Mappevurdering: Kvantiativ metode og statistikk IDR4000_1

Approx. 15,000 words

510

2024-11-22

Table of contents

Preface **5**

1 Reliability and tools for reproducible data science **6**
1.1 Introduction . . . 6
1.2 Method . . . 6
1.2.1 Standardization pre-test . . . 7
1.2.2 Equipment . . . 7
1.2.3 Protocol . . . 7
1.2.4 Post-test data preparation . . . 8
1.3 Results . . . 9
1.3.1 Calculation of typical error between test 1 and test 2 . . . 10
1.3.2 Calculation of typical error between test 3 and test 4 . . . 11
1.4 Discussion . . . 11
1.4.1 Reliability . . . 11
1.4.2 Deviance and source of error . . . 11
1.4.3 Future considerations . . . 12
1.5 Conclusion . . . 13

2 Regression models, predicting from data **14**
2.1 Introduction . . . 14
2.2 Method . . . 14
2.2.1 Part 1: Lactate thresholds . . . 14
2.2.2 Part 2: **Predicting sizes of DNA fragments, or slopes of a qPCR calibration curve** . . . 14
2.2.3 Part 3: Interpreting a regression table . . . 14
2.3 Results . . . 15
2.3.1 Part 1: Lactate thresholds . . . 15
2.3.2 Part 2: Predicting sizes of DNA fragments, or slopes of a qPCR calibration curve . . . 17
2.3.3 Part 3: Interpreting a regression table . . . 17
2.4 Discussion . . . 17
2.4.1 Part 1: Lactate thresholds . . . 17
2.4.2 Part 2: Predicting sizes of DNA fragments, or slopes of a qPCR calibration curve . . . 19
2.4.3 Part 3: Interpreting a regression table . . . 20

2.5 Conclusion . . . 20

3 Statistical inference **21**
3.1 Introduction . . . 21
3.2 Method . . . 21
3.3 Results . . . 21
3.3.1 Explain the estimate, SE, t-value, and p-value from the regression models that we created previously (m1 and m2). . . . 21
3.3.2 Discuss what contributes to the different results in the two studies (m1 and m2). . . . 22
3.3.3 Why do we use the shaded area in the lower and upper tail of the t-distribution. . . . 23
3.3.4 Calculate the standard deviation of the estimate variable, and the average of the se variable for each of the study sample sizes (8 and 40). Explain why these numbers are very similar. How can you define the Standard Error (SE) in light of these calculations? . . . 23
3.3.5 Create a histogram of the p-values from each study sample-size. How do you interpret these histograms, what do they tell you about the effect of sample size on statistical power? . . . 23
3.3.6 Calculate the number of studies from each sample size that declare a statistical significant effect. . . . 24
3.3.7 Using the pwr package, calculate the power of a one-sample t-test, with a effect size of 1.5/3, your specified significance level and sample sizes 8 and 40. Explain the results in the light of your simulations. . . . 25
3.3.8 With a significance level of 5%, how many studies would give you a "false positive" result if you did many repeated studies? . . . 25
3.4 Conclusion . . . 26

4 Study designs **27**
4.1 Introduction . . . 27
4.2 Study designs . . . 27
4.2.1 Cross-sectional study design . . . 28
4.2.2 Cohort study design . . . 28
4.2.3 Case series study design . . . 28
4.3 Selection of statistical test . . . 29
4.3.1 One-way ANOVA . . . 29
4.3.2 Person's chi-square test . . . 29
4.3.3 Two-sided t-test . . . 30
4.3.4 Odds ratio and confidence intervals . . . 30
4.3.5 Descriptive approach . . . 30
4.4 Inference . . . 30
4.5 Strength and weakness . . . 31
4.5.1 Cross-sectional study . . . 31

4.5.2 Cohort study design . . . 31
4.5.3 Case series study design . . . 32
4.5.4 General . . . 32
4.6 Future recommendations . . . 32
4.7 Conclusion . . . 33

5 Effects of resistance training volume on lean body mass and maximal strength 34
5.1 Introduction . . . 34
5.2 Methods . . . 35
5.2.1 Participants . . . 35
5.2.2 Study overview . . . 35
5.2.3 Data analysis . . . 36
5.3 Results . . . 37
5.3.1 Higher training volume results in greater regional hypertrophy . . . 37
5.3.2 Higher training volume results in greater muscular strength . . . 37
5.4 Discussion . . . 39
5.5 Conclusion . . . 40

6 Laboratory report 41
6.1 Introduction . . . 41
6.2 Materials . . . 41
6.3 Method . . . 42
6.3.1 Data Analysis . . . 42
6.4 Results . . . 43
6.4.1 Dilution Series . . . 43
6.4.2 Gene Expression . . . 43
6.5 Discussion . . . 43
6.5.1 Dilution Series . . . 43
6.5.2 Gene Expression . . . 45
6.5.3 Deviation . . . 46
6.6 Conclusion . . . 46

7 Philosophy of science 47
7.1 Provide a brief description of falsificationism and explain why Popper was motivated to develop this theory. Present one problem with the theory and assess whether the problem can be solved. . . . 47
7.2 Explain basic ideas of Bayesianism and how Bayesian probabilities can be interpreted. Present one problem with Bayesianism and evaluate how serious the problem is. . . . 48

References 50

Preface

This is the portfolio exam for IDR4000_1.

Each part has been produced by me or as group work in a class context.

Data and code for the different sections can be found in the following GitHub repository: https://github.com/loncir/mappeeksamen-idr4000-lc.git

1 Reliability and tools for reproducible data science

1.1 Introduction

Reliability is essential in both research and sports performance testing as it ensures consistent and accurate results over time. When a test is highly reliable, we can trust its findings and be confident that the experiment can be reproduced with similar outcomes. Moreover, high reliability allows researchers and coaches to track athletes' progress effectively, making sure that changes in performance are due to actual improvements rather than measurement errors. Without reliable measurements and data, it becomes difficult to draw meaningful conclusions or develop effective training programs.

The purpose of this report is to present estimates of reliability of measures collected from multiple VO_2max tests. The VO_2max test is often seen as the benchmark for assessing a person's aerobic capacity and cardiovascular fitness as it measures the maximum oxygen volume a person can use during intense exercise (Buttar, Saboo, and Kacker 2019). The test is therefore widely used in both clinical and sports settings as it is an important marker for endurance. On the other hand, its reliability can be affected by several factors, including the testing protocol, equipment calibration, participant motivation, and environmental conditions (Halperin, Pyne, and Martin 2015). Understanding these factors is critical for interpreting the results accurately and applying them effectively in training and research settings.

The VO_2max tests in our research project were performed in the physiology lab at Inland Norway University of Applied Sciences and we used Rstudio to analyze the data and to estimate if the performed tests were reliable.

1.2 Method

We performed a test-retest in our study where we gathered data from multiple VO_2max tests. 16 subjects performed two till four tests in the time span of 3 weeks. The first two tests were performed within 24 hours of each other during the first week. The last two tests were performed within 48 hours of each other during the third week.

1.2.1 Standardization pre-test

The results of a VO_2max test are influenced by a various of physiological and environmental factors, it is therefore important to set certain rules that the subject must follow in the days leading up to the test. A guideline may be sent out to the subject a couple of days before their first test with information regarding how they should prepare for test day. This guideline should include standardization practices that helps isolate the true aerobic performance of the subject and eliminates external factors that could influence the test results. Biological factors that are controlled for are exercise, hydration, sleeping schedule, caffeine and caloric intake, alcohol consumption and the time of day the test is performed. The subject then must follow the same schedule leading up to their next test to ensure the accuracy, reliability, and comparability of the test results.

1.2.2 Equipment

Specific equipment is needed to perform a VO_2max test. We had the subject perform the test on a cycle ergometer. This is a stationary bike that allows for incremental increases in resistance (W) during the test to progressively challenge the subject's aerobic capacity. We also used a system called Vyntus which measures the volume of oxygen consumed and carbon dioxide produced during the test. It has an automated software that helps us collect and analyze the data. There is also an integrated on-board pulse oximeter and automated volume and gas calibration. The subject had to wear a heart rate monitor attached to a chest strap in addition to a mouthpiece and a nose clip. The mouthpiece is connected by a hose to the mixing chamber that collects exhaled air that is analyzed by Vyntus. To measure the subject's lactate, we used a machine called Biosen which analyzes the blood and provides lactate values. Lastly, we had a stopwatch, an additional computer for our excel spread sheet and a scale to measure the subjects weight before starting the test. The whole setup allows us to monitor and measure the needed data to determine the subject's VO_2max as it gives us accurate information of the subject's oxygen consumption, heart rate, and cycling power output.

1.2.3 Protocol

The test protocol for performing a VO_2max test tells us how the test should be performed each time as standardization is important to get accurate and reliable results. A subject should therefore have the same test leader if a test is repeated. During our tests we adhered to the following protocol where a spread sheet in excel is prepared by the test leader before coming to the lab. When arriving, the test leader puts on a lab coat and ensures that Biosen (lactate measurement) is turned on. The machine then needs to be calibrated, and the results should come out to be 12 mmol/la. In the event of a calibration error, meaning that the instrument may provide inaccurate data, the standard fluid gets changed. Vyntus then needs to be calibrated by gas calibration and volume calibration. Gas calibration must be within

2.0 diff. and the volume calibration must be within 0.2 diff. Here the gas container needs to be opened, and the ventilation volume transducer (Triple-V turbine) must be attached to Vyntus. While the calibrations are in progress, the test leader puts together the mouthpiece, along with the nose clip, and attaches it with a hose to the mixing chamber. Then the cycle ergometer needs to be calibrated by positioning the crank arm straight up. This calibration is done on the computer connected to the ergometer as we use a different program here than Vyntus. The crank arm is looked at to make sure that it is 172.5 mm long and that the correct pedal type is attached. Lastly the equipment to measure lactate is collected and everything is ready for the subjects arrival.

A participant profile must be created in both Vyntus and the ergometer program. The profile includes the subjects name, date of birth, gender, height and weight. The first thing we do when the subject arrives is to measure weight, here the weight shown on the scale is subtracted by 300g. The cycling ergometer is then adjusted so that the subject is sitting comfortably on the bike. The bike settings are then saved in the cycling program to be used for future tests. Now the subject is ready for warm-up. The warm-up is 5 minutes long where the subject should have a progressively increase on the Borg Rating of Perceived Exertion (RPE) scale from 10 to 13. If the subject wants, the fan is turned on and faced towards them. The test leader then proceeds to inform about the test and Borg scale. Vyntus is then set to "measurement" and test leader makes sure that the heart rate monitor is connected to Vyntus, as well as making sure that the settings are set to 30 seconds measurements and size medium for the mouthpiece. The gas container gets closed, and Triple-V is disconnected from Vyntus and attached to the mixing chamber. Warm up is finished and the test gets started in Vyntus. The actual test starts when 1 minute has pasted in Vyntus and the VO_2max protocol starts on the cycling program. A stopwatch is also started and gets placed in front of the subject.

Throughout the test, the test leader informs about increases in resistance (W), normally 20W increases for women and 25W for men, and pushes the subject to pedal until exhaustion. The test stops when the subject gives up or the RPM drops below 60. The test leader writes down the maximum heart rate, end time, end wattage and borg score at the end of test in the excel spread sheet. Lactate is taken one minute after the test ends by wiping off the finger of the subject, poke a whole, wiping off the first drop of blood and then filling up the tube. The blood is then analyzed in Biosen. The subject is done testing and may leave, while test leader ends the test in both Vyntus and the cycling program. The bike, mouthpiece, hose and heart rate monitor get washed and the fan is faced towards the mixing chamber to let it dry.

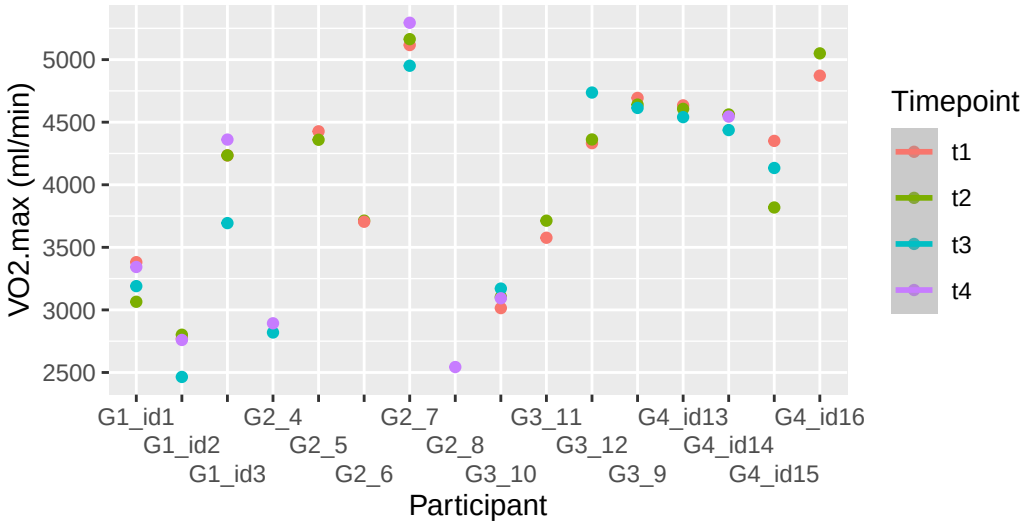
1.2.4 Post-test data preparation

Correctly gathering the data from the test is critical as it gives us insight into the subject's aerobic capacity and overall cardiovascular fitness. The report generated by Vyntus is saved onto a memory stick so it can later be saved to OneDrive. The excel spread sheet is then filled with values from the report. These values include absolute VO_2max expressed in milliliters of oxygen per minute (ml/min), giving us the total oxygen consumption of the subject which

is needed for assessing overall capacity (Buttar, Saboo, and Kacker 2019). The maximum volume of carbon dioxide (CO_2) produced which tells us how much CO_2 is being expelled by the subject's body during the test (Buttar, Saboo, and Kacker 2019). Respiratory exchange ratio (RER) which is used to estimate the ratio of carbon dioxide (CO_2) produced to oxygen (O_2) consumed during metabolism (Buttar, Saboo, and Kacker 2019). The maximum minute ventilation which is the highest volume of air the subject can move in and out of their lungs per minute (Buttar, Saboo, and Kacker 2019). The breathing frequency maximum which is the maximum number of breaths per minute the subject takes at peak exercise intensity (Buttar, Saboo, and Kacker 2019). Lastly, the lactate measurements are gathered from the Biosen machine. Maximum heart rate, end time, end wattage and borg score are already filled out from end of test while ID, period, timepoint, date, time of day, temperature, humidity, sex, age, height and weight was filled out pre-test. We then got the data from the other groups and collected it all in one excel spreadsheet. The data was then exported into RStudio where we further analyzed our results.

1.3 Results

We chose to analyze the relationship between absolute VO_2max and timepoint.



Shows the total volume of oxygen consumed by each participant at each timepo much the results varied between tests. The timepoints are color−coded where e its own color. Ideally, each participant's data points should be closely grouped to than spread apart.

Figure 1.1: VO_2max per participant per test

Table 1.1: VO_2max per participant per timepoint and difference between timepoints

Participant ID	Test 1	Test 2	Test 3	Test 4	Difference (T2 - T1)	Difference (T4 - T3)
G1_id1	3,381.50	3,065.00	3,190.00	3,343.00	-316.50	153.00
G1_id2	2,771.00	2,801.50	2,464.50	2,760.00	30.50	295.50
G1_id3	4,234.50	4,235.00	3,693.50	4,361.00	0.50	667.50
G2_4	NA	NA	2,819.50	2,893.00	NA	73.50
G2_5	4,427.00	4,359.50	NA	NA	-67.50	NA
G2_6	3,704.50	3,713.50	NA	NA	9.00	NA
G2_7	5,116.50	5,163.50	4,951.00	5,294.50	47.00	343.50
G2_8	NA	NA	NA	2,543.50	NA	NA
G3_9	4,694.00	4,640.50	4,614.00	NA	-53.50	NA
G3_10	3,014.50	3,103.50	3,170.50	3,093.00	89.00	-77.50
G3_11	3,576.50	3,713.00	NA	NA	136.50	NA
G3_12	4,332.50	4,362.00	4,737.00	NA	29.50	NA
G4_id13	4,634.50	4,606.50	4,540.50	NA	-28.00	NA
G4_id14	4,556.50	4,561.50	4,437.00	4,545.00	5.00	108.00
G4_id15	4,350.50	3,818.50	4,134.00	NA	-532.00	NA
G4_id16	4,872.00	5,050.00	NA	NA	178.00	NA

Sample sizes: t1 = 14 , t2 = 14 , t3 = 11 , t4 = 8
Data shows the total volume of oxygen consumed by each participant at each timepoint, as well as the difference between timepoints. Sample sizes is indicatet at the bottom clearly showing that not all participants completed all four tests. Missing data, indicated by NA entries, may have impacted the study results, potentially affecting reliability and overall conclusions.

TRUE

Table 1.2: Typical error between test 1 & 2

mean	sd	te	cv
4,102.1	183.5	129.8	3.2

TRUE

1.3.1 Calculation of typical error between test 1 and test 2

As shown in Table 1.2, the average VO_2max across timepoint 1 and 2 (mean) came out to be 4102.1 mL/min suggesting that, on average, the participants have a high aerobic capacity. The variability of VO_2max scores across the participants (sd) came out to be 183.5 mL/min. This shows a moderate spread in VO_2max values around the mean, implying that while the group has similar fitness levels, there is still some individual variation in performance. Additionally, the typical error (te) measures the inconsistency between the two timepoint measurements for each participant. A typical error of 129.8 means that the typical fluctuation in VO_2max readings between timepoint 1 and timepoint 2 is about 130 mL/min. Lastly, we calculated the coefficient of variation (cv) between test 1 and test 2 to be 3.2%.

Table 1.3: Typical error between test 3 & 4

mean	sd	te	cv
4,102.1	240.9	170.3	4.2

TRUE

1.3.2 Calculation of typical error between test 3 and test 4

As shown in Table 1.3, the average VO_2max across timepoint 2 and 3 (mean) came out to be 4102.1 mL/min. This is exactly the same average as between timepoint 1 and 2, indicating that, on average, the aerobic capacity of the participants did not change during the 3 weeks of testing. The variability of VO_2max scores across the participants (sd) came out to be 240.9 mL/min. This shows a spread in VO_2max values around the mean that is higher than for the previous 2 tests, implying that there is a higher individual variation in performance for test 3 and 4. Additionally, the typical error (te) came out to be 170.3 meaning that the typical fluctuation in VO_2max readings between timepoint 3 and timepoint 4 is about 170 mL/min. This is almost 40 mL/min more than between test 1 and 2. Lastly, we calculated the coefficient of variation (cv) between test 3 and test 4 to be 4.2%.

1.4 Discussion

1.4.1 Reliability

In our analysis of the reliability of VO_2max measures across four testing timepoints, we found that our results varied when comparing the first two tests and the latter two. The results for the first two tests (Table 1.2) indicated a high level of reliability and consistency between tests. The calculated coefficient of variation (CV) of 3.2% is, according to Hopkins, indicative of good reliability as it is below 5% (Hopkins 2000). Additionally, Hopkins notes that shorter intervals between tests can yield lower CV's because the individual's physiological state remains more constant, as seen with our findings (Hopkins 2000). However, as we move to the third and fourth tests, there is a slight increase in variability which could point to potential factors affecting performance, such as variations in testing conditions, or differing motivation levels over time (Jones and Carter 2000). This finding also aligns with Hopkins theory that the reliability of measures may decrease as the interval between tests increases as the time between test 3 and 4 is longer (48 hours) than between the first two tests (24 hours) (Hopkins 2000). However, the CV value for test 3 and 4 is still within a reasonable range, as it is below 5%, and can therefore be considered reliable. Our findings is similar to that of other studies as Astorino et al.found that VO_2max testing protocols generally achieve a CV of around 3-5%, emphasizing the importance of using standardized procedures (Astorino et al. 2005). All together, our findings indicates that the testing protocol is effective in capturing the participants true aerobic capacities without significant variability.

1.4.2 Deviance and source of error

In VO_2max testing several factors can introduce deviance and error, affecting the reliability of the results. Physiological variability is a primary source of error, as day-to-day changes

in hydration, nutrition, fatigue, and even motivation can lead to fluctuations in performance (Miller and Mchugh 2014). This is why we sent out a guideline to each participant controlling for biological factors. By doing so we isolated the true aerobic performance of the subject and eliminated external factors that could influence the test results. The participant did also have the same test leader for each test, who would give the same amount of encouragement and feedback every time, meaning that the external motivation would be the same for every timepoint. Internal motivation may be harder to control as this refers to the participants own drive to push themselves to exhaustion and depends on the participants own personal interest and sense of fulfillment from the activity itself (Miller and Mchugh 2014). Therefore, the difference in internal motivation from each timepoint may have influenced the performance of the participants and lead to different exertions of effort.

Additionally, tester error or inconsistencies in the protocols can contribute to variation between tests.The majority of the test leaders were new to the test protocol and administered their first ever VO_2max test during test 1. As a result, the test leader may have administered the test a little different for each timepoint as they gained more experience and knowledge. Ideally each leader should have gone through the protocol and practiced a couple of times before administering the test as an experienced test leader is likely to administer the test more consistently. Research even suggest that a skilled administrator can minimize the effects of external factors such as motivation and encouragement, which can skew results if not managed appropriately(Buchheit et al. 2010). Moreover, an experienced test leader is typically more skilled at operating testing equipment which may minimize human errors such as incorrect calibration or data handling that would otherwise introduce variability in the results. In other words, leaders with a strong technical background are better equipped to avoid these pitfalls, ensuring that the data collected is accurate and reliable (Marchetti, De Almeida, and Alvares 2016).

Variability in VO_2max measurements may also stem from the testing equipment. It is crucial that the equipment used is reliable so one can obtain accurate and consistent results. Several factors may influence the reliability of the equipment, including calibration and maintenance. We calibrated the Vyntus by gas calibration and volume calibration before each test to ensure accuracy as equipment that is not calibrated correctly can generate inaccurate measurements. According to a study by Routledge et al., consistent calibration and adherence to manufacturer guidelines can significantly reduce variability in measurements during exercise testing (Routledge 2015). Technical errors with the ergometer or the software programs can also affect VO_2max readings. Inaccuracies in these devices can lead to systematic errors that affect the reliability of the results. Reliable equipment and strict adherence to testing protocol is therefore crucial in ensuring accurate and consistent measurements.

1.4.3 Future considerations

To enhance the precision and reliability of future studies measuring VO_2max, several improvements may be recommended. To obtain reliable estimates of reliability, Hopkins suggests at

least 50 participants and at least 3 trials (Hopkins 2000). Our study fell short of this suggestion as we only had 16 participants, which may have affected the precision of our estimates. Additionally, not all participants completed all four tests, as there were only 14 for the first two, 11 for the third and only 8 for the fourth (see Table 1.1). This further complicated our analysis and may have introduced bias, suggesting a need for improved participant retention strategies in future studies.

Minimizing sources of error may be done through careful control of test conditions while ensuring consistent procedures across all timepoints. It is important to implement standard protocols, like warm-up routine, ergometer adjustments and calibration procedures, as it helps to make sure that any changes we see are accurate and not caused by inconsistent testing. Similarly, training the test leader thoroughly on equipment handling and data recording will help reduce human errors and improve the reliability of measurements.

1.5 Conclusion

In conclusion, understanding and measuring reliability is vital for producing trustworthy results from a VO_2max test. Our approach had both strengths and limitations as the initial test demonstrated good reliability, but we also saw an increase in variability between later tests. These findings indicate that there may be factors affecting the precision of our results. Addressing these sources of error is essential to improve measurement consistency and may strengthen future research. Overall, our findings emphasize the importance of standardized protocols and equipment calibration in VO_2max testing to ensure reliable and accurate assessment of aerobic performance.

2 Regression models, predicting from data

2.1 Introduction

The purpose of this report is to predict data using regression analysis performed in RStudio, as well as to interpret a regression table. The report consists of three sections. By determining the lactate threshold at blood lactate values of 2 and 4 mmol L^{-1}, we analyze the relationship between performance in watts and training intensity. We also analyzed the slope of a qPCR calibration curve and interpreted a regression table on the relationship between 3RM squat and the cross-sectional area of type II muscle fibers.

2.2 Method

2.2.1 Part 1: Lactate thresholds

In part 1 of the report, we used the dataset *cyclingstudy* from (Sylta et al. 2016) to predict two specific blood lactate thresholds at 2 and 4 mmol L^{-1}. The data analysis was conducted in RStudio.

2.2.2 Part 2: Predicting sizes of DNA fragments, or slopes of a qPCR calibration curve

In part 2 of the report, using (Schindelin et al. 2012), we analyzed an image of qPCR obtained from the experiment " "DNA extraction and analysis" ("Login - eLabFTW — Elab.inn.no"). The image analysis provided data, which we applied in RStudio to predict the slope of the qPCR calibration curve.

2.2.3 Part 3: Interpreting a regression table

In part 3 of the report, we conducted a statistical analysis of the relationship between Type II (FAST) fibers cross-sectional area (μm^2) at baseline (FAST_CSA_T1) and Squat 3 repetition maximum load (kg) at baseline (SQUAT_3RM) from the dataset of (Haun et al. 2018) and (Haun et al. 2019) to investigate whether there was a linear relationship.

Table 2.1: Lactate threshold calculations for 2 mmol L^{-1}

watt	predictions
271.8	1.999125

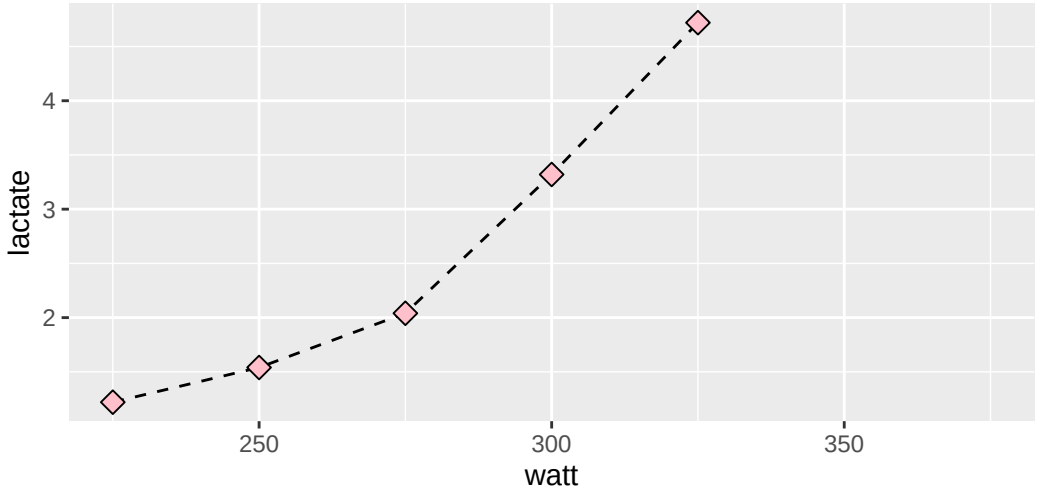
Table 2.2: Lactate threshold calculations for 4 mmol L^{-1}

watt	predictions
314.2	3.998828

2.3 Results

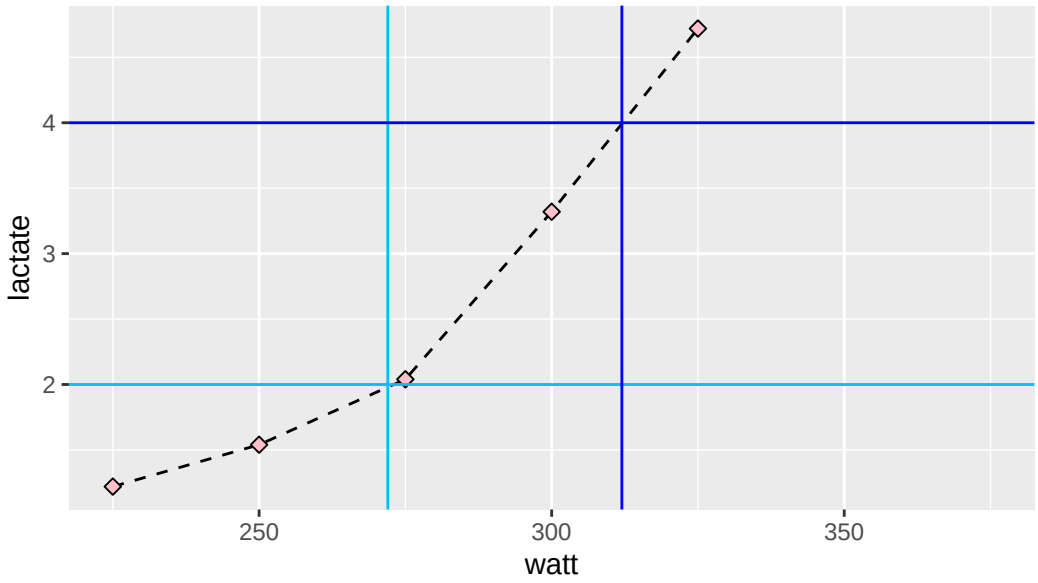
2.3.1 Part 1: Lactate thresholds

We found that a fourth-degree polynomial model was the best fit for our data (see Figure 2.4). We predicted the blood lactate threshold at 2 mmol L^{-1} to be 271.8 watts (see Table 2.1) and the blood lactate threshold at 4 mmol L^{-1} to be 314.2 watts (see Table 2.2).



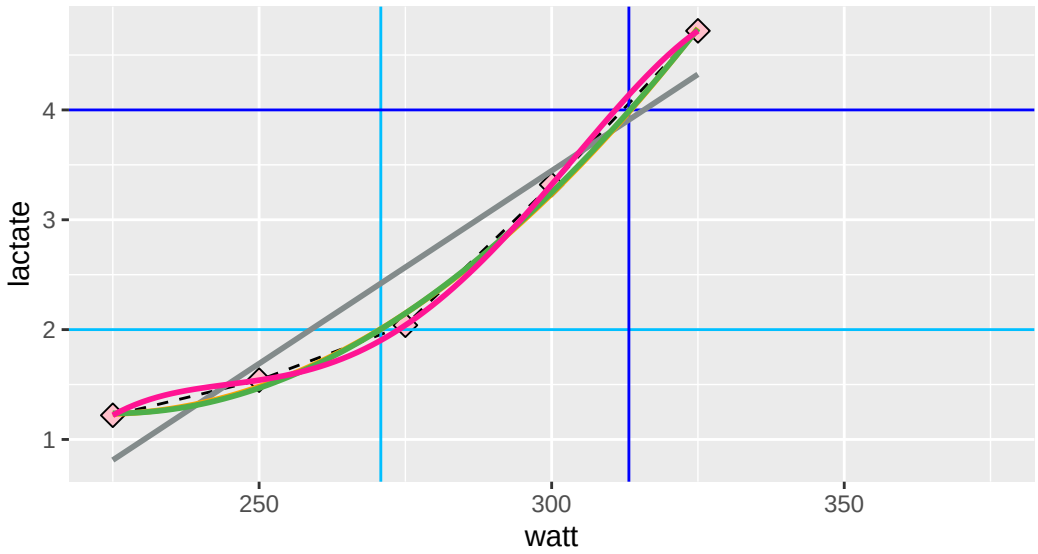
The relationship between workload (measured in watts) and lactate concentration fo at the pre−exercise time point. Each point on the graph represents lactate concentra measured at a specific workload. The dashed line connects the points for subject 3, the trend in lactate response as workload increases. A positive correlation is seen, lactate levels increase with higher workloads.

Figure 2.1: Relationship between workload (watt) and lactate



The value of x (watt) when y (lactate) is set to 2 and 4. The light blue lines is the app threshold for 2 mmol L−1. The dark blue lines is the approximate lactate threshold f

Figure 2.2: Estimated exercise intensity at 2 and 4 mmol L^{-1}



The grey line represents the linear relationship between exercise intensity (watt) and Orange is a second degree polynomial model. Green is a third degree polynomial m degree polynomial model

Figure 2.3: Curve-linear relationships between exercise intensity and blood lactate

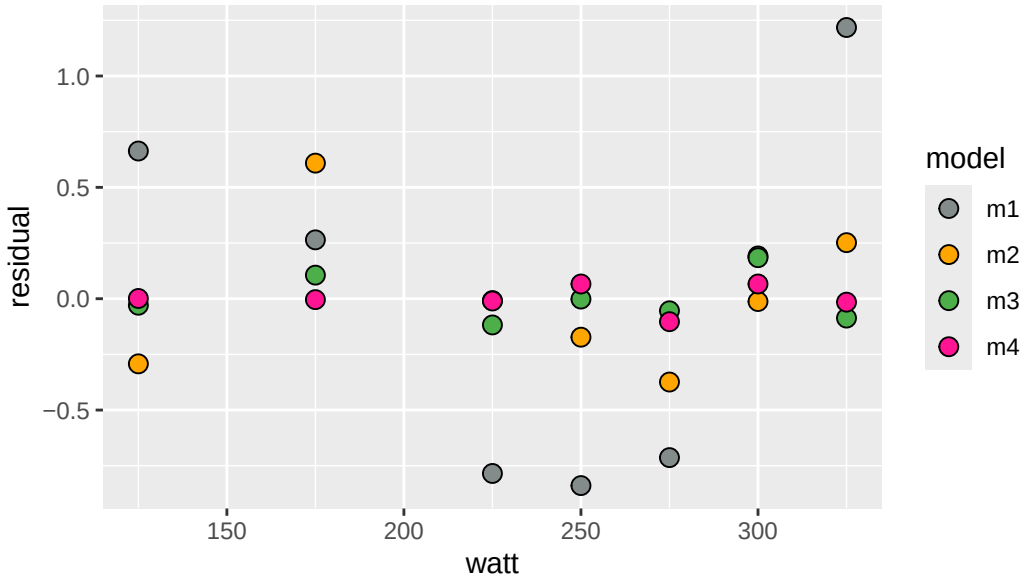


Figure 2.4: Assessing the fit of different linear models

2.3.2 Part 2: Predicting sizes of DNA fragments, or slopes of a qPCR calibration curve

2.3.3 Part 3: Interpreting a regression table

The results show no correlation between SQUAT 3RM (kg) and FAST CSA T1 (μm^2) (Estimate = 5.483, SE = 8.032, t-value = 0.683, p-value = 0.5), see Figure 2.6.

2.4 Discussion

2.4.1 Part 1: Lactate thresholds

Our results show the relationship between workload (in watts) and lactate concentration for subject 3 at the pre-exercise time point. Figure 2.1 shows a clear positive correlation, indicating that as workload increases, so does lactate concentration, which is expected in endurance exercise due to increased anaerobic metabolism (Facey, Irving, and Dilworth 2013).

Figure 2.3 contrasts the linear model (shown in gray) with polynomial regression models (in orange, green, and pink), which better capture the non-linear relationship between exercise intensity and lactate production. Notably, the linear model deviates significantly from the

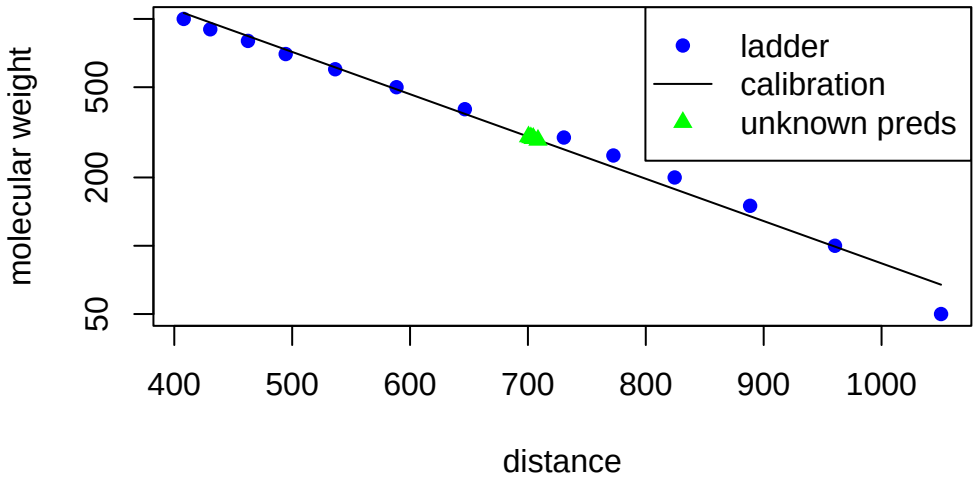


Figure 2.5: Distance vs. Molecular weight. Illustrates the relationship between molecular weight and distance as measured in a molecular weight ladder. The blue points represent the known molecular weights corresponding to their distances in the gel. The fitted line (in black) shows the calibration curve derived from the ladder data, illustrating a clear relationship. The green points represent the predicted molecular weights for unknown samples based on their measured distances

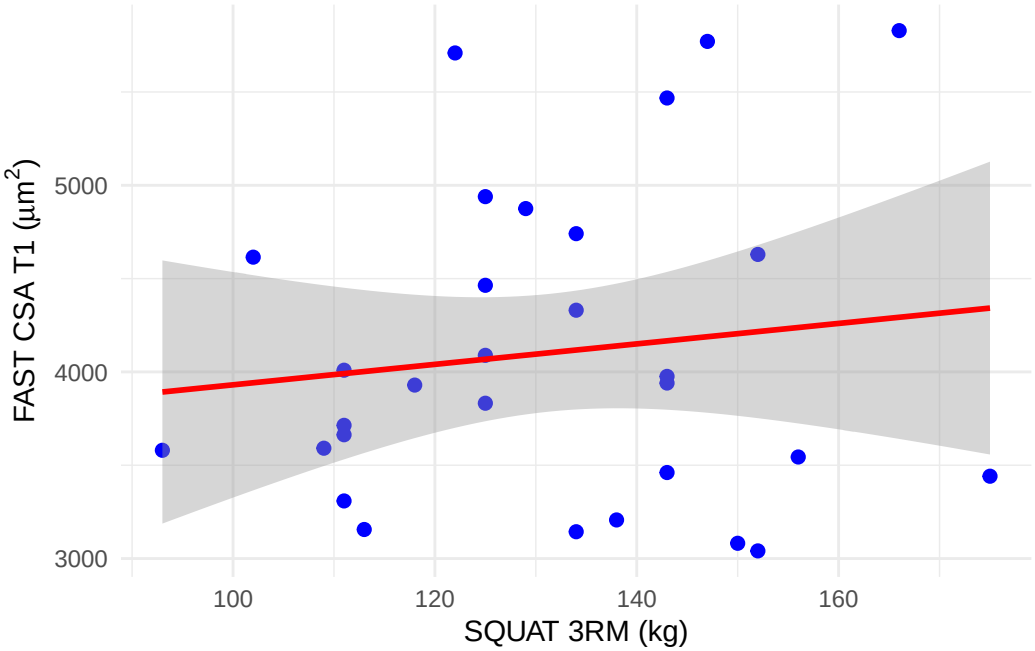


Figure 2.6: Correlation between Type II (FAST) fibers cross-sectional area and Squat 3 repetition maximum load (kg)

observed data around the 2 mmol L ^{1} mark, indicating it is a poor fit for the data. In contrast, the polynomial models closely align with the observed values, particularly the fourth-degree polynomial model (m4, pink) showcased in Figure 2.4, which demonstrates the best fit as residuals are scattered around zero. This is no surprise as polynomial models with increased degrees has more flexibility to fit the data (Hammarstrom 2024). However, the calculation of the profile curve was performed using a third-degree polynomial model (m3), as you cannot fit a fourth-degree polynomial model with only four data points (Hammarstrom 2024). This way we also prevented our model from becoming more sensitive to bad measurements. Additionally, the use of a third-degree polynomial model has previously been shown to be effective for estimating a true lactate curve, as demonstrated in a prior study by (Newell et al. 2007).

Our predictions gave us the lactate thresholds of 271.8 watts for 2 mmol L ^{1} and 314.2 watts for 4 mmol L ^{1}, providing essential insights into subject 3's physiological response to exercise, which can inform training strategies and enhance performance outcomes.

2.4.2 Part 2: Predicting sizes of DNA fragments, or slopes of a qPCR calibration curve

The calibration model describes the relationship between distance and molecular weight. After log-transforming the molecular weight, the relationship between distance and molecular weight

becomes approximately linear, making a linear regression model appropriate. The predicted molecular weights for the unknown samples are based on the fitted calibration model. These predictions provide estimates of their molecular weight based on their migration distance in the gel.

The QQ plot shows whether the differences between observed and predicted molecular weights follow a normal distribution. Ideally, the points in the QQ plot should fall along the reference line, which they appear to do, indicating that the model is well-specified and captures the relationship between migration distance and molecular weight.

2.4.3 Part 3: Interpreting a regression table

The results show no correlation between FAST_CSA_T1 and SQUAT_3RM (Estimate = 5.483, SE = 8.032, t-value = 0.683, p-value = 0.5). The standard error (SE) explains how much the mean from our sample is expected to deviate from the same mean in the population (Spiegelhalter 2019). The p-value indicates that we would observe a similar or more extreme result in 50% of cases if we repeat the study, assuming the null hypothesis is true (Spiegelhalter 2019). The t-value suggests that the difference between the sample mean and the population mean is likely to be small (Spiegelhalter 2019). The low t-value, combined with the high p-value, indicates that there is no statistical basis for claiming that the difference between the sample mean and the population mean is significant. In summary, these findings suggest that there is no basis for saying that there is a significant correlation between the increase in SQUAT_3RM weight (kg) and the increase in FAST_CSA_T1 in micrometers (μm^2). At the same time, the observed increase in FAST_CSA_T1 may instead be influenced by random variations rather than a systematic effect of increased strength.

2.5 Conclusion

In conclusion, the analysis conducted in this report provided key insights into the relationships between different physiological variables and their predictive models. In part 1, the lactate thresholds at 2 and 4 mmol L^{1} where successfully determined for subject 3, with our polynomial regression models offering a more accurate fit compared to linear models. Part 2 focused on predicting molecular weights from a calibration curve, where a strong linear fit was observed between the distance traveled in the gel and the logarithm of molecular weight. Lastly, part 3 evaluated the relationship between muscle fiber cross-sectional area and squat performance, where no significant correlation was found, suggesting that the observed difference may be due to random variations rather than any specific relationship between muscle size and squat performance. Together, these findings underline the utility of regression models in making predictions and interpreting data.

3 Statistical inference

3.1 Introduction

The study was set up as a statistical laboratory, were we performed simulations. The purpose of this rapport is to interpret and explain the results we got.

3.2 Method

We simulated a population of possible values and then drew random samples, calculated statistics and interpreted them. The population of values was regarded as the possible difference between two treatments in a cross-over study where participants performed both treatments. The values in the population were calculated as Treatment - Control. We simulated a population of one million numbers with a mean of 1.5 and a standard deviation of 3. We then made two different set of studies, one set with a sample size of 8 (samp1) and one set with a sample size of 40 (samp2). Additionally, we estimated the average value of the population.

We drew two random samples corresponding sample sizes of 8 and 40 and saved this data in data frames with the dependent variable y. Then the model were fitted as a linear model and saved as a model object. Object m1 corresponds to a sample size of 8, while m2 corresponds to a sample size of 40. Our null hypothesis is that there is no difference between the two treatments.

3.3 Results

3.3.1 Explain the estimate, SE, t-value, and p-value from the regression models that we created previously (m1 and m2).

In our model, the estimate represents the mean of the differences between the two treatments in the cross-over study. In model m1, the estimate is 1.84. This means that the average difference between the two treatments for the sample of 8 participants is 1.84. In model m2, the estimate is 1.56, meaning that the average difference between the two treatments for the sample of 40 participants is a little lower than for the sample of 8. Furthermore, the standard error (SE) provides an estimate of how much variability we expect in the sample mean if we

were to repeatedly draw samples of the same size from the population. It is calculated as the sample's standard deviation (SD) divided by the square root of the sample size. In m1, the standard error (SE) is 1.25, which tells us how much the sample mean of 1.84 might vary if we were to repeat the study multiple times with a sample size of 8. In m2, the standard error is 0.48 indicating that the sample size of 40 participants gives us a more precise estimate of the population mean. The t-value is a ratio that compares the difference between the sample mean (estimate), and the null hypothesis relative to the standard error (SE). In m1, the t.value is 1.47, meaning the observed mean difference (1.84) is 1.47 standard errors away from the null hypothesis value of 0. The t.value for m2 is almost three times bigger as it is 3.28 indicating that the observed mean difference (1.56) is 3.28 standard errors away from the null hypothesis. Lastly, the p-value tells us the probability of observing a t-value as extreme (or more extreme) than the one calculated, assuming the null hypothesis is true (i.e., no difference between treatments). In m1, the p-value is 0.185, meaning that there is an 18.5% chance of observing a difference of 1.84 or more if the true difference between the treatments was zero. Since this p-value is above the conventional threshold of 0.05, we fail to reject the null hypothesis, suggesting that the observed difference is not statistically significant. In m2, the p-value is 0.002, meaning that there is an 0.2% chance of observing a difference of 1.56 or more if the true difference between the treatments was zero. In comparison to m1, this p-value is below the conventional threshold of 0.05. We therefore reject the null hypothesis, suggesting that the observed difference is statistically significant and there is evidence that the means differ.

3.3.2 Discuss what contributes to the different results in the two studies (m1 and m2).

The two studies differ primarily in sample size, where m2 have 5 times more participants than m1. Since m1 have a small sample, the mean might fluctuate more due to random variation, whereas larger samples (m2) tend to provide a more stable and reliable estimate closer to the true population mean (Faber and Fonseca 2014). Furthermore, the larger the sample size, the smaller is the standard error, which means a more precise estimate of the population mean (Faber and Fonseca 2014). In m2, the larger sample size leads to a smaller standard error (0.48), which reduces the uncertainty around the estimate and increases the power of the test to detect differences. The t-value is influenced by both the estimate and the standard error. Even if the estimates are somewhat similar, the smaller standard error in m2 results in a larger t-value, making it more likely to detect a significent effect. Additionally, the p-value depends on the t-value. With a larger sample size, as in m2, the t-value is typically larger, leading to a smaller p-value.This means that m2 is more likely to detect significant differences than m1, where the small sample size leads to a higher p-value and lower statistical power. In conclusion, the larger sample size in m2 leads to a more precise estimate, a smaller standard error, a higher t-value, and ultimately a smaller p-value, increasing the likelihood of detecting a significant difference between the treatments.

3.3.3 Why do we use the shaded area in the lower and upper tail of the t-distribution.

The shaded area in the lower and upper tail of the t-distribution represents the probability of observing extreme values (both high and low) of the t-value under the null hypothesis. This area helps us determine the p-value, which tells us how likely it is that our observed data could occur by random chance if the null hypothesis is true.

The total shaded area in both tails represents the combined probability of observing a t-value as extreme as the one calculated, assuming the null hypothesis is true. The p-value for m1 where 0.185, meaning that 18.5% of the area is in the combined tails, representing the threshold for statistical significance.

3.3.4 Calculate the standard deviation of the estimate variable, and the average of the se variable for each of the study sample sizes (8 and 40). Explain why these numbers are very similar. How can you define the Standard Error (SE) in light of these calculations?

By calculating the standard deviation of the estimates across the 1000 studies we get a measure of how much the sample means fluctuate between different samples of the same size. For the smaller sample size of 8, the standard deviation comes out to be 1.07, while for the bigger sample size of 40 it comes out to be 0.48. Furthermore, the average standard error represents the average uncertainty of the sample mean estimate in each study. It reflects the variability in the sample means and depends on the sample size. The average standard error for the smaller sample size is 1.02, and 0.47 for the bigger sample size. Why are these numbers very similar? The standard deviation of the estimates and the average standard error are conceptually related, as the standard error estimates how much the sample mean might vary from the population mean. They are both measures of variability, but while standard error (SE) is an estimate based on the sample, the standard deviation of the estimates shows the actual observed variation across multiple studies. Therefore, in light of these calculations, we can define standard error (SE) as the expected variability from sample to sample.

3.3.5 Create a histogram of the p-values from each study sample-size. How do you interpret these histograms, what do they tell you about the effect of sample size on statistical power?

For sample size 8, the p-values are more spread out, with many values above the significance threshold of 0.05 (see Figure 3.1). This indicates that with a smaller sample, there is lower statistical power, and many studies fail to detect a statistically significant difference. For sample size 40, the p-values are more concentrated towards lower values, indicating that a larger sample size increases the likelihood of detecting significant effects. This reflects increased

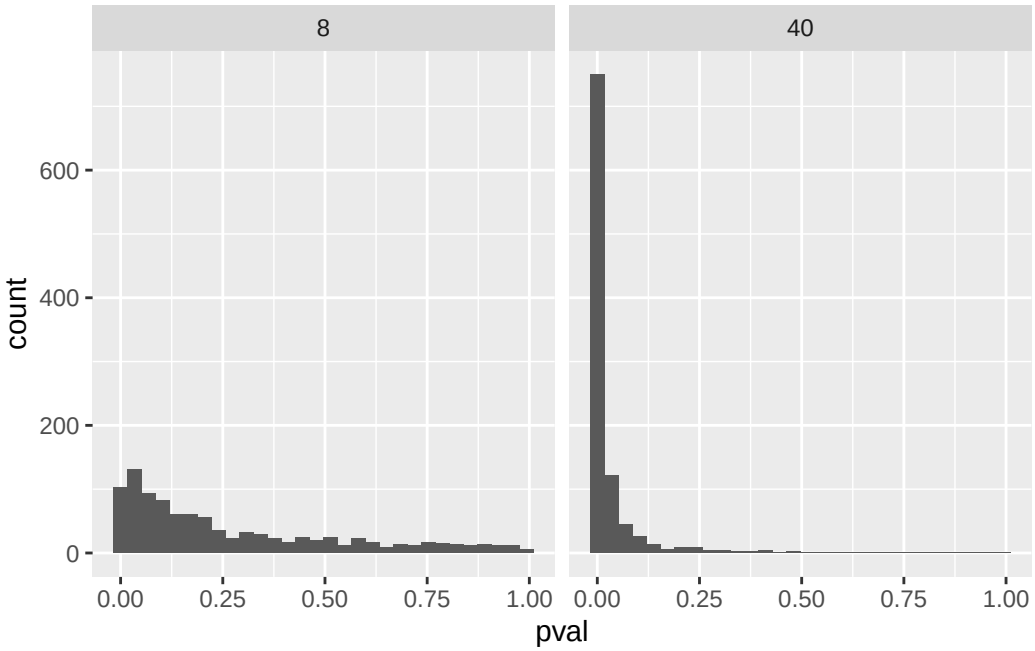


Figure 3.1: p-values from each study sample-size

Table 3.1: Statistical significant effect

n	sig_results
8	0.227
40	0.865

statistical power with larger sample sizes, meaning the test is more likely to reject the null hypothesis when a true effect exists.

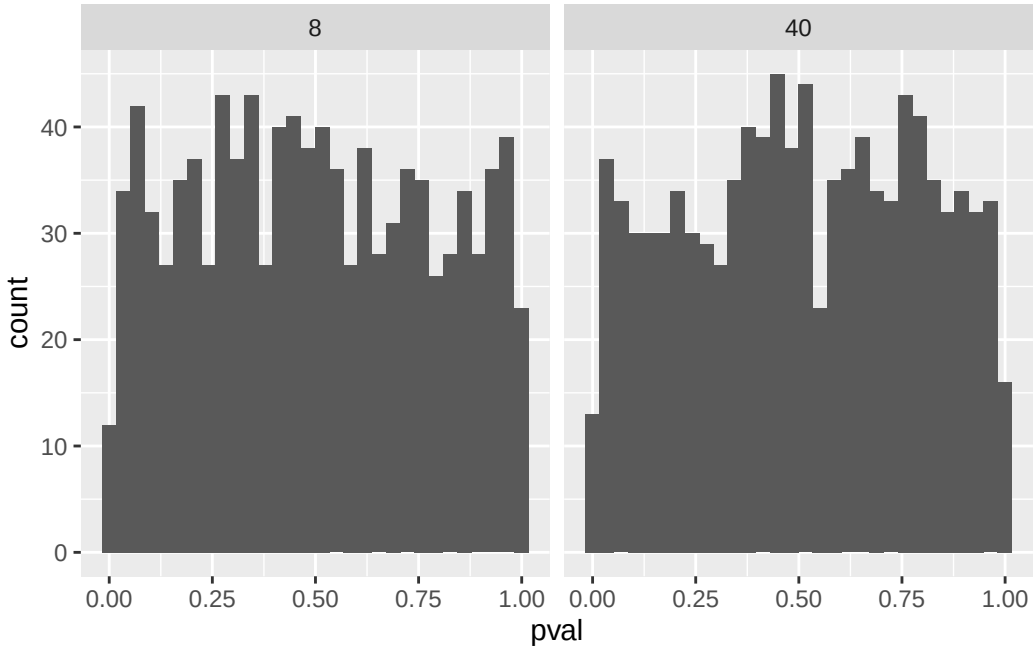
3.3.6 Calculate the number of studies from each sample size that declare a statistical significant effect.

When calculating the number of studies that declare a statistical significant effect we find that fewer studies (0.227) are likely to reach statistical significance in the sample size of 8 (see Table 3.1). This is because of the low statistical power associated with small samples. In the sample size of 40, more studies (0.865) will show a significant effect due to the higher power of larger samples, making it easier to detect true differences.

3.3.7 Using the pwr package, calculate the power of a one-sample t-test, with a effect size of 1.5/3, your specified significance level and sample sizes 8 and 40. Explain the results in the light of your simulations.

To calculate the power of a one-sample t-test, we use the effect size $= /$, where $= 1.5$ (mean) and $= 3$ (standard deviation). The effect size $= 1.5/3 = 0.5$. For a sample size of 8 we find that the power will be relatively low, 0.232, reflecting that with smaller samples, we have a lower chance of detecting a true effect when it exists (Faber and Fonseca 2014). This corresponds to fewer studies achieving significance in the simulation. For a sample size of 40 we find that the power will be much higher, 0.869, indicating that with a larger sample size, we have a higher chance of detecting a true effect. This explains why more studies with sample size 40 declare significance in the simulation.

We will now simulate a population without differences between treatment and control. The code used is very similar to the one we use above, except that we use an average effect of 0 in the population.



3.3.8 With a significance level of 5%, how many studies would give you a "false positive" result if you did many repeated studies?

To determine how many studies would give a "false positive" result with a significance level of 5%, we calculate how many studies produce a p-value less than 0.05 in this simulation. Since the population mean is set to 0 any p-value below 0.05 in this scenario would represent a false

positive. For the sample size of 8 we got 44 while for the sample size of 40 we got 49. With a significance level of 5%, we expect roughly 5% of the 1000 studies to produce p-values below 0.05 due to random variation, even though the null hypothesis is true. For 1000 studies, this means we expect around 50 false positive results for each sample size. The actual number of false positives may vary slightly due to the randomness in the simulation but should be close to 50 for each sample size (8 and 40).

3.4 Conclusion

In this study, we explored the impact of sample size on statistical results by conducting simulations across different study sizes (8 and 40 participants). We observed that larger sample sizes lead to more precise estimates, lower standard errors, and increased t-values, resulting in a higher likelihood of detecting statistically significant effects. On the other hand, smaller sample sizes had higher variability in estimates and standard errors, contributing to lower statistical power and fewer significant results. Our findings also emphasize the importance of sample size in improving the reliability of statistical tests, as demonstrated by the power calculations. Larger samples provide greater statistical power, allowing for more accurate detection of true effects while maintaining the expected false positive rate under the null hypothesis. Overall, this simulation highlights how increased sample sizes lead to more robust and reliable conclusions in research.

4 Study designs

4.1 Introduction

A good study design is crucial for producing reliable and meaningful research results. It minimizes bias and ensures that the results are valid, reproducible and generalizable. In this report we've analyzed the study design of five original research studies who tried to understand the association between menstrual cycle and musculoskeletal injury among female athletes. We describe strength and weaknesses of each study and comment on the selection of statistical tests to answer study aims. Lastly, recommendations are given regarding how future studies in this area should be designed to best answer similar questions.

The broader problem the authors are trying to resolve is similar in each study, but the specific questions they are trying to answer is angled differently. Chang et.al investigates the prevalence and relationship between hormonal contraceptive use, menstrual function and stress fractures in female collegiate athletes in the United States (Cheng et al. 2021). Bingzheng et.al is also focusing on female collegiate athletes but limits it down to just soccer players. Their study aims to examine how the menstrual cycle and sex hormones affect knee kinematics, specifically during a 90-degree cutting maneuver (Bingzheng et al. 2023). Female soccer players are also the population of Martin et.al study, specifically English national team players, as they aimed to enhance the understanding of how the menstrual cycle, including extended cycle length, may impact injury risk in said population (Martin et al. 2021). On the other hand, Ackerman et.al studied younger female athletes and investigated whether there is a difference in fracture occurrence between oligoamenorrheic athletes (AA), eumenorrheic athletes (EA), and nonathletes (NA) (Ackerman et al. 2015). The study also tried to determine the relationship between bone density, structure, and strength estimates. The last study, done by Thein-Nissenbaum et.al., investigates the prevalence of and the relationship between menstrual irregularity and musculoskeletal injuries in female high school athletes (Thein-Nissenbaum et al. 2012).

4.2 Study designs

One factor all the studies have in common is that they use an observational study design. Observational studies have two primary purposes: descriptive where one examines the dis-

tributions of predictors and outcomes in a population, and analytic where one characterizes associations between these predictor and outcome variables (Hulley et al. 2013).

4.2.1 Cross-sectional study design

Most of the studies are cross-sectional studies as the investigators makes all the measurements on a single occasion or within a short period of time. Cross-sectional studies provide info about prevalence and can be used for examining associations (Hulley et al. 2013). For instance, in the study done by Ackerman et.al, data from different groups of participants were collected within the same day (Ackerman et al. 2015). They interviewed participants to document previous fractures, then measured areal bone mass index of the spine, hip and whole body as well as assessing bone structure.

4.2.2 Cohort study design

The study done by Bingzheng et. al is also an observational study, but a cohort study as the measurements take place over a period of time (Bingzheng et al. 2023). They measured sex hormones and analyzed knee kinematics of the participants during four different phases of the menstrual cycle. More specifically is this a prospective cohort study as the participants are followed over time to observe outcomes that occur after the study begins (Hulley et al. 2013). This way of doing a study may be more beneficial than a cross-sectional study as it allows the calculation of incidence.

4.2.3 Case series study design

The study done by Martin et.al where they tried to find the association between menstrual cycle and injury risk in female soccer players, can be classified as a case series study design (Martin et al. 2021). They recorded injuries and menstrual irregularities among English national soccer team players over a period of four years. Here the group of individuals all share a specific characteristic, they are tracked over a defined period and the outcome in the group is documented. Unlike cohort studies, a case series lacks a comparison group, meaning it does not compare outcomes between exposed and non-exposed individuals or those with and without a condition (Kooistra et al. 2009). This can be seen in the study done by Martin et.al as they gathered data on injury types and menstrual irregularities, but did not include a non-injured or non-menstrual-irregularity group for comparison (Martin et al. 2021).

In summary, all studies utilize an observational study design, each suited to their specific research goals. Cross-sectional studies examine predictors and outcomes at a single time point to assess prevalence and associations (Hulley et al. 2013). Cohort study uses a prospective approach to observe outcomes over time, allowing for the calculation of incidence and temporal

relationships (Hulley et al. 2013). Lastly, case series tracks outcomes within a defined group sharing specific characteristics but lacks a comparison group, limiting its capacity to establish causal relationships (Kooistra et al. 2009).

4.3 Selection of statistical test

Selecting the appropriate statistical test is crucial for ensuring valid and reliable results. The choice depends on several factors, including the type of data you have, your research question, and the assumptions of the tests. None of the studies provided a detailed description of all the statistical tests used, and the studies varied in how many and what type of statistical tests they used to analyze the data.

4.3.1 One-way ANOVA

A one-way ANOVA compares the means of two or more groups for one dependent variable (Ross and Willson 2017). Thein-Nissenbaum et.al used a one-way ANOVA with Bonferroni post hoc test to compare mean values for continuous variables between different sport types (Thein-Nissenbaum et al. 2012). Another study that also compared different groups is the one done by Bingzheng et.al where they examine differences between menstrual phase, late follicular phase, ovulation phase and mid-luteal phase (Bingzheng et al. 2023). The authors do not explicitly mention which statistical tests they used but we can assume that they likely used an ANOVA test to compare the mean values of knee kinematics parameters between the four groups. The study done by Ackerman et.al also fails to specify exactly which statistical test they used but mentions using p-values to assess statistical significance (Ackerman et al. 2015). They compared three groups (AA, EA and NA) regarding various variables such as fracture prevalence, BMD Z-score, bone structure and strength estimates. Based on this we can assume they likely used ANOVA tests for continuous variables.

4.3.2 Person's chi-square test

The examination of cross-classified category data is common in evaluation and research, with Pearson's chi-square test representing one of the most utilized statistical analyses for answering questions about the association or difference between categorical variables (Franke, Ho, and Christie 2011). Thein-Nissenbaum used chi-square tests to examine associations between categorical variables, such as menstrual irregularity and sport type (Thein-Nissenbaum et al. 2012). Additionally, Bingzheng et.al likely used the same statistical test to examine the relationship between serum estrogen, progesterone concentrations and knee kinematics parameters (Bingzheng et al. 2023). Lastly, Cheng et.al mentions the usage of Pearson's chi-square tests to analyze categorical variables in their study, such as usage of hormonal contraceptives and menstrual irregularity (Cheng et al. 2021).

4.3.3 Two-sided t-test

A two-sided t-test, also called two-tailed t-test, is used to determine whether there is a significant difference between the means of two groups in situations where differences can occur in either direction (Nayak and Hazra 2011). In other words, this test does not presume whether the sample mean will be greater than or less than the other mean, instead, it tests for the possibility of differences in both directions. The only study mentioning the usage of a two-sided t-test is Cheng et.al who used it to analyze continuous variables (Cheng et al. 2021). However, the study does not explicitly specify which continuous variables they analyzed. Based on the information provided in the study, the only continuous variable mentioned is age. Likely, the study used a two-sided t-test to compare the mean ages between groups, such as athletes with and without previous menstrual irregularities.

4.3.4 Odds ratio and confidence intervals

Odds Ratio (OR) is a measure of association used to compare the odds of a particular outcome occurring in one group to the odds of it occurring in another group (Nayak and Hazra 2011). A confidence interval (CI) is a range of values that is likely to contain the true value of a parameter with a specified level of confidence (Spiegelhalter 2019). The lower bound and upper bound defines the range while the confidence level indicates how confident we are that the interval contains the true value (Spiegelhalter 2019). Thein-Nissenbaum et.al calculated odds ratio (OR) and 95% confidence intervals (CI) to compare injury severity between athletes with and without menstrual irregularities (Thein-Nissenbaum et al. 2012). Another study that reported OR and CI where the study done by Cheng et.al as they investigated the relationship between the different variables in their study (Cheng et al. 2021).

4.3.5 Descriptive approach

The study done by Martin et.al used a more descriptive approach and did not perform hypothesis testing (Martin et al. 2021). Instead, it focused on calculating injury incidence per 1,000 person-days for each menstrual cycle phase and compared injury incidence ratios between the phases.

4.4 Inference

The five studies analyzed in this report each contribute valuable insights into the complex relationship between the menstrual cycle and musculoskeletal injuries among female athletes. The findings highlight the multifaceted nature of this association, with varying conclusions depending on the specific study. One study emphasized that hormonal contraceptive use is common among female athletes and may mask underlying menstrual irregularities, urging

more education on this risk (Cheng et al. 2021). Another study noted that sex hormones do not have a protective effect on knee kinematics in female soccer players, suggesting that other factors, such as neuromuscular control, should be explored (Bingzheng et al. 2023). In terms of bone health, a study revealed that while weight-bearing exercise can improve bone mineral density (BMD), it may also increase the risk of stress fractures, especially in athletes with menstrual disorders (Ackerman et al. 2015). The different phases of the menstrual cycle also appeared critical, with a study showing that muscle and tendon injuries might be more likely in the days leading up to ovulation (Martin et al. 2021). Lastly, high prevalence of both menstrual irregularities and musculoskeletal injuries was found among high school athletes, with those experiencing menstrual issues having a higher proportion of severe injuries, suggesting menstrual irregularity as a potential risk factor (Thein-Nissenbaum et al. 2012). Collectively, these studies underscore the need for further research to clarify the mechanisms at play and to develop better prevention strategies.

4.5 Strength and weakness

Each study design comes with its own unique strengths and weaknesses. However, all the study designs are observational studies meaning that causal inference is challenging, and interpretation is often muddied by the influences of confounding variables (Hulley et al. 2013). It is therefore important to consider this disadvantage when choosing what statistical test to use when analyzing the data.

4.5.1 Cross-sectional study

Among the studies we reviewed, three utilized a cross-sectional design, focusing on capturing data at a single point in time to assess associations between variables. A cross-sectional study design can be ideal when the researchers don't have the time or money to do a longer study as there is no waiting around for the outcome to occur. This makes them fast and inexpensive and avoids the problem of loss to follow-up (Hulley et al. 2013). On the other hand, a cross-sectional studie measure only prevalence, rather than incidence, it is therefore important to be cautious when drawing inference about the causes of the condition. A factor that is associated with prevalence of a condition may be a cause of the condition but could also be associated with the duration of the condition (Hulley et al. 2013). In other words, it may be difficult to establish causal relationships from cross-sectional data.

4.5.2 Cohort study design

A cohort study, unlike a cross-sectional design, allows for the calculation of incidence. Here levels of the predictor are measured before the outcome occurs which establishes the time

sequence of the variables and strengthens the process of inferring the causal basis of an association (Hulley et al. 2013). It also prevents the predictor measurements from being influenced by the outcome or knowledge of its occurrence and it allows the investigator to measure variables more completely and accurately than is usually possible retrospectively (Hulley et al. 2013). However, a cohort study is more time consuming and may therefore be expensive.

4.5.3 Case series study design

A case series study design is relatively easy to conduct and do not require the complexity or cost associated with larger studies, making it simple and inexpensive. The design allows for the collection of meaningful data from a small group of cases, and researchers can generate hypotheses that can later be tested in more controlled studies (Kooistra et al. 2009). However, it does not include a comparison group, making it difficult to establish causal relationships or to compare outcomes against a standard or control. There is also no control over external variables making it impossible to definitively conclude that a particular exposure caused the outcome (Kooistra et al. 2009).

4.5.4 General

In general, most of the studies have a large sample size which increases the statistical power and gives a better representation of the population in general, meaning that there is a high generalization of the findings (Columb and Atkinson 2016). They also had an extensive data collection gathering information on a wide range of variables, such as demographics, sports participation, and hormonal contraceptives, providing a detailed picture of the health status of this population. On the other hand, most of the studies relied on self-reported data about menstrual and injury history, which may lead to inaccurate recall. They also used a subjective approach to measure some of the variables resulting in less accurate and reliable data. Such lack of clinical data makes it hard to verify the results.

4.6 Future recommendations

To better understand the association between menstrual cycle and musculoskeletal injury among female athletes, future studies should use a prospective study design where a group of individuals are observed over time. The researcher would observe the athlete before the injury occurs to better examine how specific exposures or factors, in this case the menstrual cycle, influence the outcome. Recall bias was one of the weaknesses of the studies we analyzed, a prospective study minimizes recall bias and allows for confounding factors making it highly reliable for establishing causal relationships (Hammerton and Munafò 2021). Another thing that can be done differently is how the menstrual cycle is measured. A more objective approach would be more reliable as hormonal measurements, or the use of menstrual tracking

apps can provide more precise data. Lastly, one should control for other relevant factors such as training volume, diet, stress levels and the use of birth control as they all can affect the menstrual cycle and frequency of injuries.

4.7 Conclusion

In conclusion, our analysis of the five studies highlights both the complexity of the relationship between the menstrual cycle and injuries, as well as the varying approaches taken to study it. While observational study designs were commonly used, each study presented unique strengths and limitations. Cross-sectional studies provided valuable insight into prevalence but struggled with establishing causal relationships, while prospective cohort studies allowed for a clearer understanding of incidence and causality but required more time and resources. Statistical analysis methods varied across studies, with some providing detailed descriptions while others left assumptions to be made about the tests used. Across the board, reliance on self-reported data introduced the risk of recall bias, and the lack of clinical data diminished the reliability of some findings. Future studies would therefore benefit from prospective designs, objective measures of the menstrual cycle, and better control for confounding factors to further clarify the impact of the menstrual cycle on injury risk.

5 Effects of resistance training volume on lean body mass and maximal strength

5.1 Introduction

Resistance training is a general term referring to exercise requiring one to exert force against a resistance (Kraemer et al. 2017). By using resistance, whether from weights or body weight, resistance training helps improve muscle strength, endurance, and power. This way of training have become a fundamental element in fitness and rehabilitation due to its variety of health benefits, as it enhances physical performance and improves overall quality of life. Such benefits are valuable across many fields, from sports to general health and longevity. Investigating the effects of resistance training is therefore important for understanding how to best utilize its health benefits. Different variables in resistance training, such as volume, may affect the outcome, making it essential to examine how specific training protocols contribute to muscular strength and body composition.

Resistance training volume refers to the total amount of work performed in a single resistance training session or across multiple sessions (Ostrowski et al. 1997). Volume is a crucial variable in resistance training as it influences the training adaptions, such as muscle growth, lean body mass and muscle strength. Muscle growth, or hypertrophy, involves an increase in the number of muscle fibers and thus increased muscular size while lean body mass refers to the total body mass without the weight of body fat (Kraemer et al. 2017). Muscle hypertrophy is one component of lean mass gain as gains in lean mass are often attributed to hypertrophy. Muscle strength refers to the maximum force or tension that a muscle or group of muscles can generate during a single maximal effort and is influenced by factors such as hypertrophy (KRAEMER and RATAMESS 2004). Finding the right volume balance is essential as different volumes may lead to different results. Some authors have claimed that multiple sets are necessary to optimize strength gains, while others have argued that a single set per exercise is all that is necessary, and further gains are not achieved by successive sets (Krieger 2009). Ostrowski et al. investigated the impact of different volumes of resistance training on muscle size and function over a 10-week period and found that regardless of the volume of training, all groups experienced increases in muscle size and strength (Ostrowski et al. 1997). These findings are supported by Cannon & Marino as their findings demonstrate that additional neuromuscular adaption during early-phase moderate-intensity resistance training may not be elicited through higher-volume training when training loads are matched provided that a minimal volume threshold is attained (Cannon and Marino 2010). However, evidence from several other studies

Table 5.1: Participant characteristics

	Age (years)	Stature (cm)	Body mass (kg)
Female n = 18	22 (1.3)	167.7 (6.9)	64.4 (10.4)
Male n = 16	23.6 (4.1)	182.9 (5.9)	75.8 (10.7)

Number of participants are indicated by n. Data are means and standard deviations (SD).

supports greater strength gains with multiple sets. One of the most notable scientific studies was performed in the early 1960s and found that different training volumes and intensities elicit different magnitudes of strength gains as three sets of six repetitions resulted in the greatest strength increases (RHEA et al. 2003) & (Berger 1962)). A more recent study, done by Humburg et al., support these findings as they argue that improvements for 1RM is significantly higher during a 3-set program compared to a 1-set program (Humburg et al. 2007).

As a result, there has been considerable debate over the optimal number of sets per exercise to improve musculoskeletal strength during a resistance exercise program (Krieger 2009). Additionally, much research has been performed examining strength increases with training, but they provide little insight into the magnitude of strength gains along the continuum of training volume (RHEA et al. 2003). Hence, the purpose of this study is to identify the effect of single- and multiple-set programs on lean body mass and maximal strength by using a within-participant study design.

5.2 Methods

5.2.1 Participants

The study included 34 participants between the ages of 18 and 40. To be eligible for the study one had to meat specific health criteria and be a non-smoker. People were excluded if they had any sensitivity to local anesthetics, had participated in resistance training more than once a week in the previous year, had limitations in muscle strength due to injury, or were taking medications that could interfere with training outcomes (Hammarström et al. 2020). Participants are further described in Table 5.1. Overall, the male participants are older, taller and heavier than the female participants.

5.2.2 Study overview

The 12-week training intervention consisted of full-body resistance exercises. Leg exercises were done one leg at a time, allowing for the comparison of different training volumes and its effect on each participant. One leg was randomly assigned to perform a single set of each exercise, while the other leg performed three sets. This allowed each participant to engage in

both workout protocols. Muscle strength was evaluated at the beginning of the intervention (pre), and at the end of the 12 weeks (post). Body composition, including measurements of hypertrophy and lean mass gain, were taken at the start and conclusion of the study.

5.2.2.1 Training protocol

Participants completed a standardized warm-up before each workout. The warm-up included cycling on an ergometer for 5 minutes with a rate of perceived exertion (RPE) score of 12-14 as the intensity should be moderate. Secondly, they did bodyweight exercises such as back extensions, push-ups and squats for 10 repetitions. Lastly, participants performed one set of 10 reps at 50% of their one-repetition maximum (1RM) for each exercise. Leg resistance exercises were performed as either one set (single) or three sets (multiple), with the single set completed between the second and third set in the multiple-set routine. Exercises were performed in order with unilateral leg press first, then leg curl and lastly knee extension. Rest intervals between the sets ranged from 90 to 180 seconds. The training intensity gradually increased over the course of the intervention. The participants started with 10RM for the first two weeks, then 8RM for the following three weeks, and lastly 7RM for the remaining seven weeks. One session a week with reduced loads (90%) were introduced from the ninth session. High-intensity training sessions were spaced at least 48 hours apart, while sessions with reduced loads were scheduled at least 24 hours apart from other sessions.

5.2.2.2 Maximal strength and lean mass assessments

Maximal strength for unilateral leg press was evaluated using the one-repetition maximum (1RM) method. 1RM was found by progressively increasing the weight until the participant could no longer complete a full range of motion. The successful lift with the highest load was recorded as the 1RM. Lean mass (and hypertrophy) was measured regionally and determined before and after the training intervention using a dual-energy X-ray absorptiometry (DXA) scanner and magnetic resonance imaging (MRI). MRI specifically determined the knee-extensor muscle cross-sectional area (CSA; vastus lateralis, medialis, intermedius and rectus femoris).

5.2.3 Data analysis

In this study, we used a paired t-test to compare the pre- and post-intervention measurements between single and multiple sets for each participant. The data was analyzed in RStudio.

Table 5.2: Mean lean mass change by sex

Sex	Mean single set change	Mean multiple set change	Mean difference (multiple - single)
Female	117.94	225.89	107.94
Male	220.62	360.12	139.50

Data is in grams (g) and shows the average lean body mass changes from pre- to post-intervention in male and females. The average change for the single set condition where lower compared to the multiple set condition for both sexes, indicating that multiple sets results in a greater lean mass gain. The last column (Mean difference) indicates that the males had a higher increase of lean mass between the different sets compared to the female participants

5.3 Results

5.3.1 Higher training volume results in greater regional hypertrophy

The mean difference in grams (g) for regional lean body mass change between single and multiple sets was 122.8 (95% CI: [8.6, 237.0], p-value = 0.036, t33 = 2.19).

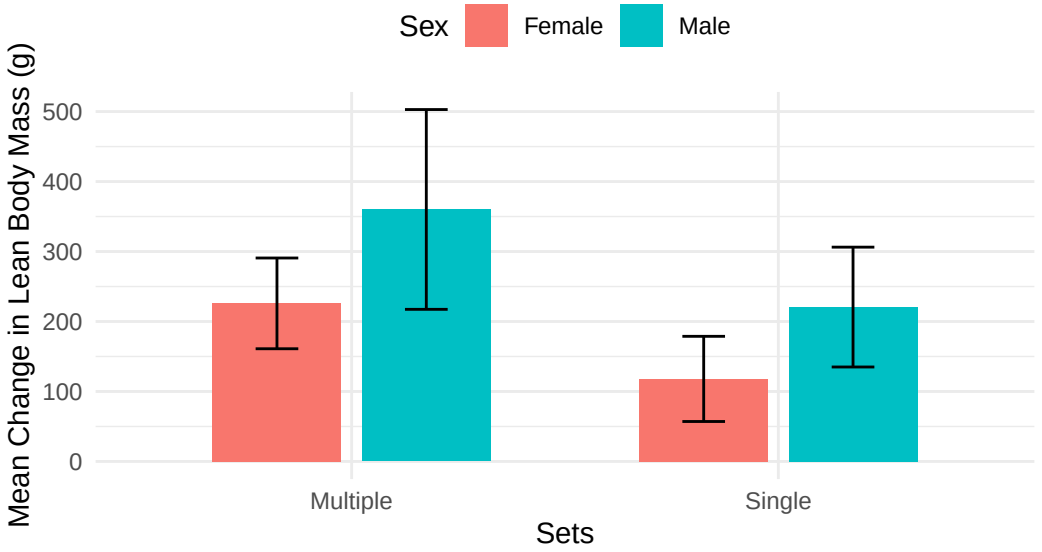


Figure 5.1: Mean lean body mass changes from pre- to post-intervention in male and female participants

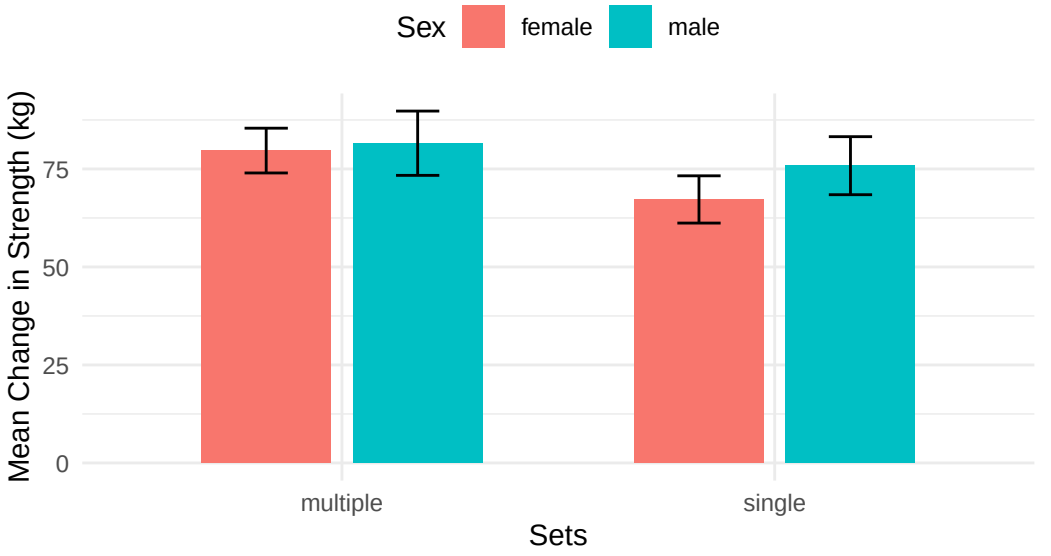
5.3.2 Higher training volume results in greater muscular strength

The mean difference in kilograms (kg) for maximal strength change between single and multiple sets was 6.8 (95% CI: [0.7, 12.9], p-value = 0.031, t30 = 2.26).

Table 5.3: Mean muscle strength change by sex

Sex	Mean single set change	Mean multiple set shange	Mean difference (multiple - single)
Female	67.22	79.69	9.22
Male	75.83	81.56	4.17

Data is in kilograms (kg) and shows the average muscle strength changes from pre- to post-intervention in male and females. The average change for the single set condition where lower compared to the multiple set condition for both sexes, indicating that multiple sets results in a greater strength gain. The last column (Mean difference) indicates that the females had a greater difference in gain of muscle strength between the sets compared to the male participants.



Multiple sets give a higher average increase in maximal strength compared to a single set in both fema The black lines represent the standard error of the mean (SE) for each condition (singe and multiple se and female), indicating the variability of the mean change values.

Figure 5.2: Muscle strength changes from pre- to post-intervention in male and female participants

5.4 Discussion

The paired t-test results for the lean mass and maximal strength data indicate that the difference in change between the single-set and multiple-set conditions was statistically significant, with a mean difference of 122.8 (95% CI: [8.6, 237.0], p-value = 0.036, t33 = 2.19) and 6.8 (95% CI: [0.7, 12.9], p-value = 0.031, t30 = 2.26) respectively. The p-value of 0.036 and 0.031 suggests that this result is unlikely to be due to random chance, as it falls below the significance threshold of 0.05. Furthermore, the 95% confidence interval (CI) demonstrates that the true mean difference in regional lean body mass gain lies between 8.6 grams and 237.0 grams, while the true mean difference in strength lies between 0.7 kilograms and 12.9 kilograms. These findings suggest that the observed effect is consistent and meaningful, but also highlighting some variability across participants.

Overall, the results indicate that both training volumes were effective in improving muscle function, with increased gains in lean body mass and maximal strength. Improvements from pre- to post-intervention where on average higher for the leg performing multiple sets of the exercise compared to the leg who did a single set. Both male and female participants demonstrated similar patterns of improvement, however, males experienced greater gains in lean mass with multiple sets, while females showed more pronounced increases in maximal strength. All in all, these findings suggest that a moderate-volume training regimen is more effective than a low-volume regimen in enhancing lean mass and maximal strength in both men and women.

Our findings align with a growing body of research that indicates higher training volumes generally lead to more significant improvements in both lean mass and muscle strength. Rhea et al. (RHEA et al. 2003), similarly reported that multiple sets resulted in superior strength gains compared to single sets. Additionally, research by Berger and Humburg et al. suggests that training volume plays a significant role in optimizing strength outcomes, particularly with multi-set protocols (Berger 1962);(Humburg et al. 2007). On the other hand, it is worth nothing that regardless of training volume, we saw that on average both sexes exhibited increases in muscle size and strength. These findings align with Ostrowski's notion that resistance training is beneficial across various volumes (Ostrowski et al. 1997). However, our results challenge his conclusion that there are no significant differences between different training volumes, suggesting that a higher volume of training may play a more significant role in optimizing muscle growth and strength development. Moreover, we found that the magnitude of the improvement in lean mass was more pronounced in male participants, whereas females showed more significant strength gains. These findings may align with those of Cannon and Marino who suggested that training volume and intensity have different impacts depending on the individual's training status and the specific adaptions targeted (Cannon and Marino 2010).

5.5 Conclusion

In conclusion, the results demonstrate that performing multiple sets leads to greater improvements in both lean mass gains and muscular strength compared to a single set. These findings are in line with previous findings (RHEA et al. 2003), (Berger 1962) and (Humburg et al. 2007) while challenging the conclusion of studies who suggest that there are no differences in outcomes between training volumes (Ostrowski et al. 1997). In other words, our study highlights the importance of training volumes in optimizing muscular adaptions as our results indicate that increasing the training volume may be more effective for both muscle size and strength gains. However, further research is needed to explore the underlying mechanisms for the differential effects of resistance training volume.

6 Laboratory report

6.1 Introduction

Analysis of gene expression through fluorescence-based real-time quantitative polymerase chain reaction (qPCR) is a well-established practice used in a variety of exercise studies (Kuang et al. 2018). This method quantifies target gene expression in biological samples, such as from blood or muscle tissue. The qPCR analysis is widely used and there are numerous protocols and methods for conducting the analysis. In our study we used the SYBR Green method where a fluorescent dye binds to the DNA during amplification. This fluorescence allows real-time tracking of the reaction, as the signal intensity correlates with the amount of DNA present (Kuang et al. 2018). The fluorescence reaches a pre-determined threshold, referred to as the cycle threshold (CT), which reflects the level of gene expression. A lower CT indicates higher gene expression, as fewer amplification cycles were required to cross the threshold (Livak and Schmittgen 2001).

In exercise physiology, qPCR is frequently used to investigate changes in gene expression that occur with training. In this study we aim to check primer efficiency and perform targeted amplification of cDNA using specific primers. Additionally, we aim to find out how strength training affects the gene expression of different muscle fiber types as participants underwent a two-week strength training program. The different muscle fiber types we looked at were type I, type IIa and type IIx. Type I fibers, slow-twitch, are highly oxidative and can sustain contraction over long periods without fatigue (Pette and Staron 2000). Type IIa fibers utilize both aerobic and anaerobic metabolic pathways, have intermediate resistance to fatigue and produce more force than type I fibers (Pette and Staron 2000). Lastly, type IIx fibers rely predominantly on anaerobic metabolism and are specialized for rapid and powerful contractions (Pette and Staron 2000).

6.2 Materials

- A real-time PCR machine (We use QuantStudio 5)
- A qPCR reaction plate
- Nuclease-free water and pipette tips
- SYBR-green Master mix

Table 6.1: Dilution series

1	2a	3a	4a	2b	3b	4b
1	1/10	1/100	1/1000	1/2	1/20	1/200
30µl	2µl	2µl	2µl	10µl	2µl	2µl
0µl	18µl	18µl	18µl	10µl	18µl	18µl

Table 6.2: Pipetting scheme

row	13	14	15	16	17	18	19	20	21	22	23
	Fp1	Fp2									
A	myhc 1	myhc 1			cmyc 1	cmyc 2a	cmyc 3a	cmyc 4a	cmyc 2b	cmyc 3b	cmyc 4b
B	myhc 1	myhc 1			cmyc 1	cmyc 2a	cmyc 3a	cmyc 4a	cmyc 2b	cmyc 3b	cmyc 4b
C	myhc 1	myhc 1			cmyc 1	cmyc 2a	cmyc 3a	cmyc 4a	cmyc 2b	cmyc 3b	cmyc 4b
D	myhc 2a	myhc 2a									
E	myhc 2a	myhc 2a									
F	myhc 2a	myhc 2a									
G	myhc 2x	myhc 2x									
H	myhc 2x	myhc 2x									
I	myhc 2x	myhc 2x									
J	myhc	myhc			cmyc 1	cmyc 2a	cmyc 3a	cmyc 4a	cmyc 2b	cmyc 3b	cmyc 4b
K	myhc	myhc			cmyc 1	cmyc 2a	cmyc 3a	cmyc 4a	cmyc 2b	cmyc 3b	cmyc 4b
L	myhc	myhc			cmyc 1	cmyc 2a	cmyc 3a	cmyc 4a	cmyc 2b	cmyc 3b	cmyc 4b

6.3 Method

Prior to the experiment, the lab manager prepared the cDNA, which was extracted from samples collected during a study where participants underwent a two-week strength training program.

First, we created a dilution series to test the primers (Table 6.1). We moved 2µl of the sample from tube 1 to tube 2a, and 10µl from 1 to 2b, then vortexed tube 2a and 2b so that the sample and water (H_2O) would mix. Next, we moved 2µl from 2a to 3a and 2µl from 2b to 3b, then vortexed tube 3a and 3b. Lastly, we moved 2µl from 3a to 4a and 2µ from 3b to 4b, then vortexed tube 4a and 4b.

We then combined a master mix consisting of 250µl SYBR-green, 50µl primer mix (MHC1, MHC2a, MHC2x, or MHCb2m), and 100µl H_2O. Subsequently, we loaded the plate with primer-specific master-mix following the outline of our pipetting scheme (Table 6.2). 8µl of the master mix was added to the wells along with 2µl of cDNA sample.

The plate was then covered with plastic and centrifuged at 1200rpm for 1 minute. The PCR samples were analyzed using real-time PCR and QuantStudio software. The PCR process consisted of 40 cycles. After the PCR process was completed, we extracted the results in the form of CT values.

6.3.1 Data Analysis

We analyzed our data in excel.

Table 6.3: Dilution series results

Dilution	Ct1	Ct2	Ct3	Avg.Ct	Sample.quan	Log...sample.quan.	Slope	Primer.Efficiency....
1	28.678	28.708	29.155	28.847	1.000	0.000	-2.6104	141.5901
1/2	29.414	29.62	29.264	29.433	0.500	-0.301		
1/10	31.776	31.416	32.413	31.868	0.100	-1.000		
1/20	33.241	32.653	Undetermined	32.947	0.050	-1.301		
1/200	Undetermined	Undetermined	34.574	34.574	0.005	-2.301		

A slope of -2.6 indicates that the primer efficiency is not optimal as it is calculated to be 142%. We can see that our observation is not linear

6.3.1.1 Dilution Series

Here we calculated the average CT value for the triplicates as well as the base-10 logarithm of the sample quantity (Log(sample.quan). The slope was then calculated by plotting Log(sample.quan) against the average CT. A linear regression was performed to find the best-fit line, and the slope was derived. Primer efficiency was then calculated from the slope. Lastly, we calculated the standard deviation (SD) among replicates and the coefficient of variation (CV).

6.3.1.2 Gene Expression

We processed our myosin heavy chain gene expression data by using the Delta-Delta-Ct ($\Delta\Delta$Ct) method. Cycle threshold (Ct) values for both the target gene and the reference gene (b2m) were measured. We calculated the average for multiple technical replicates to reduce variability. We then calculated ΔCt which quantifies the expression of the target gene relative to the reference gene within the sample. Lastly, we calculated the relative expression level (2^-$\Delta\Delta$Ct) representing the fold change in expression of the target gene in the sample relative to the calibrator.

6.4 Results

6.4.1 Dilution Series

6.4.2 Gene Expression

6.5 Discussion

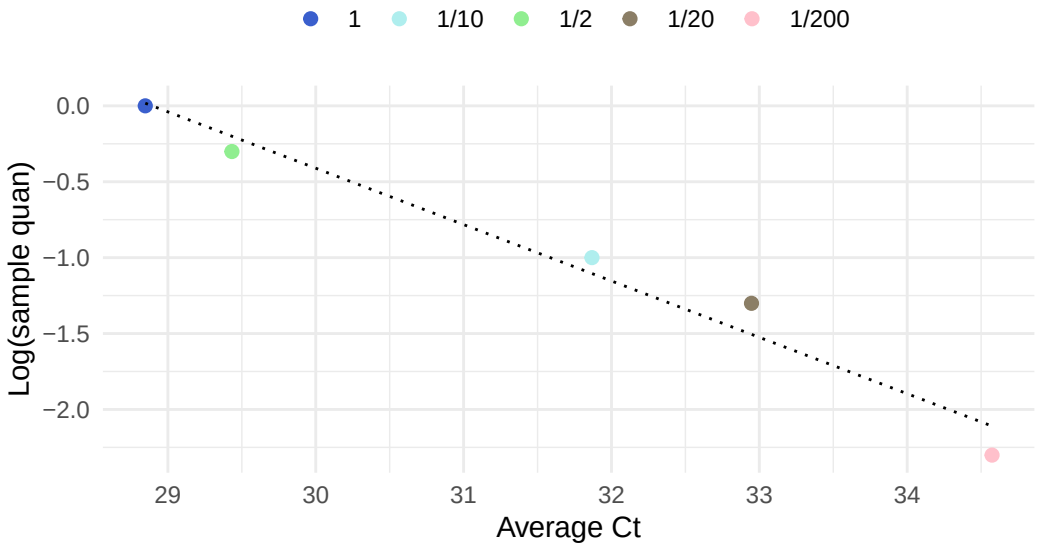
6.5.1 Dilution Series

From our dilution series we saw that as the dilution increases, Ct values rises, indicating that the cDNA concentration decreases which is to be expected in a properly diluted sample series

Table 6.4: Average Ct values, standard deviations (SD), and coefficients of variation (CV) for pooled qPCR samples at three dilution levels

Pooled.sam	Avg	SD	CV
1.0	28.847	0.2670490	0.93 %
0.5	29.433	0.1788764	0.61 %
0.1	31.868	0.5050010	1.58 %

The standard deviations (SD) are relatively low for the dilutions in the pooled sample, indicating precision in the results. The coefficient of variation (CV) is lowest for the 1/2 diluted sample at 0.61% and highest for the 1/10 diluted sample at 1.58%. This suggests that there is greater variability between the measurements for the most diluted samples. Additionally, the CV of the undiluted sample is higher than that of the 1/2 diluted sample.



The plot shows a negative linear relationship between average Ct values and log (sample quantity). T black line represents a linear regression fit to the data, emphasizing the inverse proportionality. Different colors represent the dilution level.

Figure 6.1: Relationship between the average cycle threshold values and the logarithm of sample quantities in our dilution series

Table 6.5: Cycle threshold values

Sample	Target	Ct1	Ct2	Ct3	Avg.	Ref.	Ref_Ct1	Ref_Ct2	Ref_Ct3	Ref_Avg	ΔCt	TwoΔΔC
FP6 R w0	myhc1	19.798	19.901	19.621	19.77333	b2m	24.670	24.513	24.691	24.625	-4.851	28.86
FP6 R w2pre	myhc1	18.944	19.240	19.861	19.34833	b2m	22.913	23.950	23.819	23.561	-4.212	18.53
FP6 R w0	mhc2a	21.029	21.247	20.627	20.96800	b2m	24.670	24.513	24.691	24.625	-3.657	12.61
FP6 R w2pre	mhc2a	19.549	19.304	19.580	19.47800	b2m	22.913	23.950	23.819	23.561	-4.083	16.94
FP6 R w0	myhc2x	27.019	26.898	25.907	26.60795	b2m	24.670	24.513	24.691	24.625	1.983	0.25
FP6 R w2pre	myhc2x	24.871	24.105	24.256	24.41062	b2m	22.913	23.950	23.819	23.561	0.850	0.55

Number of cycles to reach the cycle threshold (CT) has changed from week 0 to week 2. The number of cycles decreased for myhc1 from an average of 19.7 to 19.3 cycles. For myhc2a, the cycles dropped from 20.9 to 19.4, and for myhc2x it went from 26.6 to 24.4 cycles. This indicates that the gene expression increased from week 0 to week 2 as fewer cycles indicate higher gene expression.

Table 6.6: Percentage amount of gene expression for the different muscle types

100%	myhc1	myhc2a	myhc2x
41.733	69.17 %	30.22 %	0.61 %
36.034	51.44 %	47.02 %	1.54 %

The amount of gene expression for the different muscle fiber types has changed from week 0 to week 2. The expression of Muscle Fiber Type 1 (myhc1) and Muscle Fiber Type 2a (myhc2a) decreased, while Muscle Fiber Type 2x (myhc2x) increased.

(Svec et al. 2015). However, Ct values are slightly inconsistent for higher dilutions, indicated with "Undetermined" in Table 6.3. This suggests potential variability in detecting low cDNA amounts. Furthermore, we calculated the slope of the standard curve by plotting the logarithm of template concentration against the Ct values and got a slope of -2.6104. The theoretical ideal slope is -3.33 which corresponds to 100% primer efficiency (Svec et al. 2015). This further complicated our calculated primer efficiency as 141.59% is far above the acceptable range of 90-110% (Svec et al. 2015). Our findings indicate issues within our experiment, suggesting that the primer did not efficiently amplify the target during each PCR cycle as intended. This may have skewed the relative quantification of gene expression and lead to false interpretation of our results.

The results shown in Table 6.4 highlight important aspects of reproducibility and variability at different sample dilutions. As expected, the average Ct values increase with dilution since lower concentrations require more cycles to reach the detection threshold (Livak and Schmittgen 2001). The standard deviation (SD) values are increasing as the sample is diluted. This indicates that as the template concentration decreases, the variability in the amplification increases. Lastly, the coefficient of variation (CV) is relatively low for the undiluted sample and for the ½ dilution, reflecting minimal variability and high precision in these concentrations. However, the CV value rises for the 1/10 dilution, suggesting increased variability in the measurements as the sample becomes more diluted. The increased SD and CV values at higher dilutions could be indicative of challenges such as pipetting errors or reduced efficiency in template detection.

6.5.2 Gene Expression

We compared the myosin heavy chain gene expression level for pre- (w0) and post-intervention (w2pre) and found an increased expression of myhc 1 and 2a. Both genes show higher expression at 2 weeks (w2pre) compared to timepoint 0 (w0). The relative expression of these genes increases as indicated by the ΔCt and $2^{\wedge}\Delta\Delta$Ct values (Table 6.5). Although myhc 2x also shows some increase in expression after 2 weeks, the magnitude of the increase is smaller compared to myhc 1 and 2a. This suggests a more gradual or less pronounced activation of myhc 2x in response to the training intervention. Overall, our findings suggest that strength training induces changes in gene expression, with myhc 1 and 2a showing more pronounced changes compared to myhc 2x.

There are limited studies who explain the changes in myosin heavy chain (myhc) gene expression after two weeks of strength training. Wilborn and Willoughby stated that after 8 weeks of heavy strength training, gene expression for myhc 1 and 2a increased, while that for mych 2x decreased (Wilborn and Willoughby 2004). We believe that the observed changes would correspond with those of Wilborn and Willoughby, but to a lesser extent. Their findings align with our findings, as we observed an increase in gene expression for both myhc 1 and myhc 2a, but it contradicts our observation of an increase in myhc 2x. The increase in myhc2x, associated with the fastest and most explosive muscle fibers, is unexpected, as this fiber type is typically reduced with prolonged training in favor of more endurance- or strength-oriented fibers (Type I and IIa) (Pette and Staron 2000). This might indicate that our results represent an initial upregulation phase where all isoforms (myhc 1, 2a, 2x) are expressed to adapt to increased training stress. A longer training duration, such as that of Wilborn and Willoughby's 8-week period, might show a clearer downregulation of myhc 2x as type 1 and 2a fibers are prioritized for strength and endurance adoptions (Pette and Staron 2000).However, it is also possible that this reflects technical errors, such as high primer efficiency or biological variation.

6.5.3 Deviation

Our findings suggest that technical issues or inherent flaws in our primer design may have led to inaccurate results or misinterpretation of the data. A possible cause of deviation may be cross-contamination in reagents or samples which may lead to higher-than-expected amplifications. Our primer design might have been poor causing the primer to bind to unintended sequences or to themselves. Lastly, poor pipetting might have led to incorrect volumes during dilution preparation resulting in inaccurate standard curves. To address these issues, we could repeat the experiment with a new dilution series where we ensured careful pipetting to avoid errors. Additionally, negative controls could be used to confirm the absence of contamination in reagents or samples.

6.6 Conclusion

In conclusion, the results from our qPCR experiment provided valuable insights into the changes in gene expression related to muscle fiber types after two weeks of strength training. While our findings suggest an increase in gene expression for myhc 1, 2a, and 2x, we observed potential technical challenges that may have affected the accuracy of our results. Despite these limitations, our results align with previous studies on myosin heavy chain gene expression, indicating a likely up-regulation of myhc 1 and 2a with less pronounced changes for myhc 2x. Although a longer duration of training may provide further clarification on the dynamics of myosin heavy chain gene expression in response to strength training. Future experiments should address technical concerns, including pipetting errors and contamination, to get more reliable results.

7 Philosophy of science

7.1 Provide a brief description of falsificationism and explain why Popper was motivated to develop this theory. Present one problem with the theory and assess whether the problem can be solved.

Falsificationism is a philosophy of science developed by Karl Popper. It is the idea that scientific theories may be falsified but never proven, meaning that the theory can be tested and likely disproven based on observations or experience (Burke 1986). Popper meant that many scientific theories, such as psychoanalysis, simply explained every possible event and would therefore consider them unfalsifiable. He categorized such theories as pseudoscience referring to theories that claim to be scientific but lack the evidence or testability to be classified as genuine science (Thornton 2023). This view is what motivated Popper to develop his philosophy as falsification aimed to differentiate between pseudoscience and scientific theories.

Popper believed that a scientist should try to prove his own theory wrong, instead of looking for confirmation. He argued that induction is not possible and believed that it could not be rationally justified, hence, a confirmation would not be attainable as it requires induction. Furthermore, Popper argued that a good falsifiable theory should make precise and often multiple predictions (Burke 1986). These predictions could identify ways in which the theory might be false and therefore reject it or aid in making some sense of further belief that the theory is possibly true. However, this may be challenging as rejecting a theory based on a single experiment provides a very vague way of determining where the fault lies. Something may have gone wrong without the fault of the theory itself as one of the auxiliary hypotheses used in that experiment could be at fault. In other words, if an experiment where to get an unexpected result, it does not necessarily mean that the theory is wrong but could mean that the experimental setup was flawed, or a wrong assumption was made. In conclusion, Popper's way of testing and rejecting a theory makes it hard to know what exactly is being falsified, as testing a theory is not a straightforward process of disconfirmation and rejection.

A solution to the previously mentioned challenge that comes with the clean-cut falsification could be to adopt a more flexible approach. This way Popper's philosophy may still be useful as we implement an approach that focuses more on the complexity of testing in real-world science while keeping the fundamental parts of his falsification philosophy. The new approach would adjust the auxiliary hypotheses while continuing to test the theory, meaning that the

theory would not be rejected after a single experiment. In other words, a scientist would adjust their auxiliary hypothesis, continue to test their theory, and refine it based on new evidence. The original theory of Popper's falsificationism is limited when it comes to acknowledging the complexity of actual scientific practice, but this new approach may diminish the problem.

In conclusion, Popper wanted to differentiate scientific theories from pseudoscience and argued that scientific theories cannot be confirmed but can only be tested and potentially disproven. This principle of clean falsification can be challenging when determining the source of unexpected results, a more flexible approach would therefore be more ideal to enhance the functionality of falsificationism in real-world scientific practice.

7.2 Explain basic ideas of Bayesianism and how Bayesian probabilities can be interpreted. Present one problem with Bayesianism and evaluate how serious the problem is.

Bayesianism is an attempt to apply mathematical probability theory to understand the scientific process, including induction, falsification, and confirmation. Bayesians argue that all unknown quantities one wishes to draw conclusions about should be treated as random variables, and the uncertainty surrounding these quantities is described using a probability distribution (Sober 2002). Bayesianism is based on Bayes' Theorem, which provides a way to update probabilities based on new evidence. The theorem is expressed as P(H|E) = P(E|H) * P(H) / P(E) (Easwaran 2011). Easwaran further explains that the first variable of the formula, P(H|E), is called posterior probability and is what we are trying to find. It states the probability of the hypothesis (H) given the evidence (E). To find the posterior probability we need to know the likelihood, P(E|H), which is the probability of observing the evidence (E) given that the hypothesis (H) is true. In other words, how accurately the hypothesis can predict the evidence. We also need to know the prior probability, P(H), stating the probability of the hypothesis before the evidence is even considered. Lastly, P(E) is the marginal likelihood or the total probability of observing the evidence (E) under all hypotheses (Easwaran 2011). The more surprising the evidence (E), giving us a low probability, the higher the effect on the posterior probability of the hypothesis (Vassend 2024). In simple term, Bayes' Theorem allows us to incorporate prior knowledge and new evidence to produce a revised or updated probability.

Bayesian probabilities can be interpreted in several different ways. One way is where the probability is expressed as the researcher's personal belief in the truth of a hypothesis. This interpretation is called personal belief where a researcher may say that there is a 99% probability that a theory is true, based on their individual conviction (Vassend 2024). Secondly, there is collective belief where the probability represents the collective belief of most researchers. Here the percentage of probability would reflect a consensus among the majority of researchers in the field of study. Lastly, some Bayesians argue for a more objective type of probability suggesting that probability represents some universal truth, independent of individual researchers.

A central problem with Bayesianism is the determination of prior probabilities, or the subjectivity in the prior. This problem arises as different researchers may assign different priors based on their background knowledge, experience, or personal biases (Sober 2002). These priors will also lack objective evidentiary power, and researchers are more interested in objective data than in others subjective opinions (Easwaran 2011). As a result, the prior can skew the result, leading to misleading conclusions. Therefore, the problem with prior probabilities is serious as it undermines the objectivity of Bayesian analyses. Different prior probabilities can lead to radically different conclusions, even when using the same data. It is important to be aware of this issue, maintain transparency about the choice of prior probability, and focus on objective data (Sober 2002).

In summary, Bayesianism provides a framework for incorporating uncertainty and prior knowledge when exploring scientific questions using probability. The approach is based on Bayes' Theorem which facilitates the updating of probability based on new evidence. However, a significant issue lies in the determination of prior probabilities as they can introduce subjectivity and bias into analyses, resulting in misleading conclusions.

References

Ackerman, Kathryn E, Natalia Cano Sokoloff, Giovana DE Nardo Maffazioli, Hannah M Clarke, Hang Lee, and Madhusmita Misra. 2015. "Fractures in Relation to Menstrual Status and Bone Parameters in Young Athletes." *Med. Sci. Sports Exerc.* 47 (8): 1577–86.

Astorino, T A, J Willey, J Kinnahan, S M Larsson, H Welch, and L C Dalleck. 2005. "Elucidating Determinants of the Plateau in Oxygen Consumption at $o_2$max." *British Journal of Sports Medicine* 39 (9): 655–60. https://doi.org/10.1136/bjsm.2004.016550.

Berger, Richard. 1962. "Effect of Varied Weight Training Programs on Strength." *Res. Q. Am. Assoc. Health Phys. Educ. Recreat.* 33 (2): 168–81.

Bingzheng, Zhou, Zhao Xinzhuo, Jin Zhuo, Yang Xing, Li Bin, and Bai Lunhao. 2023. "The Effects of Sex Hormones During the Menstrual Cycle on Knee Kinematics." *Front. Bioeng. Biotechnol.* 11 (September): 1209652.

Buchheit, M, Y Papelier, P B Laursen, and S Ahmaidi. 2010. "Reliability of a Maximal Exercise Test for the Determination of VO2max in Well-Trained Male Runners." *Journal of Sports Sciences* 28 (9): 913–19.

Burke, T E. 1986. *Philosophy of Popper*. Manchester, England: Manchester University Press.

Buttar, K K, N Saboo, and S Kacker. 2019. "A Review: Maximal Oxygen Uptake (VO2 Max) and Its Estimation Methods." *International Journal of Physical Education, Sports and Health* 6 (6): 24–32.

Cannon, Jack, and Frank E Marino. 2010. "Early-Phase Neuromuscular Adaptations to High- and Low-Volume Resistance Training in Untrained Young and Older Women." *J. Sports Sci.* 28 (14): 1505–14.

Cheng, Jennifer, Kristen A Santiago, Zafir Abutalib, Kate E Temme, Ann Hulme, Marci A Goolsby, Carrie L Esopenko, and Ellen K Casey. 2021. "Menstrual Irregularity, Hormonal Contraceptive Use, and Bone Stress Injuries in Collegiate Female Athletes in the United States." *PM R* 13 (11): 1207–15.

Columb, MO, and MS Atkinson. 2016. "Statistical Analysis: Sample Size and Power Estimations." *BJA Education* 16 (5): 159–61. https://doi.org/10.1093/bjaed/mkv034.

Easwaran, Kenny. 2011. "Bayesianism II: Applications and Criticisms." *Philosophy Compass* 6 (5): 321–32. https://doi.org/10.1111/j.1747-9991.2011.00398.x.

Faber, Jorge, and Lilian Martins Fonseca. 2014. "How Sample Size Influences Research Outcomes." *Dental Press Journal of Orthodontics* 19 (4): 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo.

Facey, Aldeam, Rachael Irving, and Lowell Dilworth. 2013. "Overview of Lactate Metabolism and the Implications for Athletes." *American Journal of Sports Science and Medicine* 1 (3): 42–46.

Franke, Todd Michael, Timothy Ho, and Christina A. Christie. 2011. "The Chi-Square Test." *American Journal of Evaluation* 33 (3): 448–58. https://doi.org/10.1177/1098214011426594.

Halperin, Israel, David B. Pyne, and David T. Martin. 2015. "Threats to Internal Validity in Exercise Science: A Review of Overlooked Confounding Variables." *International Journal of Sports Physiology and Performance* 10 (7): 823–29. https://doi.org/10.1123/ijspp.2014-0566.

Hammarstrom, D. 2024. "12 Linear and Curve-Linear Relationships, and Predictions" 12.

Hammarström, Daniel, Sjur Øfsteng, Lise Koll, Marita Hanestadhaugen, Ivana Hollan, William Apró, Jon Elling Whist, Eva Blomstrand, Bent R. Rønnestad, and Stian Ellefsen. 2020. "Benefits of Higher Resistance-Training Volume Are Related to Ribosome Biogenesis." *The Journal of Physiology* 598 (3): 543–65. https://doi.org/10.1113/jp278455.

Hammerton, Gemma, and Marcus R. Munafò. 2021. "Causal Inference with Observational Data: The Need for Triangulation of Evidence." *Psychological Medicine* 51 (4): 563–78. https://doi.org/10.1017/s0033291720005127.

Haun, Cody T, Christopher G Vann, C Brooks Mobley, Shelby C Osburn, Petey W Mumford, Paul A Roberson, Matthew A Romero, et al. 2019. "Pre-Training Skeletal Muscle Fiber Size and Predominant Fiber Type Best Predict Hypertrophic Responses to 6 Weeks of Resistance Training in Previously Trained Young Men." *Front. Physiol.* 10 (March): 297.

Haun, Cody T, Christopher G Vann, Christopher B Mobley, Paul A Roberson, Shelby C Osburn, Hudson M Holmes, Petey M Mumford, et al. 2018. "Effects of Graded Whey Supplementation During Extreme-Volume Resistance Training." *Front. Nutr.* 5 (September): 84.

Hopkins, Will G. 2000. "Measures of Reliability in Sports Medicine and Science." *Sports Medicine* 30 (1): 1–15. https://doi.org/10.2165/00007256-200030010-00001.

Hulley, Stephen B, Steven R Cummings, Warren S Browner, Deborah G Grady, and Thomas B Newman. 2013. *Designing Clinical Research.* 4th ed. Philadelphia, PA: Lippincott Williams; Wilkins.

Humburg, Hartmut, Hartmut Baars, Jan Schröder, Rüdiger Reer, and Klaus-Michael Braumann. 2007. "1-Set Vs. 3-Set Resistance Training: A Crossover Study." *The Journal of Strength and Conditioning Research* 21 (2): 578. https://doi.org/10.1519/r-21596.1.

Jones, Andrew M., and Helen Carter. 2000. "The Effect of Endurance Training on Parameters of Aerobic Fitness." *Sports Medicine* 29 (6): 373–86. https://doi.org/10.2165/00007256-200029060-00001.

Kooistra, Bauke, Bernadette Dijkman, Thomas A. Einhorn, and Mohit Bhandari. 2009. "How to Design a Good Case Series." *Journal of Bone and Joint Surgery* 91 (Supplement_3): 21–26. https://doi.org/10.2106/jbjs.h.01573.

KRAEMER, WILLIAM J., and NICHOLAS A. RATAMESS. 2004. "Fundamentals of Resistance Training: Progression and Exercise Prescription." *Medicine & Science in Sports & Exercise* 36 (4): 674–88. https://doi.org/10.1249/01.mss.0000121945.36635.61.

Kraemer, William J., Nicholas A. Ratamess, Shawn D. Flanagan, Jason P. Shurley, Janice S. Todd, and Terry C. Todd. 2017. "Understanding the Science of Resistance Training: An Evolutionary Perspective." *Sports Medicine* 47 (12): 2415–35. https://doi.org/10.1007/

s40279-017-0779-y.

Krieger, James W. 2009. "Single Versus Multiple Sets of Resistance Exercise: A Meta-Regression." *Journal of Strength and Conditioning Research* 23 (6): 1890–1901. https://doi.org/10.1519/jsc.0b013e3181b370be.

Kuang, Jujiao, Xu Yan, Amanda J. Genders, Cesare Granata, and David J. Bishop. 2018. "An Overview of Technical Considerations When Using Quantitative Real-Time PCR Analysis of Gene Expression in Human Exercise Research." Edited by Ruslan Kalendar. *PLOS ONE* 13 (5): e0196438. https://doi.org/10.1371/journal.pone.0196438.

Livak, K J, and T D Schmittgen. 2001. "Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2(-Delta Delta C(T)) Method." *Methods* 25 (4): 402–8.

"Login - eLabFTW — Elab.inn.no." https://elab.inn.no/experiments.php?mode=view&id=83.

Marchetti, P H, A J De Almeida, and R Alvares. 2016. "Variability in Metabolic Responses to Exercise and Implications for Testing." *Sports Medicine* 46 (6): 887–95.

Martin, Dan, Kate Timmins, Charlotte Cowie, Jon Alty, Ritan Mehta, Alicia Tang, and Ian Varley. 2021. "Injury Incidence Across the Menstrual Cycle in International Footballers." *Front. Sports Act. Living* 3 (March): 616999.

Miller, B F, and M Mchugh. 2014. "Factors Affecting the Variability of Physiological Measures." *Sports Medicine* 44 (8): 1–10.

Nayak, BarunK, and Avijit Hazra. 2011. "How to Choose the Right Statistical Test?" *Indian Journal of Ophthalmology* 59 (2): 85. https://doi.org/10.4103/0301-4738.77005.

Newell, John, David Higgins, Niall Madden, James Cruickshank, Jochen Einbeck, Kenny McMillan, and Roddy McDonald. 2007. "Software for Calculating Blood Lactate Endurance Markers." *Journal of Sports Sciences* 25 (12): 1403–9. https://doi.org/10.1080/02640410601128922.

Ostrowski, Karl J, Greg J Wilson, Robert Weatherby, Peter W Murphy, and Andrew D Lyttle. 1997. "The Effect of Weight Training Volume on Hormonal Output and Muscular Size and Function." *The Journal of Strength & Conditioning Research* 11 (3): 148–54.

Pette, D, and R S Staron. 2000. "Myosin Isoforms, Muscle Fiber Types, and Transitions." *Microsc. Res. Tech.* 50 (6): 500–509.

RHEA, MATTHEW R., BRENT A. ALVAR, LEE N. BURKETT, and STEPHEN D. BALL. 2003. "A Meta-Analysis to Determine the Dose Response for Strength Development." *Medicine & Science in Sports & Exercise* 35 (3): 456–64. https://doi.org/10.1249/01.mss.0000053727.63505.d4.

Ross, Amanda, and Victor L. Willson. 2017. *Basic and Advanced Statistical Tests*. SensePublishers. https://doi.org/10.1007/978-94-6351-086-8.

Routledge, F S. 2015. "Calibration and Quality Control in Exercise Testing." *Journal of Sports Science and Medicine*.

Schindelin, Johannes, Ignacio Arganda-Carreras, Erwin Frise, Verena Kaynig, Mark Longair, Tobias Pietzsch, Stephan Preibisch, et al. 2012. "Fiji: An Open-Source Platform for Biological-Image Analysis." *Nat. Methods* 9 (7): 676–82.

Sober, E. 2002. "Bayesianism - Its Scope and Limits." In *Proceedings-British Academy*, 113:21–

38. OXFORD UNIVERSITY PRESS INC.

Spiegelhalter, David. 2019. *The Art of Statistics*. La Vergne, TN: Basic Books.

Svec, David, Ales Tichopad, Vendula Novosadova, Michael W. Pfaffl, and Mikael Kubista. 2015. "How Good Is a PCR Efficiency Estimate: Recommendations for Precise and Robust qPCR Efficiency Assessments." *Biomolecular Detection and Quantification* 3 (March): 9–16. https://doi.org/10.1016/j.bdq.2015.01.005.

Sylta, Øystein, Espen Tønnessen, Daniel Hammarström, Jørgen Danielsen, Knut Skovereng, Troels Ravn, Bent R Rønnestad, Øyvind Sandbakk, and Stephen Seiler. 2016. "The Effect of Different High-Intensity Periodization Models on Endurance Adaptations." *Med. Sci. Sports Exerc.* 48 (11): 2165–74.

Thein-Nissenbaum, Jill M, Mitchell J Rauh, Kathleen E Carr, Keith J Loud, and Timothy A McGuine. 2012. "Menstrual Irregularity and Musculoskeletal Injury in Female High School Athletes." *J. Athl. Train.* 47 (1): 74–82.

Thornton, Stephen. 2023. "Karl Popper." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N Zalta and Uri Nodelman, Winter 2023. Metaphysics Research Lab, Stanford University.

Vassend, O. 2024. "KvantMet: Vitenskapsfilosofi Dag 4. Høgskolen i Innlandet" 4.

Wilborn, Colin D, and Darryn S Willoughby. 2004. "The Role of Dietary Protein Intake and Resistance Training on Myosin Heavy Chain Expression." *Journal of the International Society of Sports Nutrition* 1 (2). https://doi.org/10.1186/1550-2783-1-2-27.