pathway-associated metabolic sets indicated a > 10-fold enrichment for ammonia recycling and protein biosynthesis.
- Subsequent comparison of the sedentary hearts post-I/R and exercise-trained hearts post-I/R further identified significant differences in three metabolites (**oleic acid**, pantothenic acid, and campesterol) related to pantothenate and CoA biosynthesis ($p \leq 1.24E{-}05$, FDR $\leq 5.07E{-}4$).

Significantly altered metabolites comparing sedentary and exercise-trained hearts after ischemia-reperfusion injury by t-test analysis. Heatmap of t-test significant metabolites from sedentary and exercise-trained hearts after IR injury. N = 10 biological replicates/group.



Parry *et al.*’s conclusion

- Their study found novel mechanisms in which exercise-induced cardio-protection occurs in ischemia-reperfusion that complement both the mitochondrial stabilization and antioxidant mechanisms recently described.
- These findings also link protein synthesis and degradation (protein quality control mechanisms) with exercise-linked cardio-protection and mitochondrial susceptibility for the first time in cardiac I/R.

Issues with the Parry *et al*.

- No evidence of randomization or blocking
    - couldn't operate on 40 animals in a day, so there could be a day effect, time-of-day effect
- Why use the lowest value of the group where the missingness had occurred? Whether all missing values were in the same group?
- t-test is meaningless when one group has zero frequency for Oleic acid.
- There was no mention of the sex or age of the rats since adult male hearts are more susceptible to ischemia/reperfusion (I/R) injury as compared to pre-menopausal female hearts.

Overall Conclusions

- Brainstorm with your colleagues and senior faculty to decide on the experiment
- Experiment should be designed with consultation with the statistician and metabolomics assays provider
- Good design and good analytic methods can lead to reduced sample size and lead to valid meaningful results

Before you start experiment …
Remember the quote from Fisher:

Sir Ronald Aylmer Fisher (17 Feb 1890 - 29 Jul 1962)

“To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.”

Bagheri, Tiwari *et al*. A lipidome-wide association study of the lipoprotein insulin resistance index. Lipids Health Dis. 2020

- **Background:** The lipoprotein insulin resistance (LPIR) score is shown to predict insulin resistance (IR) and type 2 diabetes (T2D) in healthy adults. However, the molecular basis underlying the LPIR utility for classification remains unclear.
- **Objective:** To identify small molecule lipids associated with variation in the LPIR score, a weighted index of lipoproteins measured by nuclear magnetic resonance, in the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN) study (n = 980).
- LPIR is a combined weighted score of six lipoprotein subclasses or size parameters (VLDL, LDL, and HDL mean particle size; and levels of large VLDL, small LDL, and large HDL particle numbers).

Bagheri, Tiwari *et al*.

- **Methods:** Linear mixed-effects models were used to test the association between the LPIR score and 413 lipid species and their principal component analysis-derived groups. Significant associations were tested for replication with homeostatic model assessment-IR (HOMA-IR), a phenotype correlated with the LPIR score ($r = 0.48$, $p < 0.001$), in the Heredity and Phenotype Intervention (HAPI) Heart Study ($n = 590$).
- **Results:** In GOLDN, 319 lipids were associated with the LPIR score (false discovery rate-adjusted p-values ranging from 4.59×10^{-161} to 49.50×10^{-3}). Factors 1 (triglycerides and diglycerides/storage lipids) and 3 (mixed lipids) were positively ($\beta = 0.025$, $p = 4.52 \times 10^{-71}$ and $\beta = 0.021$, $p = 5.84 \times 10^{-41}$, respectively) and factor 2 (phospholipids/non-storage lipids) was inversely ($\beta = -0.013$, $p = 2.28 \times 10^{-18}$) associated with the LPIR score. These findings were replicated for HOMA-IR in the HAPI Heart Study ($\beta = 0.10$, $p = 1.21 \times 10^{-02}$ for storage, $\beta = -0.13$, $p = 3.14 \times 10^{-04}$ for non-storage, and $\beta = 0.19$, $p = 8.40 \times 10^{-07}$ for mixed lipids).

Bagheri, Tiwari *et al*.

- **Conclusions:** Non-storage lipidomics species show a significant inverse association with the LPIR metabolic dysfunction score and present a promising focus for future therapeutic and prevention studies.