UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Steatosis quantification in Non-Alcoholic Fatty Liver Disease: A statistical approach for comparing different image processing procedures

Andreia Sofia Pedro Mindouro

Mestrado em Bioestatística

Trabalho de Projeto orientado por:
Professora Doutora Lisete Sousa
Doutor Tiago Paixão

2025

***“Our greatest weakness lies in giving up.
The most certain way to succeed is
always to try just one more time.”***

Thomas Edison

For my son Afonso and my daughter Alice,

the light of my life.

Acknowledgments

I would like to start by expressing my immense gratitude to all the incredible people who made the realization of this project possible. To everyone who supported me not only technically but also through encouragement and motivation — my deepest thanks.

To my internal supervisor, Professor Doctor Lisete Sousa, and my external supervisor, Doctor Tiago Paixão: your unwavering belief in this project and your constant encouragement have been instrumental in shaping my professional growth.

A sincere thank you to those who directly supported me in the decision to pursue this Master's degree, who proposed the idea, shared the data that made this project possible, and enabled me to balance work and study. Thank you, Mafalda Casanova, Doctor Pedro Faísca, and my Head and friend Joana Rodrigues, for your trust from day one. To Joana in particular: thank you not only for your professional support, but also for being a true friend throughout this journey. Your empathy, patience, and constant encouragement made all the difference, you are a valuable part of my life.

To José Marques, for creating the macros to run the software and for your efforts in helping me understand the image processing behind the database scope. To my friend and colleague André Barros — thank you for your invaluable support during challenging times, and for our philosophical statistical moments over coffee. I am grateful for your knowledge, time, and friendship.

To my adventurous and brave friend Ana Laura, who inspires me every day and pushes me forward, and whose daily support has meant more than words can say.

To Susana Ladeiro, my charismatic friend who has always been there for me; and to João Sobral, Cathy, and João Costa, thank you for your friendship and support.

To the fantastic Advanced Microscopy team, thank you for helping me navigate the image software — Gabriel, Mariana, Alexandre, and Manuel — and a special thanks to Patricia for her trust and friendship. To Carolina for her kindness and support.

Of course, I am incredibly grateful for the opportunity to work at the *Gulbenkian Institute for Molecular Medicine*. This institution has not only enriched my professional life but also brought extraordinary individuals into my circle. It has become my second home, and I am deeply thankful to be a part of such a unique and inspiring community.

Thanks to the *Faculdade de Ciência*s for the resources, guidance, and the opportunity to meet people who made this path more enjoyable. Thank you for your continuous support and for helping me grow: Érica, Joana Ribeiro, Inês, Isabel, Maria João, Ricardo, and Tomás — your support has been instrumental in my development.

To all the professors who were part of my academic journey, thank you not only for the knowledge you shared but also for sparking my interest in statistics, particularly biostatistics. I believe you have given me valuable tools not only for the development of this thesis but also for my future career. In particular, to Professors Fernanda Diamantino and Helena Mouriño — thank you for being enthusiastic and inspiring teachers.

And because the last shall be first — I want to thank the most encouraging and persistent friend, who, during COVID, wore the most fashionable masks, always cheered us up with food, and made the fastest and most delicious meals. The one who always carries snacks in her bag for afternoon energy boosts, who can talk for over an hour on the phone, and who opens the doors of her home to me. The best woman and friend, Sandra — I want to be like you when I grow up. You are the best gift this Master's has brought me.

I am also grateful for the presence of my dear friends Raquel, Andreia Pontífice, Andreia Dias, Andreia Pedrosa, Joana, and Vanessa — thank you for always being there. You are family by heart.

Lastly, the essential pillar of my life — my family. Without you, none of this would have been possible. You've cared for my children, offered unwavering support during my absence, and brought comfort through the most difficult moments. I am deeply thankful to my mother, Dina; my brother, Cláudio; my father, Augusto; my grandmother, Adelaide; and my sister-in-law, Marina. You are the reason I've come this far.

To Carlos — my love, best friend, and the father of my children — thank you for your unwavering support throughout these 16 years together. And to my children, Afonso and Alice — you are my unconditional love and my strength.

Although my beloved grandfather José passed away in 2021, he will always be our guiding star.

Thank you all for being a ray of sunshine, even on my darkest days.

Andreia Mindouro

Gulbenkian Institute *for* Molecular Medicine

Steatosis quantification in Non-Alcoholic Fatty Liver Disease:

A statistical approach for comparing different image processing procedures

Andreia Sofia Pedro Mindouro

2025

This thesis was fully performed at *Gulbenkian Institute for Molecular Medicine*, under the direct supervision of Dr Tiago Paixão in the scope of the Master Programme in Biostatistics of the *Faculdade de Ciências da Universidade de Lisboa*.

Resumo

O fígado é um órgão essencial nos vertebrados, desempenhando funções vitais no metabolismo e no sistema imunitário. Entre as suas múltiplas funções metabólicas, destaca-se o metabolismo lipídico. Alterações ou desequilíbrios neste processo metabólico podem levar a uma perturbação da homeostase do fígado através da acumulação irregular de gordura neste órgão, condição conhecida como esteatose hepática.

A esteatose hepática pode ser causada por diversos fatores, como obesidade, resistência à insulina, diabetes tipo 2, síndrome metabólica, estilo de vida sedentário, dieta inadequada, níveis elevados de colesterol e triglicerídeos, deficiências nutricionais, uso de medicamentos, hepatite, desnutrição e consumo crónico de álcool. Divide-se em dois grandes tipos: esteatose hepática alcoólica e não alcoólica. O estudo apresentado neste projeto concentra-se em casos de doença hepática crónica relacionada com a Esteatose hepática não alcoólica (EHNA).

A EHNA tem uma prevalência global de cerca de 25% e está relacionada com distúrbios metabólicos, na sua maioria causados por obesidade e resistência à insulina. Adicionalmente, a esteatose hepática encontra-se frequentemente associada a indivíduos com uma dieta rica em gorduras, os quais apresentam uma maior prevalência da condição. A doença é definida pela presença de uma acumulação anormal de lípidos (esteatose) no interior do citoplasma dos hepatócitos, sob a forma de vacúolos lipídicos, presente em mais de 5% das células hepáticas (hepatócitos), na ausência de consumo significativo de álcool ou outras causas de esteatose, incluindo hepatite viral ou lesão hepática induzida por medicamentos/toxinas. Estes vacúolos podem variar em número e tamanho, representando uma alteração morfológica. Esta condição pode ser benigna, especialmente se detetada num estádio inicial, uma vez que é reversível em muitos casos. No entanto, se não for tratada, a doença do fígado gordo não alcoólica pode progredir para quadros mais graves, como esteato-hepatite (inflamação do fígado), fibrose e, eventualmente, cirrose.

Para o diagnóstico do fígado gordo não alcoólico, deve considerar-se o historial clínico da pessoa e recorrer inicialmente a métodos de diagnóstico não invasivos, como análises ao sangue e exames de imagem. No entanto, nem sempre estes métodos são suficientemente precisos para determinar a extensão da doença, sendo nesses casos necessária a realização de um método considerado invasivo: a biópsia hepática. A biópsia permite uma avaliação detalhada da estrutura celular do fígado, identificando a quantidade de gordura presente nos hepatócitos e outros sinais de inflamação, necessários à classificação da esteatose e consequente diagnóstico.

Apesar dos métodos invasivos fornecerem uma resposta mais precisa relativamente à percentagem de esteatose no fígado, apresentam limitações importantes, como o risco de complicações, o custo elevado e a natureza invasiva do procedimento. O *gold standard* de diagnóstico é um método semi-quantitativo, realizado através de uma biópsia hepática, em que o patologista atribui um score categórico de acordo com o número de hepatócitos com vacúolos. Embora não requeira equipamento especializado, este método está sujeito a uma sobrestimação da percentagem relativa de hepatócitos com esteatose, além de estar sujeito a variações entre observadores, sendo que cada categoria abrange uma ampla faixa percentual de

esteatose. Face a estas limitações, torna-se evidente a necessidade de quantificar a esteatose hepática de forma objetiva.

Neste sentido, surgem novas tecnologias computacionais como: algoritmos de *machine learning*, um ramo da inteligência artificial que permite aos computadores aprenderem padrões a partir de dados e técnicas avançadas de análise de imagem, capazes de quantificar de forma objetiva a esteatose. Entre estas abordagens, destacam-se: (i) segmentação automática por segmentação da componente de saturação da imagem, e (ii) classificação supervisionada por *Random Forest* – Weka (*Waikato Environment for Knowledge Analysis*), que integra atributos morfológicos como circularidade, área e solidez.

Estas ferramentas permitem uma avaliação mais objetiva e padronizada do fígado, minimizando a subjetividade da análise dependente do observador (normalmente o patologista), oferecendo resultados mais consistentes e fiáveis. O uso de dados quantitativos nestas análises é vantajoso, pois fornece informações mais precisas sobre o grau de esteatose e contribui para diagnósticos e prognósticos mais rigorosos.

No contexto do desenvolvimento de novas técnicas de análise, a utilização de modelos animais tem sido fundamental na investigação da EHNA. Animais, como ratinhos, são frequentemente usados para estudar o desenvolvimento da doença e testar potenciais tratamentos, dado que estes modelos conseguem replicar com precisão as condições metabólicas que levam ao aparecimento da doença em humanos.

Neste estudo experimental, o modelo animal utilizado foi o ratinho, sendo submetido a uma dieta controlada, rica em ácidos gordos, para induzir a esteatose hepática. Com base neste modelo animal, propõe-se a utilização de amostras do fígado para analisar a precisão de dois *plugins* de análise automatizada de imagem: *Weka* e *Saturation*, ambos disponíveis na plataforma *FIJI*. O uso de softwares automáticos permite recolher mais informação celular e realizar uma quantificação mais precisa de várias medidas dos componentes celulares, como dos vacúolos. Os resultados obtidos por estes *plugins* foram usados para avaliar a percentagem de esteatose em imagens histológicas do fígado, com base em diversas variáveis da base de dados.

O objetivo principal deste estudo é determinar qual dos dois *plugins* oferece maior precisão na quantificação da esteatose hepática. Durante esta avaliação, o processo será otimizado, determinando-se o número ideal de imagens histológicas necessárias para uma análise fiável. E ainda a melhor combinação ou combinações de ampliação e resolução para uma avaliação precisa da percentagem de gordura hepática. Pretende-se ainda avaliar se o tamanho dos vacúolos lipídicos está relacionado com o grau de esteatose. A resposta a estes objetivos envolveu a aplicação de diversos métodos estatísticos, nomeadamente: o teste de Wilcoxon, para comparar pares de amostras dependentes; o teste sobre o coeficiente de correlação, para avaliar associações entre variáveis; os modelos lineares generalizados mistos com distribuição binomial negativa (GLMM), para modelar o número de vacúolos considerando efeitos aleatórios; a análise de variância, para comparar a qualidade de ajustamento entre modelos; e, por fim, as técnicas de reamostragem bootstrap, para estimar intervalos de confiança e avaliar a estabilidade das estimativas em diferentes combinações de imagens.

Responder a estes objetivos permitirá continuar o desenvolvimento e orientação da forma de classificação da esteatose hepática, essencial para avaliar o grau da doença e, em contexto de transplante, prever complicações e definir estratégias preventivas e terapêuticas. Este aspeto é especialmente relevante perante o aumento da prevalência da doença, associado à obesidade, diabetes e alterações no estilo de vida.

Os resultados obtidos permitiram identificar que apesar das semelhanças entre os métodos, o *plugin Weka* identificou um maior número de vacúolos, embora por vezes com maior dispersão nos dados. A combinação ideal de ampliação e resolução, onde uma menor ampliação (maior área de visualização),

10x, combinada com uma alta resolução (melhor detalhe), 40x ajuda a uma precisa identificação dos vacúolos. Quanto ao número mínimo de imagens necessárias para uma estimativa fiável, um mínimo de 3 imagens são fiáveis, mas 4 imagens permitem uma quantificação mais precisa. Foi ainda possível detetar uma associação entre o tamanho dos vacúolos e o grau de esteatose. Destas conclusões surgiu uma nova pergunta, o número de imagens depende do grau da doença? Com este estudo, demonstra-se que o uso de ferramentas quantitativas e automáticas facilita a padronização, reprodutibilidade e replicabilidade na quantificação da EHNA. Neste contexto, corroboram-se outros estudos ao mostrar que a utilização de métodos automáticos permite uma identificação mais precisa dos vacúolos e, por conseguinte, uma melhor avaliação da percentagem de esteatose em fígados de ratinhos.

A utilização de ferramentas estatísticas tornou-se essencial na análise de dados biológicos, permitindo apresentar resultados mais fiáveis e precisos. Neste cenário, a Bioestatística assume um papel central no processo de investigação, apoiando os investigadores em todas as fases do estudo. A aplicação correta de métodos estatísticos, como por exemplo: (i) os Modelos Lineares Generalizados Mistos, com diferentes distribuições e (ii) a possibilidade de reamostragem através de técnicas *bootstrap*, permite uma análise rigorosa dos dados e uma interpretação fundamentada dos resultados, facilitando a comunicação científica e a tomada de decisões baseadas em evidência.

É importante destacar a relevância deste estudo, assim como a necessidade da sua continuidade, incorporando mais dados: parâmetros bioquímicos e demográficos relativos ao modelo animal, o que poderá viabilizar uma futura aplicação na avaliação de fígados humanos. A possibilidade de expandir este estudo a um maior número de variáveis poderá permitir a construção de um modelo que, com base em variáveis clínicas e demográficas, fornecendo uma estimativa precisa da percentagem de esteatose, que permita definir o grau da doença e orientar o tratamento adequado (prognóstico). Poderá ainda evoluir para uma ferramenta capaz de prever a mortalidade com base nos parâmetros do animal.

Palavras chave: Fígado; Esteatose hepática não alcoólica; Quantificação; GLMM Binomial Negativa; Bootstrap.

Abstract

The liver plays a crucial role in metabolism and immunity, aiding in nutrient metabolism, detoxification, protein production, digestion, and vitamin storage. The liver is essential for overall well-being. Hepatic steatosis, or fatty liver, is primarily caused by obesity, type 2 diabetes, poor diet, and chronic alcohol consumption. A specific form, Non-alcoholic fatty liver disease (NAFLD), is linked to metabolic issues and high-fat diets. NAFLD is characterized by fat accumulation in liver cells without alcohol involvement, ranging from benign steatosis to more severe conditions like cirrhosis.

Diagnosing NAFLD typically involves blood or imaging tests, though liver biopsy remains the most accurate method. However, the standard scoring system for assessing steatosis based on biopsy results is prone to inaccuracies due to variability among observers. Animal models, particularly mice, are commonly used in research to study NAFLD progression and evaluate treatments.

New techniques such as machine learning and quantitative data analysis have improved the accuracy of NAFLD diagnosis. In this study, mouse liver samples will be used to compare two automated image analysis plugins, Waikato Environment for Knowledge Analysis (Weka) and Saturation, to determine their accuracy in assessing steatosis. Weka applies machine learning techniques, specifically the Random Forest classifier, to distinguish vacuoles from other cellular components based on morphological features. The study also aims to identify the optimal combinations of magnification, resolution, and the number of images for accurate analysis, and to assess whether there is a pattern in vacuole size related to the steatosis percentage. The appropriate application of statistical methods, such as generalized linear mixed models and bootstrap techniques, enables robust data interpretation and supports scientific communication and evidence-based decision-making.

Weka detected more vacuoles than Saturation, especially at low magnification and high resolution. Three to four images were sufficient for reliable estimates, and vacuole size was associated with steatosis severity. Accurate classification of steatosis is vital, as NAFLD is a significant manifestation of metabolic syndrome and poses increased mortality risks, particularly from cardiovascular diseases. Early diagnosis and targeted treatment are key to improving patient outcomes and addressing this global health concern.

Keywords: Liver; Non-alcoholic fatty liver disease; Quantification; GLMM Negative Binomial; Bootstrap.

Contents

List of Figures **XI**

List of Tables **XIV**

1 Introduction **1**

2 Non-alcoholic fatty liver disease **4**

2.1 Liver . . . 4

2.2 Liver Steatosis, Non-alcoholic fatty liver disease . . . 6

2.3 Histopathology . . . 12

2.4 Semi-quantitative Pathologist assessment . . . 13

2.5 Quantitative Methods, FIJI Is Just ImageJ . . . 14

2.5.1 Saturation . . . 16

2.5.2 Waikato Environment for Knowledge Analysis . . . 16

2.5.3 Illustrative Example of Visual vs. Automated Steatosis Assessment . . . 17

3 Study Design and Description **19**

3.1 Ethical Aspects Study . . . 20

3.2 Objectives . . . 20

3.3 Variables Description and Outcomes . . . 21

4 Statistical Methodology **25**

4.1 Spearman’s Rank Correlation . . . 26

4.2 Wilcoxon Signed-Rank Test . . . 27

4.3 Models . . . 29

4.3.1 Exponential Family . . . 30

4.3.2 Generalized Linear Models . . . 31

4.3.3 Generalized Linear Mixed Model . . . 34

4.3.4 Model Comparison . . . 35

4.4 Bootstrap Method . . . 36

5 Results and Discussion **38**

5.1 Descriptive Analysis . . . 38

5.1.1 Number of vacuoles for Weka and Saturation: . . . 38

5.1.2 Area of Vacuoles for WEKA and Saturation: . . . 39

5.1.3 Percentage of steatosis for Weka and Saturation: . . . 41

5.2 Evaluation of the performance: Weka versus Saturation . . . 42

5.2.1 Spearman's rank correlation . . . 43
5.2.2 Wilcoxon signed-rank test . . . 47
5.3 Generalized Linear Mixed Model . . . 48
5.3.1 Generalized Linear Mixed Model (GLMM) Univariable Analysis . . . 49
5.3.2 GLMM Multivariable Analysis . . . 52
5.3.2.1 Variance Inflation Factor . . . 53
5.3.2.2 Aditive Model . . . 53
5.3.2.3 Model with Interactions, Program x Magnification . . . 55
5.3.2.4 Model with Interactions, Magnification x Resolution . . . 57
5.3.2.5 ANOVA . . . 58
5.4 Number of Images per Liver . . . 60
5.4.1 Combinatorial Analysis of Image Subsets . . . 62
5.4.2 Bootstrap Analysis . . . 68
5.5 Quantitative Evaluation of Vacuole Size in Mild and Severe Hepatic Steatosis . . . 73

6 Conclusion **75**

References **77**

Appendices **83**

A Brunt system to grade NASH activity **84**

B Quantile-Quantile Plots **86**
B.1 Quantile-Quantile plot of the variable Percentage of Steatosis . . . 86
B.2 Quantile-Quantile plot of the variable Number of vacuoles . . . 86
B.3 Quantile-Quantile plot of the variable Area of vacuoles . . . 87

C Residual Analysis- Model with Interactions, Program x Magnification **88**

List of Figures

2.1 Schematic representation of the liver acinus according to Rappaport's model. Adapted from (Lau et al., 2021). 5

2.2 Anatomy of the mouse liver: The mouse liver has four lobes: Left (largest), right (hemisected), median, and caudate. The gallbladder fundus protrudes below the central isthmus of the medial lobe when viewed from the usual ventral perspective. Adapted from Ali et al. (2019). 6

2.3 Hepatic steatosis results from an imbalance in lipid storage and lipolysis or secretion (adapted from (Gong et al., 2023)). 7

2.4 Global heat map of changing NAFLD prevalence. The prevalence of NAFLD increased from 1990 to 2019 in all countries, except Niger (decreased from 2010 to 2019), Afghanistan, Chad, Guinea, Mozambique, Sierra Leone and Uganda (decreased from 1990 to 2010). NAFLD, non-alcoholic fatty liver disease. Adapted from Wong et al., 2019. 8

2.5 Inflammation associated with NAFLD. Adapted from Younossi et al. (2016). 9

2.6 The Pathogenesis of NAFLD. Adapted from Fang et al. (2022). 10

2.7 The Multisystem Impact of NAFLD 11

2.8 TWS pipeline for pixel classification. Image features are extracted from an input image using FIJI-native methods. Next, a set of pixel samples is defined and represented as feature vectors, and a Weka learning scheme is trained on those samples and finally applied to classify the remaining image data. Adapted from (Arganda-Carreras et al., 2017) 17

2.9 Example of an H&E-stained liver image processed using Weka (40× magnification, 80× resolution). Detected vacuoles are highlighted with green. 17

2.10 H&E liver images at 10x Magnification 18

2.11 H&E liver image, Magnification and Resolution 18

3.1 Study Design 20

3.2 Description of the data collected for this study 22

4.1 Illustration of the bootstrap resampling process. 36

5.1 Distribution of the number of vacuoles identified by each program. The box plots compare the results obtained using the Saturation and Weka methods. 39

5.2 Comparison of Weka and Saturation for the variable number of vacuoles. The histogram represents the distribution of observations, while the overlaid density curve illustrates the estimated probability density function. 39

5.3 Distribution of the area of vacuoles identified by each program. The box plots compare the results obtained using the Saturation and Weka methods. 40

5.4 Comparison of Weka and Saturation for the variable area of vacuoles. The histogram represents the distribution of observations, while the overlaid density curve illustrates the estimated probability density function. 40
5.5 Distribution of the percentage of steatosis identified by each program. The box plots compare the results obtained using the Saturation and Weka methods. 41
5.6 Distribution of the number of vacuoles detected by Saturation and Weka across different combinations of magnification and resolution. Each box plot represents the distribution of *Vacuole_nr* values for a given condition, with individual data points overlaid. The figure illustrates the variability and the differences in plugin performance in vacuole detection under varying imaging conditions. 42
5.7 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 10×10) 44
5.8 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 10×20) 44
5.9 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 10×40) 45
5.10 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 20×20) 45
5.11 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 20×40) 45
5.12 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 20×80) 46
5.13 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 40×40) 46
5.14 Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 40×80) 46
5.15 Box plots showing vacuole count distribution across different image magnification levels. 50
5.16 Boxplots showing the distribution of vacuole counts across different image resolution levels. 51
5.17 Multiplicative effects of the variables *program*, *magnification*, and *resolution* on the expected vacuole count, expressed as exp(β), with error bars representing confidence intervals.. 51
5.18 Multiplicative effects of each predictor on the expected vacuole count, estimated from a generalized additive linear mixed model with negative binomial distribution. The bars represent exponentiated coefficients (exp(Estimate)), which indicate how each variable influences the mean vacuole count relative to the reference levels (Saturation program, 10× magnification, and resolution 10). Error bars represent 95% confidence intervals. . 55
5.19 Predicted number of vacuoles across resolution levels (10, 20, 40, 80) for each magnification level (10×, 20×, 40×), based on a negative-binomial GLMM including the Program x Magnification interaction. 57
5.20 Predicted number of vacuoles across resolution levels (10, 20, 40, 80) for each magnification level (10×, 20×, 40×), based on a negative-binomial GLMM including the Program x Magnification interaction. 58
5.21 Distribution of the median steatosis percentages across different image combinations for specimen 2. 62

5.22 Specimen 5 and 6 image analysis 63
5.23 Specimen 7 and 9 image analysis 63
5.24 Specimen 13 and 14 image analysis 64
5.25 Specimen 15 and 17 image analysis 64
5.26 Specimen 1 and 4 image analysis 65
5.27 Specimen 8 and 10 image analysis 66
5.28 Specimen 11 and 12 image analysis 66
5.29 Distribution of the median steatosis percentages across different image combinations for specimen 16. 67
5.30 Distribution of the median steatosis percentages across different image combinations for specimen 3. 67
5.31 Bootstrapped median values of the percentage steatosis for specimens 3 and 5. 69
5.32 Bootstrapped median values of the percentage steatosis for specimens 6 and 7. 70
5.33 Bootstrapped median values of the percentage steatosis for specimens 9 and 10. 70
5.34 Bootstrapped median values of the percentage steatosis for specimens 14 and 17. 71
5.35 Scatter plot showing the average area of each vacuole (in μm^2) across individual specimens. Each dot represents an image, coloured by steatosis severity (Ratio: Mild vs Severe). The horizontal distribution reflects vacuole size, and the vertical axis distinguishes specimens. Specimens with Severe steatosis consistently show higher vacuole area values compared to those with Mild steatosis. 74

B.1 Quantile-Quantile plot of the variable Percentage of steatosis. 86
B.2 Quantile-Quantile plot of the variable number of vacuoles. 87
B.3 Quantile-Quantile plot of the variable area of vacuoles. 87

C.1 DHARMa Residual Diagnostics: Quantile-Quantile Plot and Residuals vs. Predicted. . . 89

List of Tables

3.1 Database variables 23

3.2 Number of Images per Sample and Magnification 24

4.1 Types of alternative hypotheses for Spearman's rank correlation 27

4.2 p-values for alternative hypotheses using the normal approximation 27

4.3 Alternative hypotheses for the Wilcoxon Signed-Rank Test. 28

5.1 Spearman correlation coefficients r_s and p-values by AR group 44

5.2 Wilcoxon Test for paired samples for comparing Weka vs Saturation 47

5.3 Univariable GLMM (Negative Binomial) results for each qualitative predictor. Estimates are on the log scale; exponentiated estimates represent multiplicative effects on the expected number of vacuoles. 49

5.4 Multiplicative effects (exp(β)) of predictors on mean vacuole count from the Negative Binomial (nbinom2) mixed model. All predictors are statistically significant ($p < 0.001$). 54

5.5 Multiplicative effects (exp(β)), 95% confidence intervals, and p-values from the Negative Binomial (nbinom2) mixed model with interaction Program x Magnification (`glmmTMB`). 56

5.6 Multiplicative effects (exp(β)), 95% confidence intervals, and p-values from the Negative Binomial (nbinom2) mixed model with `Magnification x Resolution` interaction (`glmmTMB`). 57

5.7 Model comparison using likelihood ratio tests (Additive GLMM and GLMM including interaction Program x Magnification) 59

5.8 Model comparison using likelihood ratio tests (Additive GLMM and GLMM including interaction Magnification x Resolution) 59

5.9 Number of images per specimen at 10x magnification x 40x resolution 60

5.10 Frequency of Lowest Variability and Closest Median to the True Value Across Bootstrap 50 Replicates (k = 3, 4, 5). Variability inferred from Bootstrap 95% CI, and the true value corresponds to the global median. 72

5.11 Bootstrap results for all specimens: standard error (SE), and 95% confidence interval by number of images. 72

A.1 Comparison of key histological features and scoring criteria among the Brunt, NASH CRN, and SAF systems for grading and staging NAFLD/NASH. Adapted from (Kleiner et al., 2005). 85

List of Abbreviations and Acronyms

AIC	Akaike Information Criterion
ALD	Alcoholic Liver Disease
ANOVA	Analysis of Variance
AR	Magnification x Resolution
BIC	Bayesian Information Criterion
CT	Computed Tomography
EHNA	Esteatose hepática não alcoólica
FIJI	Fiji Is Just ImageJ
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
H&E	Haematoxylin and Eosin
LM	Linear Models
LRT	Likelihood Ratio Test
MLE	Maximum Likelihood Estimation
MRI	Magnetic Resonance Imaging
NAFLD	Non-alcoholic fatty liver disease
NAS	Non-alcoholic fatty liver disease Activity Score
NASH	Non-alcoholic steatohepatitis
NLMMs	Nonlinear mixed models
OCT	Optimal Cutting Temperature compound
PEM	Protein-energy malnutrition
RGB	Red Green and Blue
SVM	Support Vector Machines
VIF	Variance Inflation Factor
Weka	Waikato Environment for Knowledge Analysis

Chapter 1

Introduction

The liver, a vital organ in humans and animals, plays a significant role in metabolism and immunity (e.g., Kupffer cells, specialized macrophages found in the hepatic sinusoids of the liver). It metabolizes nutrients, detoxifies harmful substances, and produces important proteins. It also helps in digestion and stores essential vitamins. Given its multifaceted functions, maintaining its health is crucial for overall well-being.

The leading causes of hepatic steatosis, or fatty liver, are numerous and diverse. These include obesity, insulin resistance, type 2 diabetes, metabolic syndrome, sedentary lifestyle, poor diet, high levels of cholesterol and triglycerides, chronic alcohol consumption, nutritional deficiencies, medication, hepatitis, and malnutrition. Despite this wide range of causes, this work focuses on a specific disease that is not alcohol-related, namely Non-alcoholic fatty liver disease (NAFLD), which is generally linked to metabolic issues and a high-fat diet. NAFLD is defined as the presence of fat in $\geq 5\%$ of hepatocytes, in the absence of other causes of liver disease, hepatocellular injury, or fibrosis (Brunt, 2010; Younossi et al., 2016). Hepatic steatosis can be benign and, if detected in an early stage, is a reversible condition. It has a spectrum that ranges from simple steatosis to progressive inflammation (steatohepatitis), and eventually to fibrosis and cirrhosis.

The diagnosis of NAFLD can be made with non-invasive methods such as blood tests or imaging tests, but these methods are not precise in determining the severity of the condition. A more accurate yet invasive alternative is a liver biopsy. In this procedure, a pathologist assesses the biopsy through the visual examination of histological liver tissue slides. This evaluation is typically performed using a widely employed semi-quantitative grading system. One commonly used grading system is the NAFLD Activity Score Non-alcoholic fatty liver disease Activity Score (NAS), which grades key histological features such as steatosis, lobular inflammation, and hepatocellular ballooning. For instance, steatosis is graded as an interval percentage: 0 (<5%), 1 (5–33%), 2 (34–66%), or 3 (>66%). While this system provides a structured approach to classifying disease severity, it is inherently subjective. This work focuses on assessing the degree of steatosis, excluding other cellular findings included in the NAS.

The subjectivity of this assessment arises from its dependence on the visual interpretation of the pathologist's histological slides. This can lead to variability due to intra-observer (the same pathologist giving different assessments at different times) and inter-observer (different pathologists providing inconsistent assessments of the same slide) differences. Furthermore, the scoring system uses broad categories, such as 5%–33% for grade 1 steatosis, reducing sensitivity to small but clinically relevant changes. As a result, while this method is systematic, it introduces potential inaccuracies that can affect clinical decision-making. Efforts to standardize training and implement automated or AI-assisted grading methods aim to reduce this subjectivity and improve reliability (Arganda-Carreras et al., 2017; Hallou et al., 2021).

With new techniques such as training algorithms and machine learning—particularly Random Forest-based classifiers applied to pixel-level segmentation in histological images—it is now possible to develop more accurate methods for assessing NAFLD, reducing the risk of misinterpretation (Arganda-Carreras et al., 2017; Hallou et al., 2021). The main advantage of using quantitative data is that it can provide objective, reliable, and reproducible results. Quantitative data can be analysed using rigorous and standardized methods, including mathematical formulas, statistical models, and dedicated image analysis software.

The classification of the degree of steatosis is imperative for clinical prognosis, particularly as NAFLD is now seen as the hepatic manifestation of the metabolic syndrome and is associated with an increased risk of mortality, mainly due to cardiovascular complications, which constitutes a global health problem (Chalasani et al., 2018). Early and accurate diagnosis and targeted interventions are needed to mitigate the long-term risks associated with this condition and improve patient outcomes.

The use of animal models in research has been pivotal in advancing scientific understanding and facilitating medical progress. These models offer critical insights into biological processes, disease mechanisms, and therapeutic strategies, enabling studies that are otherwise impractical or ethically unfeasible in human subjects. The lack of approved pharmacotherapies for NAFLD/Non-alcoholic steatohepatitis (NASH) underscores the urgent need for robust preclinical animal models that accurately reflect the pathophysiological features of human NAFLD/NASH.

In this study, mouse models are employed due to their cost-effectiveness and suitability for genetic manipulation, making them a widely used tool in research. However, these models possess inherent limitations, particularly their inability to fully recapitulate the complex characteristics of human NAFLD/NASH. As highlighted by Fang et al. (2022), each mouse model presents distinct strengths and weaknesses, emphasizing the necessity of developing novel models and advanced methodologies that more precisely replicate the mechanisms and progression of human NAFLD. By leveraging liver samples from animal models exhibiting steatosis (mice subjected to a controlled hypercaloric diet), we propose to:

1. Compare two automated image analysis plugins, Weka and Saturation, both implemented in Fiji Is Just ImageJ (FIJI), to determine which is more accurate;

2. Determine the optimal magnification vs. resolution configuration for reliable vacuole detection;

3. Establish the minimum number of histological images per liver needed for stable quantification;

4. Evaluate whether vacuole size is associated with steatosis percentage.

To achieve these goals:

- For objective (1), automated workflows were developed using FIJI macros for both plugins, supported by logistic regression classifiers for vacuole segmentation. The performance of each plugin was compared using descriptive statistics for each variable, Spearman's rank correlation, and the Wilcoxon signed-rank test.

- Objective (2) was addressed using generalized linear mixed models (GLMM) with a negative binomial distribution, accounting for overdispersion and within-subject variability across image configurations.

- For objective (3), bootstrap resampling techniques were applied to estimate confidence intervals and assess the stability of steatosis estimates across different numbers of images per sample.
- Lastly, objective (4) was examined through graphical analysis, using scatter plots to evaluate the association between the average area of vacuoles and the severity of the disease.

This thesis is structured into six chapters. In Chapter 1, a brief introduction to the topic is presented to facilitate the interpretation and understanding of the results, along with a description of the project and its intended objectives.

In Chapter 2, Non-Alcoholic Fatty Liver Disease begins by introducing the liver's essential role in the human body, highlighting its key functions in metabolism, detoxification, digestion, and immunity. It then provides an overview of hepatic steatosis (fatty liver), explaining its definition, underlying causes, and impact on health. A specific focus is given to NAFLD, covering its prevalence, characteristic features, progression through various disease stages, and associated comorbidities. The discussion extends to the challenges in diagnosing NAFLD, outlining current diagnostic methods and their limitations. It further explores histopathological techniques used to evaluate liver tissue, particularly the semi-quantitative assessment performed by pathologists. It introduces quantitative methods for analyzing liver steatosis, specifically the Weka and Saturation image analysis techniques, which offer potential advantages over traditional visual assessment methods.

Chapter 3, Study Design and Description, outlines the study design employed to achieve the study objectives. It comprises a detailed description of the methods, variables, study instruments, procedures, and analytical techniques used. Also focuses on the ethical aspects and objectives, and explores the dataset used.

In Chapter 4, Statistical Methodology, we will focus on the presentation of theoretical concepts related to the statistical model used, namely the Generalized Linear Mixed Model (GLMM) with a negative binomial distribution, which was used to account for overdispersion in the count data. A combinatorial strategy was applied to explore multiple image subsets and assess their predictive power and non-parametric bootstrap techniques.

Chapter 5, Results and Discussion, presents the analysis of the results obtained for the defined objectives.

Finally, Chapter 6, Conclusion, is dedicated to the discussion of the results, the conclusions of the study, and subsequent final remarks.

Chapter 2

Non-alcoholic fatty liver disease

2.1 Liver

The liver is one of the vital organs in an organism, responsible for numerous essential functions necessary to maintain overall health and well-being. Both macroscopic and microscopically, the liver's appearance in colour, size, location, and shape varies amongst species. In humans, it is the largest internal organ and gland, located in the right upper quadrant of the abdomen. In mice, the liver is composed of four lobes: left, right, median, and caudate, located in the cranial part of the abdominal cavity, and occupy the entire subdiaphragmatic region.

The liver plays a crucial role in many physiological processes, such as:

- **Lipid metabolism** - where it processes and converts nutrients into an energy source by synthesizing cholesterol and triglycerides;
- **Bile production** - a digestive fluid that is stored in the gallbladder, crucial for fat digestion and absorption;
- **Protein synthesis** - in which the liver synthesizes various proteins like albumin and clotting factors;
- **Urea cycle** - by converting excess amino acids into urea;
- **Carbohydrate metabolism** - through regulation of blood glucose, in which the liver helps regulate blood sugar levels by storing excess glucose as glycogen (in a process called glycogenesis) when blood sugar levels are high and releasing glucose into the bloodstream when blood sugar levels are low;
- **Detoxification** - as the name implies, detoxing the blood from harmful substances, as observed through the hydrogen sulfide metabolism regulation and homeostasis (hepatic lipotoxicity);
- **Storage and management of essential nutrients**, like vitamins and minerals;
- **Immunological defence** - through the presence of Kupffer cells, which are specialized macrophages, essential for defending against various pathogens, contributing to the body's innate immune response. In addition to their immunological role, Kupffer cells are involved in recycling red blood cells and filtering toxins from the bloodstream. They also regulate inflammatory responses by releasing cytokines, thereby maintaining liver homeostasis and overall immune balance. Recent studies reveal that they also play a significant role in insulin resistance and the progression of non-alcoholic steatohepatitis (NASH).

Due to the illustrated importance of the liver in assuring an organism's health, it is important to devise mechanisms for diagnostic, prognostic, and even therapeutic interventions of the more common liver diseases, (Manikat & Nguyen, 2023; Trefts et al., 2017).

Rappaport, 1954 proposed the hepatic acinar model, distinguishing zones within the liver based on gradients of oxygen, nutrients, and toxins. This model divides the hepatic lobules into three zones around the portal tract and the central vein, Figure 2.1:

- Zone 1, known as the periportal zone, is located closest to the portal triad (portal vein, hepatic artery, and bile duct). It receives the most highly oxygenated and nutrient-rich blood and is the most active in processes such as gluconeogenesis, β-oxidation, and cholesterol synthesis.
- Zone 2 represents an intermediate region between Zones 1 and 3, exhibiting mixed characteristics.
- Zone 3, also known as the centrilobular zone, is located near the central vein. As the last region to receive blood flow, it is supplied with the least oxygenated blood, making it particularly vulnerable to hypoxia, toxic injury, and necrosis.

Zone 3 is especially relevant in this study, as it is the primary site where steatosis often initiates in several liver diseases, particularly NAFLD, making it a critical area for detecting early pathological changes. Moreover, Zone 3 hepatocytes are heavily involved in glycolysis, lipogenesis, and drug metabolism, further contributing to their susceptibility to injury under metabolic stress (Rappaport, 1954). Although Rappaport's acinar model was originally based on human liver anatomy, its fundamental principles apply to rodent models as well, despite some differences in lobular structure and enzyme expression. In this work, we did not partition images by zone; instead, we used a random sampling approach so that each field of view could include a mixture of zones, thereby avoiding bias from manual selection of specific regions.

The liver comprises various cell types with different embryological origins; however, this work will focus on the hepatocyte. Hepatocytes are the primary epithelial cells of the liver, constituting the majority of its volume and performing many of its essential functions. These cells are critical for metabolic processes, detoxification, and immune regulation, playing a central role in maintaining liver homeostasis. Recent research, as highlighted by Gong et al. (2023), underscores the pivotal role of hepatocytes in liver inflammation, demonstrating how these cells release pro-inflammatory factors in response to chronic liver injury.

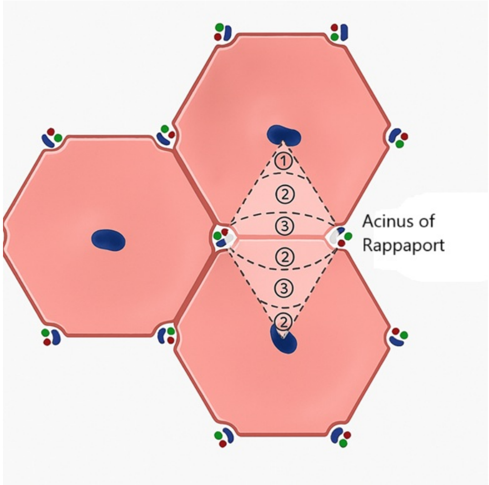


Figure 2.1: Schematic representation of the liver acinus according to Rappaport's model. Adapted from (Lau et al., 2021).

Given their significance, this study will focus on the pathological processes affecting hepatocytes and examine their metabolic responses to these disturbances.

Animal models are indispensable for elucidating pathogenic mechanisms due to their cost-effectiveness, rapid disease progression, controlled experimental conditions, and suitability for genetic engineering. Mouse models are particularly valuable for studying non-alcoholic fatty liver disease (NAFLD) because of significant anatomical, physiological, and genetic similarities between mouse and human livers. These parallels enable researchers to replicate the progression of human NAFLD and investigate its underlying mechanisms.

Notably, mice share key similarities with humans in liver architecture (not anatomical because mice liver as four lobes, as show in Figure 2.2), metabolic functions, and genetic regulation. For example, mouse models exhibit liver inflammation and fibrosis, replicating crucial features of human non-alcoholic steatohepatitis (NASH). Gene-set enrichment analyses reveal up to 90% similarity in liver gene expression changes between mice and human NASH patients, underscoring the relevance of these models in studying inflammatory aspects of NAFLD.

However, mouse models have limitations, as they cannot fully replicate the complexity of human NAFLD/NASH. Differences in lipid metabolism, immune responses, fibrosis patterns, and genetic variability lead to discrepancies in the disease's appearance, progression, and outcomes between species. Despite these challenges, the morphological evaluation of NAFLD/NASH in animal models remains a cornerstone of pathologic analysis, as the morphological changes often reflect common endpoints of pathogenic pathways, regardless of the primary etiology.

This study focuses on quantifying steatosis in mouse models by assessing vacuole distribution using advanced imaging plugins. Future research will be necessary to adapt these methods for evaluating human liver samples, bridging the gap between preclinical findings and clinical applications (Denk et al., 2019; Fang et al., 2022).

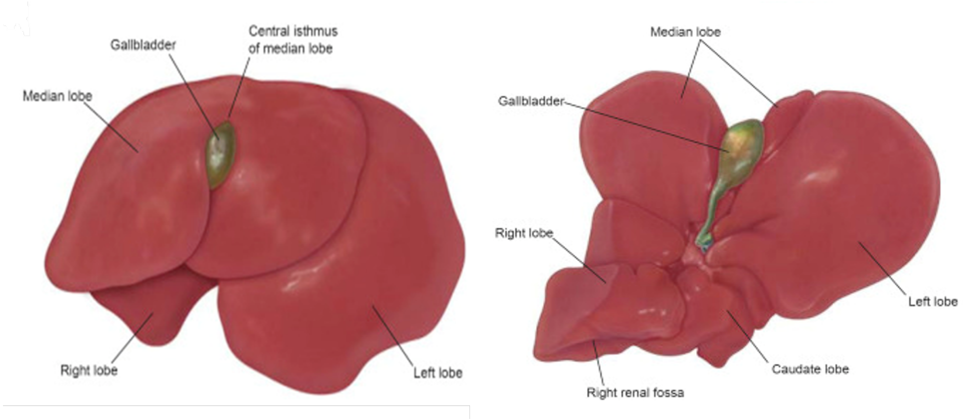


Figure 2.2: Anatomy of the mouse liver: The mouse liver has four lobes: Left (largest), right (hemisected), median, and caudate. The gallbladder fundus protrudes below the central isthmus of the medial lobe when viewed from the usual ventral perspective. Adapted from Ali et al. (2019).

2.2 Liver Steatosis, Non-alcoholic fatty liver disease

Lipids are a broad group of organic molecules that have numerous biological functions in the human body, such as acting as structural components of cell membranes, serving as energy storage sources, and participating in signalling pathways (Fahy et al., 2011). The three main types of lipids are triacylglycerols (also known as triglycerides), phospholipids, and cholesterol, all of which can accumulate in all cells.

Steatosis is an abnormal accumulation of excess fat within cells or organs. If present in the liver, it is called hepatic steatosis or fatty liver disease, microscopically seen as lipid droplets in the parenchyma cells (Brunt, 2010; Fong et al., 2000). This accumulation is particularly notorious in the hepatocytes.

When lipids accumulate, they give origin to steatosis Figure 2.3. In a microscopic view, they give the liver a gross view of fatty livers, which are pale brown or yellow, heavy, and greasy. Microscopically, liver steatosis is initially observed as small white round vacuoles within the cytoplasm and beside the nucleus (lipidic droplets). As the fat accumulation progresses, the small vacuoles cluster into prominent ones, causing displacement of the nucleus.

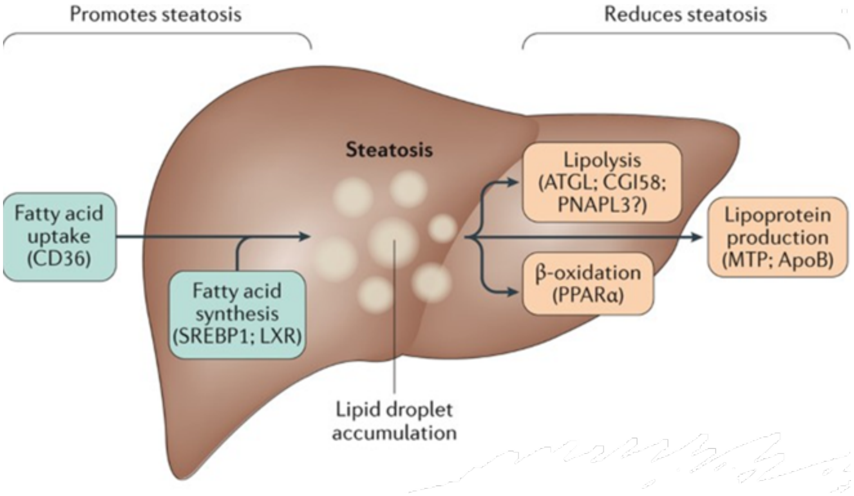

Promotes steatosis
Reduces steatosis
Steatosis
Fatty acid uptake (CD36)
Fatty acid synthesis (SREBP1; LXR)
Lipolysis (ATGL; CGI58; PNAPL3?)
β-oxidation (PPARα)
Lipoprotein production (MTP; ApoB)
Lipid droplet accumulation

Figure 2.3: Hepatic steatosis results from an imbalance in lipid storage and lipolysis or secretion (adapted from (Gong et al., 2023)).

Hepatic steatosis is defined as the intrahepatic triacylglycerol accumulation of at least 5% of liver weight or 5% of hepatocytes containing lipid vacuoles. If detected early, hepatic steatosis can be reversible. It implies a healthy lifestyle, physical activity, and dietary habits (Nassir et al., 2015).

When the liver becomes steatotic, the liver functions related to detoxification, digestion, glucose and lipid metabolism, protein synthesis, immune regulation, nutrient storage, and hormone balance are compromised.

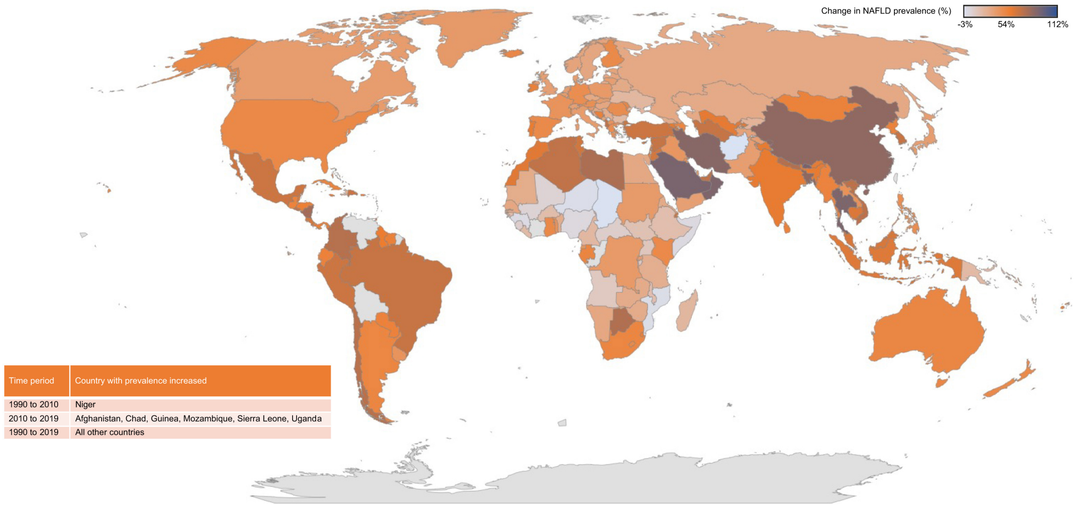
Hepatic steatosis causes many health problems, which can lead to NAFLD and Alcoholic Liver Disease (ALD) (Kumar et al., 2015).

NAFLD, first described by Ludwig et al., 1980, represents a clinically significant condition characterized by fat accumulation in at least 5% of hepatocytes. NAFLD encompasses a spectrum of benign, reversible pathological changes, such as liver steatosis without necro-inflammatory injury, which may progress to more severe and irreversible pathological states (Cobbina & Akhlaghi, 2017). It is distinctly differentiated from other liver disorders by the absence of concurrent liver diseases and histological evidence of hepatocellular injury or fibrosis (Brunt, 2005; Drew, 2017). Unlike alcohol-induced cirrhosis, NAFLD is unrelated to alcohol abuse and excludes secondary causes of hepatic fat accumulation, such as prolonged exposure to steatogenic medications or monogenic hereditary disorders (Angulo, 2002; Chalasani et al., 2018; Li et al., 2010; Younossi et al., 2004).

As stated by Leow et al., 2023, (NAFLD), steatosis typically originates in the perivenular region (zone 3). The variability observed is mainly in the extent of lipid accumulation, rather than in the distribution pattern. As the disease progresses, fat deposition can extend to other hepatic zones; however, the initial perivenular pattern generally remains consistent. This zonal preference is believed to be linked to regional differences in oxygenation and metabolic activity, which make zone 3 hepatocytes more susceptible to lipid accumulation and oxidative stress. Understanding this distribution is crucial for accurate histopathological evaluation and disease staging.

Accordingly, Schwabe et al., 2012 reports that approximately 24% of NAFLD patients may develop active lesions characterized by hepatocyte injury, cell death, and inflammation, leading to (NASH) and cirrhosis. This fatty liver disease is a prominent cause of liver disease mortality, with a concerning annual increase in prevalence worldwide, making it a significant public health issue (Chew et al., 2024; Clark et al., 2002; Vanderbeck et al., 2013). Type 2 diabetes mellitus and cardiometabolic outcomes in metabolic dysfunction-associated steatotic liver disease population (Chew et al., 2024).

Younossi et al., 2016, indicate considerable heterogeneity among studies for both NAFLD prevalence and incidence. Several studies suggest differences in the prevalence and severity of NAFLD by race or ethnicity, which may be linked to differences in lifestyle, diet, metabolic comorbidity profile, and genetic background, among others. Ethnic differences from different parts of the world are still relatively scarce and specific to a few countries, Figure 2.4. Younossi et al., 2016 estimated the NAFLD prevalence at 25.2% and the NASH prevalence at 3% to 5%, using imaging tests for diagnosis. To corroborate these two studies, Cholongitas et al. (2021) performed a meta-analysis which validated that the prevalence is similar to the global rates (>25%) in European adults and higher if taking into account patients with obesity and/or metabolic syndrome. Recent data from Portugal estimate the prevalence of NAFLD at 17%, based on a population-based study that excluded other causes of liver disease through clinical and laboratory criteria (Leitão et al., 2020). Although time series data are lacking, this estimate is consistent with the global trend of increasing prevalence, and the growing burden of NAFLD is expected to mirror the rise in obesity and metabolic syndrome in the Portuguese population. It is becoming the most common liver disease. Consequently, according to Nassir et al., 2015, 70% of overweight and diabetic adult individuals suffer from NAFLD, and the values rise to 90% morbidly in obese adults. NAFLD is also engaging attention as a significant form of liver disease present in pediatric populations, which in some patients advances to more severe liver diseases in adulthood (Kleiner et al., 2005). Surprising data shows the presence in 3 to 10% of normal-weight children and 50% of obese children, according to an Italian study (Crte et al., 2012).



Time period	Country with prevalence increased
1990 to 2010	Niger
2010 to 2019	Afghanistan, Chad, Guinea, Mozambique, Sierra Leone, Uganda
1990 to 2019	All other countries

Figure 2.4: Global heat map of changing NAFLD prevalence. The prevalence of NAFLD increased from 1990 to 2019 in all countries, except Niger (decreased from 2010 to 2019), Afghanistan, Chad, Guinea, Mozambique, Sierra Leone and Uganda (decreased from 1990 to 2010). NAFLD, non-alcoholic fatty liver disease. Adapted from Wong et al., 2019.

This is particularly evident in developed countries, where diets rich in processed foods and sugar, combined with high obesity rates, contribute significantly to the burden of NAFLD (Younossi et al., 2016). High-calorie diets, sedentary lifestyles, and ageing of the global population are all risk factors for chronic liver disease, leading to the current obesity epidemic, with increasing prevalence of metabolic syndrome among adults and children, and global rise of NAFLD and NASH (Cholongitas et al., 2021; Sayiner et al., 2016).

NAFLD efficiently progresses to NASH. This development is microscopically illustrated by steatosis, hepatocyte injury (e.g., ballooning), which could progress to cirrhosis, fibrosis, lobular and portal inflammation and necrosis, Mallory bodies, liver-related complications and even hepatocellular carcinoma (Angulo, 2002; Brunt, 2010; Huby & Gautier, 2022; Mendler et al., 2005). Much progress has been made in understanding the pathogenesis of NASH and the complex and multifactorial molecular pathways during development and progression (Hardy et al., 2016; Machado & Diehl, 2016).

It is important to note that in (NAFLD), chronic inflammation caused by hepatic fat accumulation triggers the production of acute-phase inflammatory proteins in Figure2.5. These proteins are key components of the systemic inflammatory response and play a crucial role in the disease's progression. Elevated levels of these proteins can exacerbate liver injury, leading to more severe conditions such as NASH, fibrosis, and cirrhosis. Additionally, this chronic inflammatory state is associated with increased oxidative stress and insulin resistance, which further promote liver damage and metabolic complications. Understanding the role of acute-phase proteins in NAFLD is essential for developing targeted therapeutic strategies to prevent disease progression (Koruk et al., 2003).

Distinguishing between NAFLD and NASH is crucial because these conditions have different pathogeneses and outcomes. It is therefore crucial to know the disease stage to distinguish between mild and severe phenotypes of NAFLD, NASH, and cirrhosis. Identification of the disease stage will ensure the choice of correct treatment and follow-up (Angulo, 2002; Li et al., 2010; Mendler et al., 2005; Vanderbeck et al., 2013). Early disease recognition and precautionary measures are thus imperative to prevent progression and control the possible advanced outcome of the disease in terms of disability and death (Munsterman et al., 2019).

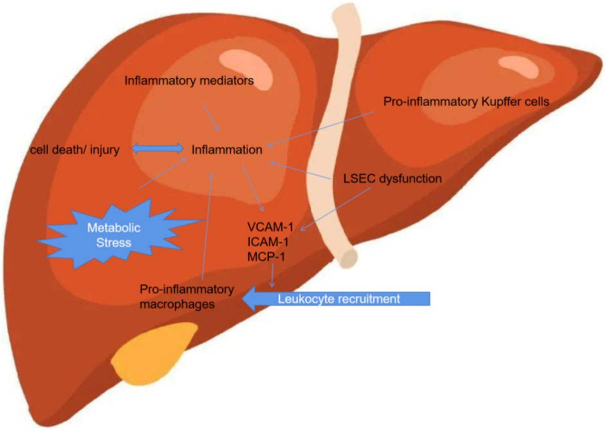


Figure 2.5: Inflammation associated with NAFLD. Adapted from Younossi et al. (2016).

NAFLD can progress through five stages Figure 2.6, but the order may vary per individual.

1. **Steatosis (simple fatty liver)** deposition in at least 5% of hepatocytes – is a mostly harmless build-up of fat in the liver cells that may only be diagnosed during tests carried out for other motives.

2. **Non-alcoholic steatohepatitis (NASH)** – a more severe form of NAFLD, where the liver has become inflamed.

3. **Fibrosis*** – where persistent inflammation causes scar tissue around the liver and nearby blood vessels, but the liver can still function normally.

4. **Cirrhosis*** – the most severe stage, occurring after years of inflammation, where the liver shrinks and is lumpy due to extensive fibrosis; this damage is permanent and can lead to liver failure and liver cancer.

5. **Hepatocellular carcinoma** – liver cancer.

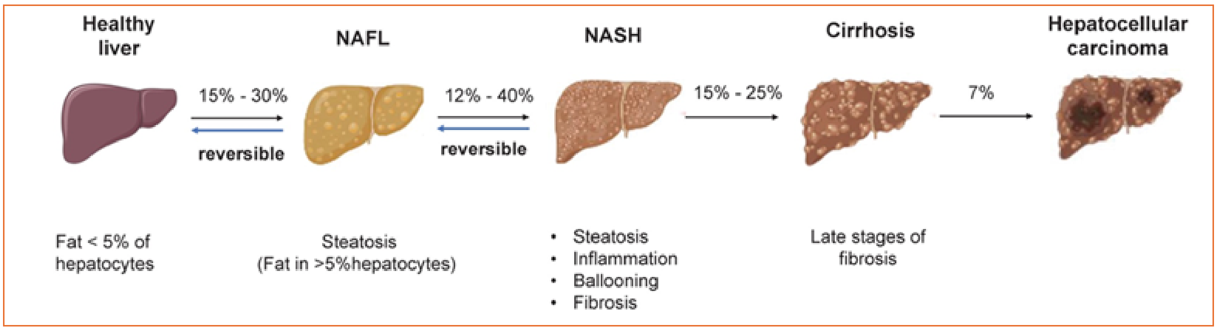
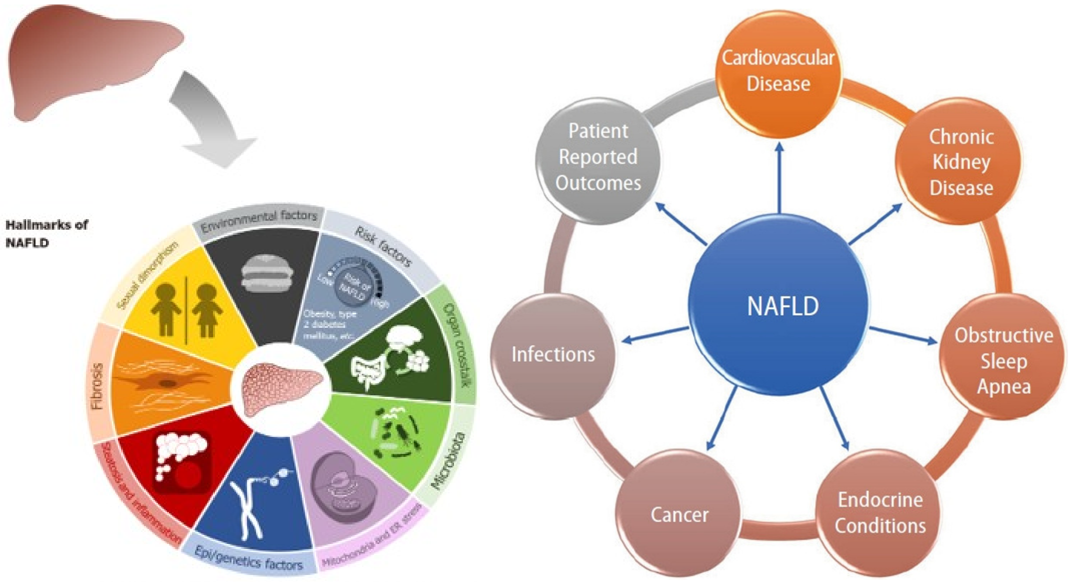


Figure 2.6: The Pathogenesis of NAFLD. Adapted from Fang et al. (2022).

NAFLD has been associated with a range of non-liver comorbidities, including anthropomorphic variables such as weight, sociodemographic variables like age and gender, or comorbidities such as obesity and genetics. Frequently, risks associated with metabolic comorbidities, for instance, overweight or obesity, might be a consequence of a sedentary lifestyle, fast-food consumption among children and adults, type 2 diabetes mellitus- insulin resistance, hypertension, atherosclerosis, and systemic micro-inflammation, jejunoileal bypass, and dyslipidemia. Therefore, the manifestation of metabolic syndrome remains controversial, as it is unclear whether NAFLD is a cause or a consequence of glucose intolerance and insulin resistance (Angulo, 2002; Chalasani et al., 2018; Clark et al., 2002; Fong et al., 2000; Kleiner et al., 2005; Manikat & Nguyen, 2023; Nassir et al., 2015; Tanaka et al., 2019; Younossi et al., 2004).



(a) Hallmarks of NAFLD. Adapted from (Martin-Grau et al., 2022).

(b) Associative relationships between NAFLD and various health conditions.

Figure 2.7: The Multisystem Impact of NAFLD

Steatosis due to triglycerides has been observed in a range of clinical conditions, including nutritional, metabolic, alcohol consumption, obesity,NASH, type 2 diabetes mellitus, syndrome X, and drug-induced processes (Fong et al., 2000).

The leading causes of triglyceride accumulation are toxins, protein-energy malnutrition, type 2 diabetes mellitus (insulin resistance), hepatitis C virus infection, Wilson disease, anoxia, obesity, high fat and carbohydrate intake, alcohol, drugs, and abnormal lipid metabolism. Furthermore, it is well recognized that older age, ethnicity and race, starvation, and many everyday medications may also contribute to ectopic hepatic lipid accumulation (Brunt & Tiniakos, 2010; Fong et al., 2000; Momose et al., 2011; Promrat et al., 2004). The leading causes of triglyceride accumulation are toxins, Protein-energy malnutrition (PEM), type 2 diabetes mellitus (insulin resistance), hepatitis C virus infection, Wilson disease, anoxia, obesity, high fat and carbohydrate intake, alcohol, drugs, and abnormal lipid metabolism. Furthermore, it is well recognized that older age, ethnicity and race, starvation, and many everyday medications may also contribute to ectopic hepatic lipid accumulation (Brunt & Tiniakos, 2010; Fong et al., 2000; Momose et al., 2011; Promrat et al., 2004).

The degree of steatosis is a critical factor in determining an organ's transplantability. Livers with more than 30% fat content have a 25% chance of developing primary non-function, significantly reducing the survival of liver grafts after transplantation (de Graaf et al., 2011; Fiorini et al., 2004; Qayyum et al., 2012).

Often, NAFLD is diagnosed based on the analysis of the clinical historical and complementary exams, such as imaging studies and blood tests, which normally detect if the liver enzymes are high. The diagnosis is usually done progressively, starting from the less invasive methods, and accordingly, these results go far beyond other methods.

- **Blood tests (biomarkers)** - A panel of liver function tests, possibly combined with genetic testing, can be used as an initial assessment. However, blood tests do not always detect the presence of NAFLD.

- **Imaging tests** - An ultrasound, Computed Tomography (CT) scan, and Magnetic Resonance

Imaging (MRI) can show liver damage.

- **Biopsy** - The most invasive test performed on an individual. A small sample of liver tissue is collected using a thin needle and analysed by a pathologist. In this case, the pathologist examines the liver microscopically to analyse the cellular features present and determine the stage of the disease.

Despite its limitations, liver biopsy remains a valuable diagnostic tool due to its ability to provide a definitive diagnosis through detailed histological evaluation. It facilitates the assessment of disease severity by determining the extent of liver damage and enables the differentiation between simple steatosis and non-alcoholic steatohepatitis (NASH). Furthermore, it aids in excluding other liver diseases, such as autoimmune hepatitis and hemochromatosis, and offers prognostic value by serving as a critical predictor of disease progression. The degree of fibrosis identified through biopsy is particularly significant in predicting outcomes in patients with NAFLD, although fibrosis is not the primary focus of this study. Consequently, early detection of NAFLD, coupled with appropriate lifestyle modifications, has the potential to halt or reverse disease progression (Fernando et al., 2019).

2.3 Histopathology

Histology is the study of tissues and their microstructure, whereas Pathology is the study of disease. Similarly, Histopathology is the study of tissues affected by disease. Histopathology is extremely useful in making a diagnosis and determining the severity and progress of a condition. Disease processes affect tissues and species in distinct ways, depending on the tissue type, the disease itself, and how it has progressed.

This biology area is also used extensively in biomedical research to identify the causes and possible treatments for disease. This type of research may take place in hospitals, universities, research institutes, and pharmaceutical companies. Biomedical scientists may receive different samples, from a tiny biopsy to an intact organ. This study field may include live and post-mortem samples. The conventional work of a biomedical scientist involves the preparation of tissues for microscope observations by the Pathologist. This technique requires knowledge from biology to chemistry and work experience. The tissue goes through various stages along the way:

- **Fixation** – The main goal is to preserve the tissue, cells, and subcellular components from autolysis or putrefaction. This means keeping all body components as close as possible to their living state. Upon cell death, enzymes are released that begin to break down the tissue components. Therefore, it is carried out immediately after removing the tissue after surgical pathology or immediately after death. A physical or chemical agent can be used for this process.

- **Embedding** – This is a delicate process that requires precision and care. The tissue, which needs to be sectioned into thin sections, is placed in a medium that will solidify around it, creating a block that is hard enough to be sectioned. This medium could be paraffin (a type of wax) or, depending on the purpose, Optimal Cutting Temperature (Optimal Cutting Temperature compound (OCT)) compound.

- **Sectioning** – This involves using a device that allows the block to be sectioned into thin sections. Different devices are available according to the technique. Normally, for paraffin sections, the biomedical scientist uses a microtome and produces 3 μm thick sections. These sections are then placed on a slide for later viewing by the pathologist.

- **Staining** – Plays a vital role in histology, as most of the cells and cellular elements are transparent. Staining the tissue is crucial, as it enhances visibility and improves the quality of histological analysis. According to the tissue findings, many staining techniques are used. Haematoxylin and Eosin (H&E) is the main stain used routinely and provides a general view of tissue architecture and cell morphology (Bancroft & Gamble, 2008).

In this study, tissue samples were collected from murine models subjected to a high-fat diet. The entire process, from sample collection to slide preparation, was designed to replicate the conditions commonly encountered in biopsies collected during the diagnostic process. The samples were processed following the methodology detailed in this thesis, subsequently undergoing microscopic evaluation and scoring by a pathologist. These same samples were then imaged and analysed using the tested plugins.

2.4 Semi-quantitative Pathologist assessment

Steatosis is routinely evaluated in liver biopsies because it is valuable for establishing accurate diagnosis, prognosis, and treatment planning. The current gold standard for liver steatosis quantification is a semi-quantitative method obtained by pathologists, a rapid assessment that does not require any special equipment. Microscopically, the pathologists evaluate the extent of fat accumulation in cells, and the presence of inflammation, fibrosis, and ballooning degeneration in the liver by a histologic evaluation of the haematoxylin and eosin staining slides. The hepatic steatosis is characterized semi-quantitatively (percentage of hepatocytes containing lipid droplets) by the Pathologist.

This method involves a standardized score system among pathologists to grade steatosis, ensuring consistency and accuracy in diagnosis and research. Two grading systems are commonly used to stage and grade the severity of NAFLD: the NAFLD NAS and the Brunt Score System. However, they focus on different aspects of the disease and are used in distinct contexts.

- **NAS** comprises 14 histological features, 4 of which are evaluated semi-quantitatively, including grading steatosis (0–3), lobular inflammation (degree of inflammatory infiltration in liver tissue, 0–2), hepatocellular ballooning (liver cell injury characterized by ballooning of hepatocytes, 0–2), and fibrosis (0–4). Another nine features are recorded as present or absent (Kleiner et al., 2005).

  Steatosis is graded as follows:

  - Grade 0 (healthy): <5% of hepatocytes contain fat.
  - Grade 1 (mild): 5% to 33% of hepatocytes contain fat.
  - Grade 2 (moderate): 34% to 66% of hepatocytes contain fat.
  - Grade 3 (severe): >66% of hepatocytes contain fat.

  This grading provides a clear, semi-quantitative measure of the extent of steatosis. In **Appendix A**, there is a table with the comparison of the essential elements of NAFLD/NASH Grading and Staging Systems.

- **Brunt Score System for NAFLD** grades steatosis similarly. It provides a comprehensive assessment based on a combination of key histological lesions: steatosis, ballooning, and intra-acinar and portal inflammation. This score was developed to reflect both the location and extent of fibrosis. The fibrosis score is derived from the extent of zone 3 perisinusoidal fibrosis, with possible additional portal/periportal fibrosis and architectural remodeling. Fibrosis stages are as follows:

- Stage 1: Zone 3 perisinusoidal fibrosis.
- Stage 2: As above with portal fibrosis.
- Stage 3: As above with bridging fibrosis.
- Stage 4: Cirrhosis.

It is also used explicitly for grading and staging non-alcoholic steatohepatitis (NASH) according to the level of inflammation and liver injury (Kleiner et al., 2005).

The Brunt score categorizes liver disease into stages (from normal liver to cirrhosis) by assigning a numerical grade to each parameter (Brunt & Tiniakos, 2010; Brunt et al., 1999). The stages include:

- Grade 0: No NAFLD (normal liver).
- Grade 1: Mild NAFLD.
- Grade 2: Moderate NAFLD.
- Grade 3: Severe NAFLD.
- Grade 4: Cirrhosis (end-stage liver disease).

The distinction between the scoring systems lies in their focus and application: the NAFLD Brunt Score emphasizes the assessment of liver fibrosis and overall damage, aiding in classifying the severity of the disease from mild to cirrhosis. In contrast, the NAFLD NAS concentrates on disease activity and injury levels, evaluating inflammation, steatosis, and ballooning degeneration. NAS is commonly used in clinical trials and is valuable for tracking disease progression or response to treatment. Together, these scoring systems complement each other by offering a comprehensive evaluation of liver health and disease stage, helping to determine the need for therapeutic intervention (Chalasani et al., 2018; Sanyal et al., 2011).

Unfortunately, the consistency in such assessment remains imprecise and subject to intra-observer and inter-observer variations (Gawrieh et al., 2011). Another issue is that some hepatocytes are too small, numerous, and indistinctive to be counted at low magnification, while overestimation of steatosis tends to increase with progressing severity. Also, each score has a wide percentage range of steatosis see Figure 2.10. Therefore, accurate and reproducible identification of the hepatic steatosis percentage is paramount in determining response to therapy.

With new and more advanced techniques, there is significant potential for improvement in this area, offering hope for a more precise and reliable assessment method in the future.

2.5 Quantitative Methods, FIJI Is Just ImageJ

It is important to note that the semi-quantitative method, while widely used, is not without limitations. It relies on a score that is susceptible to interobserver and intra-observer variability, making it prone to inaccuracy. Automatic methods for hepatic steatosis quantification present a compelling alternative. They offer an objective assessment of steatosis, potentially reducing human bias, increasing accuracy, and reliability, and providing continuous grading. These advantages can significantly enhance the perception of clinicians and researchers in their respective applications. Moreover, the use of quantitative methods allows for the generalization, and homogenization of findings to a larger population, a crucial step in making broader inferences and predictions.

Bio-image analysis is:

- **Objective** – which reduces the influence of personal biases and opinions. This ensures that the findings are based on measurable and observable features;
- **Reproducible** – quantitative research follows systematic procedures, making it easier for other researchers to replicate studies and verify results. This enhances the credibility and reliability of the results;
- **Reliable** – on precision and accuracy because these methods allow for precise data measurement and analysis. Statistical tools can accurately quantify relationships, differences, and effects, leading to more reliable, evidence-based research ((Kemmer et al., 2023)).
- **Colour** – In colour digital images, each pixel has 3 values corresponding to Red Green and Blue (RGB), which vary between 0-255, and the combination of the 3 gives the final colour in the image;
- **Resolution** – This is dependent on how the image was acquired, magnification, and camera specifications. Resolution is the ability to distinguish between two different objects in an image; therefore, higher resolution allows for the identification of smaller structures. This is limited by physics to around 350 nm.

A pixel or "picture element" is the smallest unit of a digital image, its value storing information about the original sample, overall light intensity for black and white images or the red, green, and blue RGB light values in colour images. Image resolution can also affect analysis.

FIJI is a widely adopted software platform specifically developed for the analysis of biological images. Given its critical role in the quantitative analyses conducted in this study, a detailed introduction will be provided in the subsequent subchapter.

FIJI is a powerful and versatile open-source software package specifically developed to enhane the capabilities of ImageJ2 for scientific image processing and analysis. As an integrated platform, FIJI provides a comprehensive suite of tools designed to facilitate a wide range of tasks, from basic image adjustments, such as brightness and contrast optimization, to advanced functionalities including microscopy image analysis, three-dimensional (3D) reconstruction, and quantitative measurements (Pietzsch et al., 2015; Schindelin et al., 2012; Schmid & Mair, 2014).

One of the distinguishing features of FIJI is its ability to utilize macros—scripts or automated workflows that enable users to streamline repetitive tasks and standardize analyses. These macros are not only highly efficient but also inherently shareable, fostering collaboration and reproducibility among researchers (Huisman & Schutte, 2017).

FIJI is widely adopted within the scientific community, particularly in disciplines requiring sophisticated analysis of biological images and other research data. It has become an indispensable tool for researchers working with complex datasets derived from techniques such as fluorescence microscopy, electron microscopy, and other imaging modalities (Cai & Zhou, 2016). The platform's extensibility, user-friendly interface, and community-driven development further enhance its appeal, making it a cornerstone in modern scientific image analysis workflows (Schindelin et al., 2012).

It is possible to develop or utilize open-source plugins that enhance and customize image analysis to meet the specific requirements of each experiment or project. These plugins provide flexibility, adaptability, and an efficient approach to handling diverse experimental conditions and objectives.

In this project, two specific plugins (Saturation and Weka) were selected for evaluation and testing. The selection was based on several factors: their popularity both within the institute and in the broader scientific community, the availability of personnel with expertise in their use, and their reputation as user-friendly tools that facilitate efficient analysis. Additionally, these plugins are supported by an extensive array of online resources, including documentation, tutorials, and active user forums, which make them accessible and practical for researchers at various levels of proficiency. This combination of features ensures that the chosen plugins align well with the project's goals, providing both reliability and adaptability in image analysis tasks.

2.5.1 Saturation

Imaging technologies that collect data from images and perform more detailed analyses are transforming the world of bioscience. Techniques such as noise reduction and removal, image segmentation, and filtering are some of the strongest tools that FIJI can perform. In FIJI, Intensity Saturation, Clipping, or Over Exposure refer to situations where the pixel values in an image exceed or are capped at certain limits, leading to a loss of detail in those regions.

Intensity saturation works through threshold segmentation. The image histogram, which shows the distribution of pixel intensities/values, is crucial to understanding intensity saturation.

The binary thresholding selects a value to divide the image into two subsets of pixels. The algorithm's role is to calculate where to place this dividing line for a given image and, essentially, which intensity level to use as a decision point. When pixel values in an image reach the maximum or minimum limits of the image's dynamic range, some areas appear completely white (for maximum saturation) or completely black (for minimum saturation). This can lead to the loss of details in those regions. The image is then split into 'foreground' or 'background.' An 8-bit scale image has pixel intensity values ranging from 0 (black) to 255 (white), with any values that hit 255 being fully saturated. Whether this step is done with a resource to an algorithm or manually, it is important to note that applying a filter before segmentation and post-processing can significantly improve the result.

2.5.2 Waikato Environment for Knowledge Analysis

The trainable Weka Segmentation is a powerful FIJI plugin that combines machine-learning algorithms with selected image features to produce pixel-based segmentations. Commonly used techniques in image segmentation include supervised learning algorithms such as Random Forest, Support Vector Machines (SVM), and more recently, deep learning methods (e.g., convolutional neural networks). However, in the context of Weka, Random Forest remains one of the most widely adopted approaches due to its robustness and ease of implementation in biomedical image analysis (Arganda-Carreras et al., 2017).

The segmentation process is typically supervised, meaning that the user provides annotated examples of each class (e.g., vacuole vs. background), and the algorithm learns to classify pixels accordingly. Weka simplifies this by allowing the user to create and train a classifier interactively, refining the model based on visual feedback. The training process involves calculating image features (such as mean, median, variance, and kurtosis, among others), which are then used by the algorithm to distinguish between structures of interest. For example, the Random Forest classifier, a popular ensemble learning method, works by constructing multiple decision trees from randomly selected subsets of features and samples.

Each tree makes a prediction, and the final output is determined by majority vote, increasing accuracy and reducing overfitting (Breiman, 2001).

In addition to supervised learning, the plugin also offers unsupervised segmentation techniques (e.g., clustering) for exploratory analysis or cases where labelled data is not available. Its user-friendly interface and integration with FIJI make it a valuable tool for histological image analysis, as illustrated in Figure2.8 and Figure2.9.

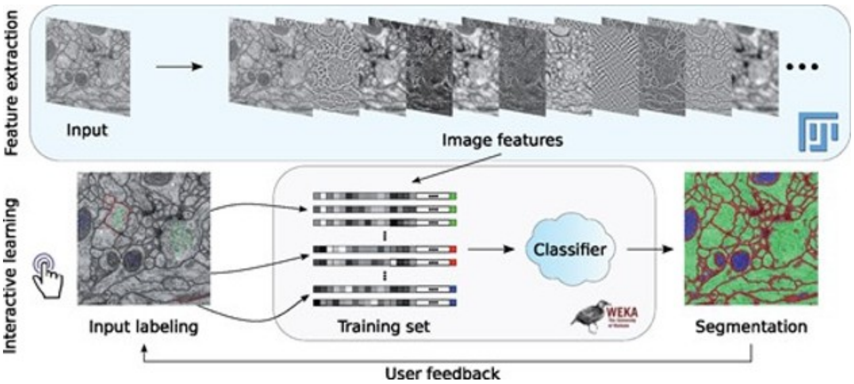


Figure 2.8: TWS pipeline for pixel classification. Image features are extracted from an input image using FIJI-native methods. Next, a set of pixel samples is defined and represented as feature vectors, and a Weka learning scheme is trained on those samples and finally applied to classify the remaining image data. Adapted from (Arganda-Carreras et al., 2017)

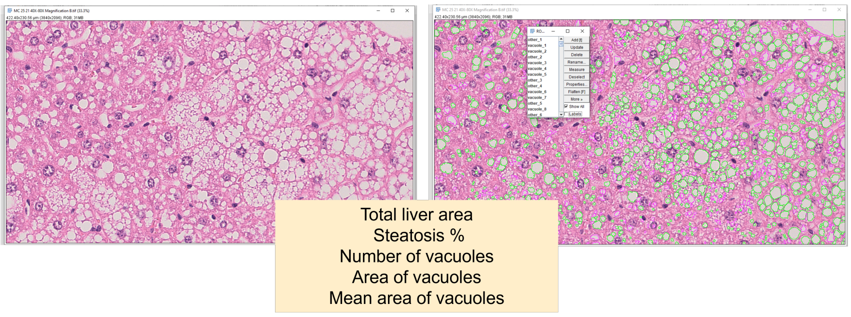


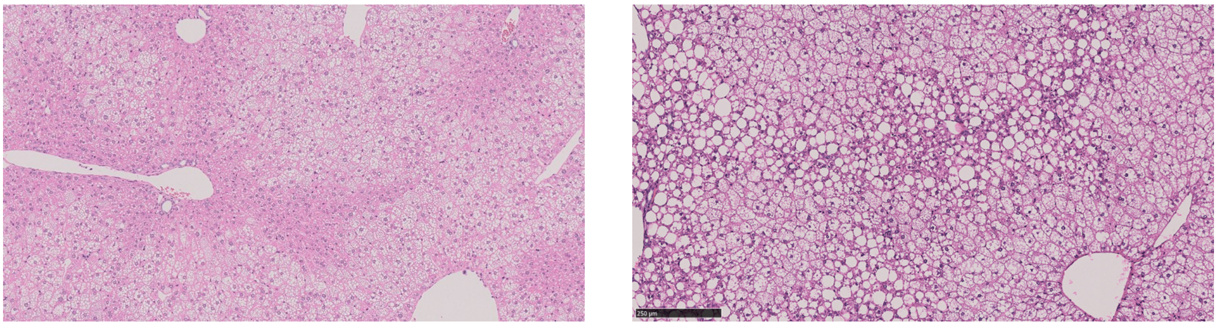
Figure 2.9: Example of an H&E-stained liver image processed using Weka (40× magnification, 80× resolution). Detected vacuoles are highlighted with green.

2.5.3 Illustrative Example of Visual vs. Automated Steatosis Assessment

To illustrate the limitations of visual scoring systems, consider the two histological images in Figure2.10. At first glance, the tissue samples appear substantially different in terms of fat content. Nevertheless, both were assigned the same score—grade 3 (i.e., $>$ 66% steatosis)— by a trained pathologist, based on semi-quantitative assessment.

Specifically, Figure2.10a was estimated at approximately 70% steatosis, and Figure2.10b at approximately 100%, highlighting the wide range of variability within each score category. This reinforces the concern that substantially different degrees of steatosis may be grouped under the same classification, potentially reducing diagnostic precision.

When analysed with quantitative methods—using the Saturation and Weka plugins— steatosis levels were calculated as 16% and 24%, respectively. This discrepancy underscores the potential of automated image analysis to detect finer differences that may be missed in conventional visual scoring.



(a) Figure A.

(b) Figure B.

Figure 2.10: H&E liver images at 10x Magnification

Not only is it important to standardize and establish a reliable method for steatosis assessment, but the combination of magnification and resolution also plays a crucial role in the accuracy and reproducibility of that evaluation, particularly when using automated image analysis techniques. Inconsistent or suboptimal imaging parameters can significantly affect the detection and quantification of lipid droplets, leading to variability in results and potentially compromising the reliability of the analysis. The magnification may enhance the visibility of finer histological details, while appropriate resolution ensures that the image retains sufficient clarity for accurate segmentation and feature extraction. As illustrated in Figure 2.11, the choice of magnification and resolution directly influences the performance of automated systems by determining the quality of input data used for classification or measurement tasks. Therefore, selecting optimal imaging settings is essential to ensure robust and consistent steatosis assessment.

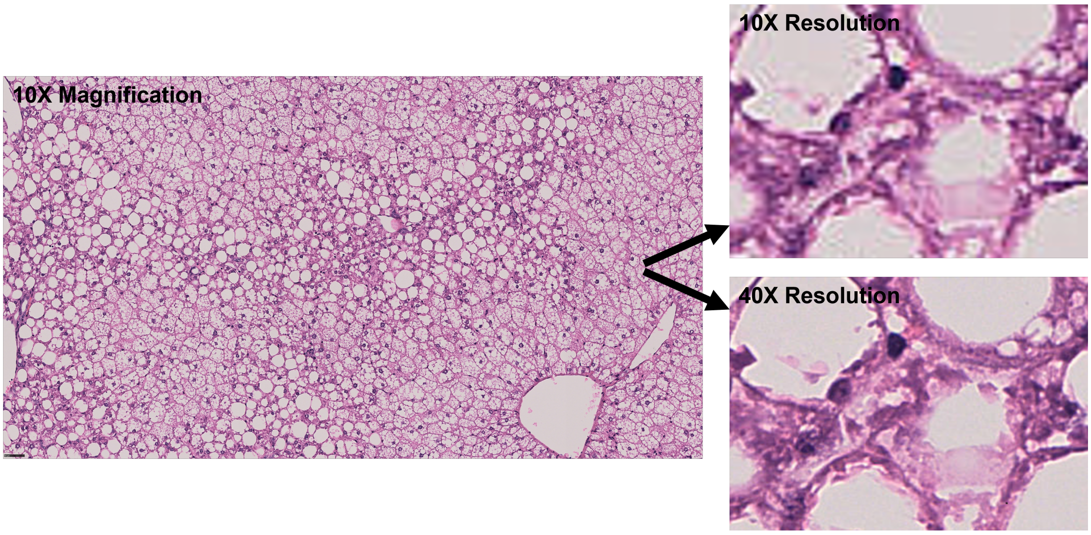


Figure 2.11: H&E liver image, Magnification and Resolution

Chapter 3

Study Design and Description

This chapter provides a detailed presentation of the data utilized in this thesis. A comprehensive understanding of the data is essential to contextualize the application of the statistical methods discussed. By exploring the characteristics and structure of the dataset, this chapter aims to establish a foundation for how these methods are tailored to the data's specific requirements and challenges. This experimental study was conducted in collaboration with the Histopathology Facility and the Advanced Imaging Unit at the Gulbenkian Institute for Molecular Medicine. The dataset analyzed in this study was provided by the Histopathology Facility and consists of 17 mouse liver samples. A schematic representation of the workflow study design as outlined in Figure 3.1.

Sample Selection A total of seventeen paraffin-embedded liver tissue blocks were obtained from the Histopathology Facility Biobank. Selection was carried out by a Veterinary Pathologist based on the inclusion criteria of confirmed fatty liver induced by a hypercaloric diet. To ensure a representation of different stages of steatosis, the samples were pre-classified using the semi-quantitative score. These tissue blocks were obtained from previous experiments conducted at the Institute, all approved by the Directorate-General for Food and Veterinary. Importantly, no animals were specifically tested for this study. Permission to use the slides was obtained from the original researchers responsible for each experiment. Furthermore, no additional information about the mice, such as demographic characteristics, was available.

Sample Preparation and Imaging All liver samples were sectioned at a thickness of 3 µm and stained using Haematoxylin and Eosin (H&E). The stained slides were scanned using a NanoZoomer-SQ Digital Slide Scanner (Hamamatsu Photonics) at 40x magnification. The resulting digital images were reviewed using NDP.view 2 software, and images were saved at multiple magnifications (10x, 20x, 40x), and export with different resolutions (10x, 20x, 40x, and 80x) for detailed analysis.

Image Analysis and Plugin Application For each scanned image, random regions of interest were selected for further examination. Two image analysis plugins—Weka and Saturation—were applied to each selected region using FIJI software. Custom macros were developed for both plugins by the technician José Serrado Marques from the Advanced Imaging Unit, based on the methods described by Munsterman et al., 2019. The macros enabled the separation of steatotic hepatocytes, quantify vacuoles in liver, from other similar objects using a statistical classifier based on logistic regression, available at Marques, 2021. The study by Munsterman et al. (2019) utilized a logistic regression model to classify segmented objects as either vacuoles or non-vacuoles, based on their morphological features. In this context, the response variable was the probability of an object being a vacuole (binary outcome: vacuole = 1, non-vacuole = 0), and the explanatory variables were four shape descriptors:size (in μm^2), circularity, roundness, and solidity. The resulting classifier was defined as:

$$\text{logit}(p) = -16.2 + 0.00272 \times \text{size}\,\mu m^2 + 5.81 \times \text{circularity} + 7.054 \times \text{roundness} + 10.3 \times \text{solidity}$$

where p represents the probability that the object is classified as a vacuole.
To refine vacuole detection, both the logit function and a size threshold were implemented into the macros, using the parameters derived from Munsterman et al. (2019).

Assessment and Scoring The semi-quantitative scoring was performed by a Veterinarian Pathologist, who reviewed the outputs of the Weka and Saturation plugins. The pathologist evaluated the results based on the degree of steatosis observed in each image, ensuring consistent classification of fatty liver severity. For each plugin, the macros were programmed to identify and count vacuoles, measure their area, and extract shape descriptors. These features were summed to provide a comprehensive profile of vacuole morphology within the tissue samples. All measurements were compiled into a table format for post-analysis.

This design enabled a systematic assessment of fatty liver samples while optimizing automated image analysis techniques for histopathological research. It established standardized classification criteria, ensuring consistent measurement outputs across all analysed samples.

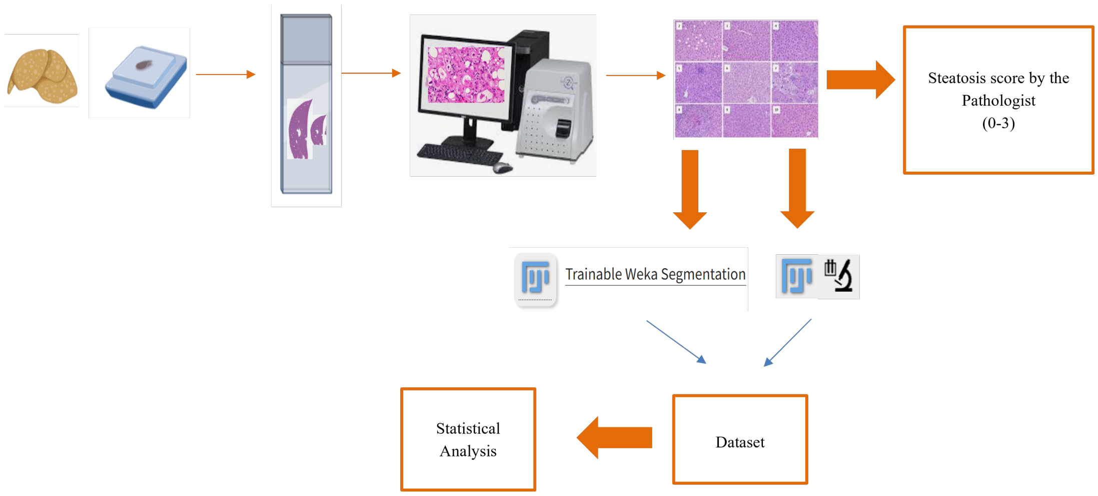

Steatosis score by the Pathologist (0-3)
Trainable Weka Segmentation
Statistical Analysis
Dataset

Figure 3.1: Study Design

3.1 Ethical Aspects Study

This study was approved by the Ethical Committees of the involved institutions and does not require approval by the National Committee for Data Protection. It was done in collaboration with the Histopathology Facility of *Gulbenkian Institute for Molecular Medicine*, which shared the database for this study.

3.2 Objectives

The primary purpose is to evaluate an automated quantification method for hepatic steatosis in liver histological images. This method can significantly impact the field by providing a more accurate assessment of hepatic steatosis based on the percentage of the vacuole's area and size.

To achieve this main goal, the following objectives were defined:

- Using the variables from the database, conduct a comprehensive evaluation of the performance of two distinct types of image plugins, Weka and Saturation, both from FIJI. This thorough analysis will provide a robust foundation for our research.

- Define the adequate magnification and resolution for vacuole identification, allowing the plugins to identify the vacuoles precisely. It is important to consider that multiple combinations of magnification and resolution may exist. Additionally, the time each plugin takes to generate results should also be taken into account.

- Determine the ideal number of histological images per liver relevant to the steatosis percentage assessment. To optimize the analysis, it is important to define the minimum number of images required. However, if processing more images enhances accuracy depending on the steatosis stage, increasing the number of images should be considered. In this case, it is also essential to evaluate the time each plugin takes to process each image. Balancing the number of images with processing time is crucial for achieving both accurate and efficient results.

- Analyse the difference between the area of the vacuoles and the percentage and the steatosis score. It might be interesting to verify if there is a vacuole pattern according to steatosis degree.

3.3 Variables Description and Outcomes

> *"It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."*
>
> – A. Conan Doyle, *A Scandal in Bohemia*

This section provides an overview of the variables included in the database compiled by Mafalda Casanova, along with the reorganized variables used in this study. The description takes into account the paired nature of the data, highlights the categories excluded due to missing data (10x80), and concludes with a summary table describing the database used in the analysis. See tables 3.1, and 3.2.

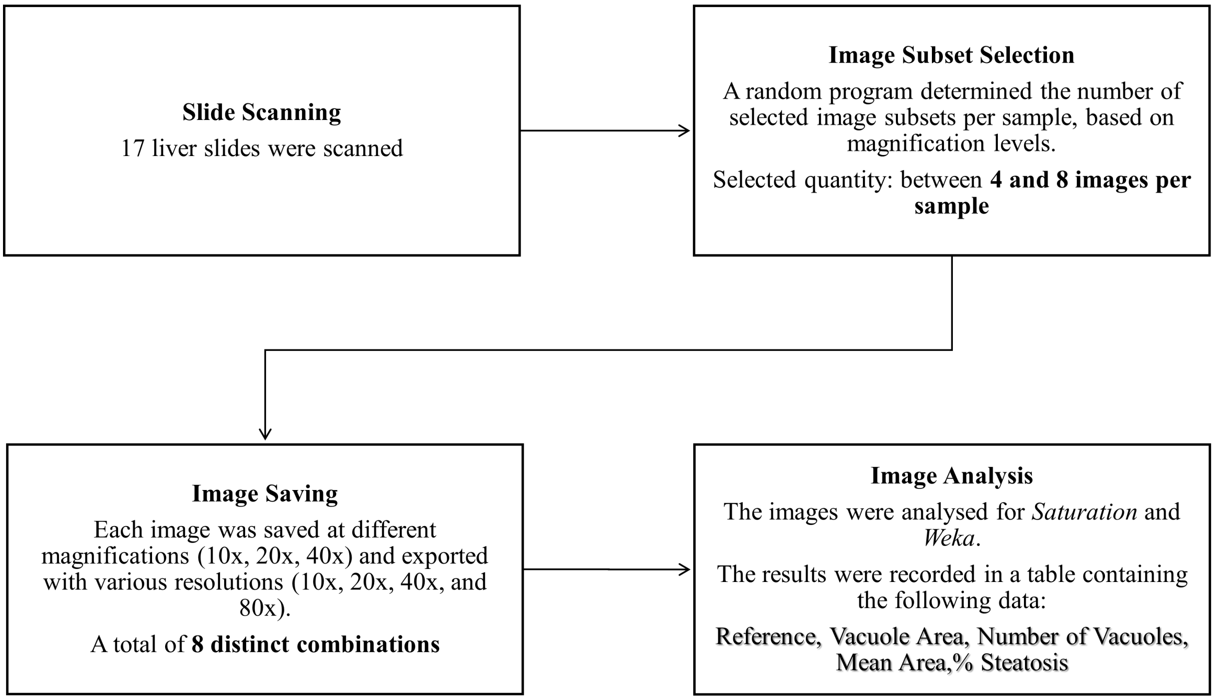

Slide Scanning
17 liver slides were scanned
Image Subset Selection
A random program determined the number of selected image subsets per sample, based on magnification levels.
Selected quantity: between 4 and 8 images per sample
Image Saving
Each image was saved at different magnifications (10x, 20x, 40x) and exported with various resolutions (10x, 20x, 40x, and 80x).
A total of 8 distinct combinations
Image Analysis
The images were analysed for Saturation and Weka.
The results were recorded in a table containing the following data:
Reference, Vacuole Area, Number of Vacuoles, Mean Area,% Steatosis

Figure 3.2: Description of the data collected for this study

Original Database Variables:

- **Reference**: A unique code assigned to each mouse liver sample to maintain anonymity **(Qualitative Nominal Variable)**.
- **Area of Vacuoles**: Represents the total area occupied by all vacuoles in the corresponding image **(Quantitative Continuous Variable)**.
- **Vacuole_nr**: The count of vacuoles present in a given image **(Quantitative Discrete Variable)**.
- **Mean_area**: The average area of vacuoles, calculated as *Area of Vacuoles / Vacuole_nr* **(Quantitative Continuous Variable)**.
- **Percentage_steatosis**: Indicates the percentage of liver steatosis, calculated as *Area of Vacuoles / Total Liver Area* **(Quantitative Continuous Variable)**.
- **Steatosis Score**: A semi-quantitative score assigned by the pathologist to indicate the extent of steatosis **(Qualitative Nominal Variable)**.

New Database Variables:

- **Code**: A combination of the liver sample number and a letter representing the order of images within the same liver sample. Constructed by merging two variables: *specimen_nr* and *specimen_letter*. The values range from "1A" to "17AB" **(Qualitative Nominal Variable)**.
- **Vacuole_nr_mean**: The average number of vacuoles per liver sample across all images **(Quantitative Continuous Variable)**.
- **Magnification**: Refers to the magnification level used during image acquisition (10x, 20x, 40x) **(Qualitative Nominal Variable)**.

- **Resolution**: Indicates the image resolution, recorded as 10x, 20x, 40x, or 80x (**Qualitative Nominal Variable**).
- **Ar**: A combined variable representing Magnification × Resolution to capture different image scaling levels (**Qualitative Nominal Variable**).

Table 3.1: Database variables

Code	Value
Specimen_nr	1 – 17
Specimen_letter	A – X
*Vacuoles_area**	$\mathbb{R}^+$
Vacuole_nr	$\mathbb{N}$
*Mean_area**	$\mathbb{R}^+$
*Percentage_steatosis**	$[0, 100]$
*Vacuole_nr_mean**	$\mathbb{R}^+$
Steatosis_score	1, 2, 3, 4
Program	Weka, Saturation
Magnification	10, 20, 40
Resolution	10, 20, 40, 80
AR	10x10, 10x20, 10x40, 20x20, 20x40, 20x80, 40x40, 40x80

*Quantitative variable

For the variable percentage of steatosis, values were initially expressed as percentages within the interval [0, 100]. In some parts of this work, these values were converted to a [0, 1] scale for standardization and modeling purposes.

Table 3.2: Number of Images per Sample and Magnification

Specimen_nr	*Magnification*	*Image subset_nr*	*Specimen_nr*	*Magnification*	*Image subset_nr*
1	10x	6	9	10x	5
	20x	5		20x	6
	40x	8		40x	6
2	10x	4	10	10x	6
	20x	4		20x	5
	40x	6		40x	5
3	10x	7	11	10x	6
	20x	6		20x	7
	40x	6		40x	7
4	10x	6	12	10x	6
	20x	6		20x	7
	40x	6		40x	7
5	10x	5	13	10x	5
	20x	6		20x	5
	40x	6		40x	5
6	10x	5	14	10x	5
	20x	5		20x	6
	40x	5		40x	6
7	10x	5	15	10x	5
	20x	5		20x	6
	40x	5		40x	5
8	10x	6	16	10x	6
	20x	6		20x	6
	40x	5		40x	6
			17	10x	5
				20x	6
				40x	6

Chapter 4

Statistical Methodology

To quantify liver steatosis using two automatic image processing plugins, Weka and Saturation, and to analyse their performance, specifically in terms of magnification and resolution, the number of liver images required, and the distribution patterns of the vacuole sizes, a methodological plan was structured into three distinct phases, each corresponding to a specific objective.

Initially, Spearman's rank correlation was employed to assess the strength and direction of the monotonic relationship between the differences in the mean number of vacuoles detected by the Weka and Saturation plugins (Conover, 1980), given its suitability for both ordinal and numerical variables, as well as its robustness to outliers and deviations from normality. This project aims to evaluate and compare the performance of both plugins. To further investigate this objective, a Wilcoxon signed-rank test was performed on the paired differences in the mean number of vacuoles to determine whether statistically significant differences exist between the two plugins while considering the combination of magnification and resolution. This non-parametric test was chosen due to the non-normal distribution of the paired differences, as verified by visual inspection (histograms and Q-Q plots) and confirmed by the Shapiro-Wilk test, which violated the assumptions required for the paired t-test (Razali & Wah, 2011). Subsequently, a series of univariable GLMM were fitted, each including a single independent variable—namely, the program (Weka or Saturation), magnification, or resolution-to assess their individual associations with the number of vacuoles. These models, assuming a Negative Binomial distribution to account for overdispersion, included a random effect to account for the repeated measures structure of the data (Zuur et al., 2009). This approach allowed for a consistent modelling framework while isolating the effect of each covariate individually and identifying the most influential variables, which were then considered for inclusion in the construction of the final multivariate model. Based on these results, a multivariate analysis was performed using a GLMM with a Negative Binomial distribution to identify the factors contributing to variations in vacuole counts.

Building upon the findings from the first phase, a combinatorial analysis was carried out to find the minimum number of liver images required to achieve a reliable and stable assessment of hepatic steatosis. A bootstrap analysis was applied to derive confidence intervals, thereby providing a more robust and accurate estimate of the variability in the image analysis results.

Finally, the third phase focused on determining whether the distribution of vacuole sizes follows a specific morphological pattern and whether the percentage of hepatic steatosis influences this distribution. For this purpose, vacuole areas were compared between two groups — mild and severe steatosis — to assess potential differences in morphological characteristics.

4.1 Spearman's Rank Correlation

When the analysis involves studying two random variables simultaneously and examining the relationship between them, it is referred to as bivariate analysis, often represented by the pair (X,Y).

The concepts of correlation and regression were first developed by Francis Galton in Natural Inheritance (1889). Building on Galton's work, Spearman (1904) introduced the ordinal correlation coefficient, denoted by r_s, based on the use of *ranks* instead of raw values (Velosa & Pestana, 2008).

Unlike the Pearson correlation, which measures the linear relationship between variables using raw data, the Spearman coefficient assesses the strength and direction of a monotonic association by comparing the ranks of the data.

The Spearman rank correlation coefficient will be calculated in Chapter 5.2.1, where the relationship between the differences in the mean number of vacuoles for both plugins is analysed.

To compute the Spearman correlation coefficient, the values of the two variables are first converted to ranks. In the presence of ties, all tied values are assigned the average of the ranks they would occupy if there were no ties.

Given two random variables, X and Y, and a sample of size n, the raw values x_i and y_i (for $i = 1,\ldots,n$) are replaced by their respective ranks. Let v_i be the rank of x_i in the sample of x's, and s_i be the rank of y_i in the sample of y's. Let $\bar{v}$ and $\bar{s}$ denote the mean ranks of $(v_1,\ldots,v_n)$ and $(s_1,\ldots,s_n)$, respectively:

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_i \quad \text{and} \quad \bar{s} = \frac{1}{n}\sum_{i=1}^{n} s_i$$

The Spearman correlation coefficient is then given by:

$$r_s = \frac{\sum_{i=1}^{n}(v_i-\bar{v})(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{n}(v_i-\bar{v})^2\sum_{i=1}^{n}(s_i-\bar{s})^2}}$$

When there are no tied values in the data, a computationally simpler formula can be used. Let d_i denote the difference between the ranks of the i-th individual or item in the two orderings. The Spearman rank correlation coefficient r_s is given by:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where d_i represents the difference between the ranks of corresponding values of X and Y, that is, $d_i = v_i - s_i$.

The coefficient r_s ranges from -1 to 1, where:

- $r_s = 1$ indicates a perfect positive correlation,
- $r_s = -1$ indicates a perfect negative correlation (i.e., the variables are exactly inversely related),
- $r_s = 0$ suggests no correlation.

The magnitude of r_s indicates the strength of the relationship, with values closer to ±1 reflecting a stronger association. Values near 0 indicate a weak or non-existent relationship between both variables (Murteira, 2024; Velosa & Pestana, 2008).

Consider a sample of bivariate data (X_i,Y_i), for $i = 1,\ldots,n$, drawn from a continuous population X,Y. Let V_i = rank of X_i in the sample of X_i's and S_i = rank of Y_i in the sample of Y_i's. Assuming that X and Y have a continuous distribution, there are no ties. Let $D_i = V_i - S_i$.

To assess whether the Spearman correlation is significantly different from zero, the following hypotheses are tested:

H_0: There is no monotonic association between the variables X and Y; that is, $\rho = 0$.
Depending on the context and the expected direction of association, different alternative hypotheses can be formulated. Table 4.1 summarizes the possible forms of H_A used in the Spearman rank correlation test.

Table 4.1: Types of alternative hypotheses for Spearman's rank correlation

H_A	**Description**
$\rho \neq 0$	Two-tailed (any monotonic correlation)
$\rho > 0$	One-tailed (positive monotonic correlation)
$\rho < 0$	One-tailed (negative monotonic correlation)

The sample Spearman rank correlation coefficient is denoted by r_s, while R_s denotes the corresponding random variable used in the theoretical distribution under the null hypothesis. The Spearman rank correlation coefficient for the random variable (R_s) is then:

$$R_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$$

According to Conover, 1980, under the null hypothesis H_0, the expected value of the Spearman correlation coefficient is, $E(R_s) = 0$ and $\text{var}(R_s) = \frac{1}{n-1}$.

Then, for large n, the distribution of $\sqrt{n-1}R_s$ can be approximated by the distribution $\text{N}(0,1)$ (Daniel, 2005). For $n \leq 30$, there are tables of critical values obtained from the exact distribution of R_s.

The decision whether to reject H_0 is based on the p-value, calculated according to the type of alternative hypothesis being tested. Table 4.2 summarizes the formulas used to compute these p-values under the normal approximation.

Table 4.2: p-values for alternative hypotheses using the normal approximation

H_A	p-value
Any monotonic association ($\rho \neq 0$)	$2 \min\{P(Z \leq z), P(Z \geq z)\}$
Positive monotonic association ($\rho > 0$)	$P(Z \geq z)$
Negative monotonic association ($\rho < 0$)	$P(Z \leq z)$

where z is the observed value of the test statistic

$$Z = \sqrt{n-1}R_s, \quad Z \sim \text{N}(0,1).$$

4.2 Wilcoxon Signed-Rank Test

The Wilcoxon test is a non-parametric statistical method used to assess whether there is a significant difference between the central tendencies of paired samples or one sample. When data are paired, the observations are inherently related, which is a key consideration in applying this test. Developed by

Frank Wilcoxon (1945), this test is particularly valuable for analysing continuous or ordinal data that are non-normally distributed. It is robust to data skewness, the presence of outliers, and is suitable for small sample sizes (e.g., n <30). The Wilcoxon test operates on ranked data rather than raw values, making it less sensitive to extreme outliers that could otherwise distort results in parametric analyses.

Let X and Y represent two variables from which random samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ are drawn, respectively.

In symmetric populations, where the mean is equal to the median (Daniel, 2005), the Wilcoxon test indirectly assesses the mean value by testing the equality of the medians. In such cases, the null hypothesis (H_0) posits that the median of the differences between paired observations is zero, indicating no difference between the two samples. Conversely, the alternative hypothesis (H_A) states that the median of these differences is not zero.

Formally, the hypotheses for the Wilcoxon signed-rank test can be stated as:

$$H_0 : \chi_{\frac{1}{2}} = x_0$$

Depending on the direction of the effect being tested, different forms of the alternative hypothesis may be considered. Table 4.3 summarizes the possible alternative hypotheses for the Wilcoxon signed-rank test.

Table 4.3: Alternative hypotheses for the Wilcoxon Signed-Rank Test.

H_0	**Description**
$\chi_{\frac{1}{2}} \neq x_0$	Two-tailed (the median is different from x_0)
$\chi_{\frac{1}{2}} > x_0$	One-tailed (the median is greater than x_0)
$\chi_{\frac{1}{2}} < x_0$	One-tailed (the median is less than x_0)

where $\chi_{\frac{1}{2}}$ denotes the population median.

The test compares the ranks of the differences between paired values, considering their magnitudes but ignoring their signs. By focusing on ranks rather than raw differences, the test evaluates whether there is a systematic shift in the distribution of differences between two conditions (Daniel, 2005; Murteira, 2024).

Let $d_k = x_k - y_k$, for $k = 1, \ldots, n$, be the difference between each pair of matched observations from the two samples.

The absolute values $|d_k|$ are then computed and sorted in ascending order. Each $|d_k|$ is assigned a rank $r^*(d_k)$, while retaining the original sign of d_k. If there are tied absolute differences, they are assigned the average of the ranks they would occupy if distinct (Velosa & Pestana, 2008).

Then, for large sample sizes, the distribution of the Wilcoxon signed-rank statistic:

$$W = \sum_{k=1}^{n} r^*(d_k)$$

can be approximated by a normal distribution with mean

$$\mu_W = \frac{n(n+1)}{4}$$

and standard deviation

$$\sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}.$$

The standardized test statistic is then given by:

$$Z = \frac{W - \mu_W}{\sigma_W}$$

under the null hypothesis H_0, $Z \sim \mathcal{N}(0,1)$ (Conover, 1980; Daniel, 2005). In the presence of ties or zero differences, corrections to σ_W may be required.

In this study, the same liver image was analysed using Weka and Saturation metrics. To assess the consistency of these evaluations, the Wilcoxon test will be performed and the results will be presented in Chapter 5.2.2.

4.3 Models

In various fields of research, particularly in health, problems are common, and the primary objective is to examine the relationship between variables. More specifically, the focus is often on analysing how one or more explanatory (independent) variables influence a variable of interest, known as the response (dependent) variable. To address such problems, a range of modelling techniques can be applied, with regression models being among the most widely used (McCullagh & Nelder, 1989).

Regression models were initially developed assuming the response variable followed a normal distribution. This framework dominated statistical modelling until the mid-20th century. However, it became evident that the normal linear model was not adequate for many practical situations, especially when dealing with non-linear relationships or non-normally distributed data. In response to these limitations, Nelder and Wedderburn, 1972 introduced Generalized Linear Model (GLM) — a unified framework that extends classical linear models to accommodate a broader class of response distributions and link functions, thus providing greater flexibility in modeling diverse types of data.

The GLM allows for estimating regression coefficients and quantifying the impact of independent variables on the dependent variable. It enables the assessment of both the direction and strength of the relationship, along with its statistical significance (McCullagh & Nelder, 1989).

According to McCullagh and Nelder, 1989, GLM includes a broad class of models where the response variable follows a distribution from the exponential family, encompassing models such as linear regression, logistic regression, Poisson regression, and others. Some common special cases of GLM include the following:

- Classic linear regression model;
- Logistic regression model;
- Poisson regression model;
- Analysis of variance and covariance model;
- Log-linear model for multidimensional contingency tables;
- Probit model for proportion studies, etc.

All of the models mentioned above share a linear regression structure and are characterized by the fact that the response variable follows a distribution within the exponential family, a class of distributions with specific mathematical properties. Despite their flexibility, GLM present certain limitations: they require the preservation of linearity in the predictor structure, assume that the response distribution belongs to the exponential family, and rely on the assumption of independence among observations. However, GLM are versatile in that both response and explanatory variables can be measured on nominal, ordinal, or continuous scales (Dobson & Barnett, 2008). The GLM framework has played a pivotal role in modern statistics due to its flexibility and broad applicability across disciplines (McCullagh & Nelder, 1989). Its popularity among both specialists and non-specialists stems from the wide range of statistical models it unifies and the increasing availability of user-friendly software for data analysis. However, applying a GLM requires that the response variable follows a distribution from the exponential family. Therefore, a formal definition of this family is presented next.

4.3.1 Exponential Family

A random variable Y is said to follow a distribution belonging to the exponential family (or exponential dispersion family) if its probability density function (for continuous distributions) or probability mass function (for discrete distributions) can be written in the general form:

$$f(y \mid \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\},$$

where θ and ϕ are scalar parameters, representing the canonical (location) and dispersion (scale) parameters, respectively, and $a(\cdot)$, $b(\cdot)$, and $c(\cdot,\cdot)$ are known real-valued functions.

In this formulation, θ is the canonical parameter (i.e., the natural parameter of the distribution), and ϕ is typically assumed to be known and constant across observations. In some contexts, particularly in linear models, corresponding to the residual variance. It is also assumed that $b(\cdot)$ is differentiable and that the support of the distribution does not depend on the parameters.

In many applications, the function $a(\cdot)$ takes the form $a(\phi) = \phi/w$, where w is a known positive constant called the weight, which may vary from observation to observation. In this case, the expression simplifies to:

$$f(y \mid \theta, \phi, \omega) = \exp\left\{ \frac{\omega}{\phi}(y\theta - b(\theta)) + c(y, \phi, \omega) \right\},$$

where ω is the weight associated with the observation.

Mean and Variance

To derive the mean and variance of a distribution in the exponential family, we begin by considering the log-likelihood of a single observation:

$$\ell(\theta; \phi, y) = \ln(f(y \mid \theta, \phi)).$$

The score function is a random variable, as a function of Y:

$$S(\theta) = \frac{\partial \ell(\theta; \phi, Y)}{\partial \theta}$$

For regular exponential families, the expected value of the score function is:

$$\mathbb{E}[S(\theta)] = 0,$$

and

$$\mathbb{E}\left[S^2(\theta)\right] = \mathbb{E}\left[\left(\frac{\partial \ell(\theta;\phi,Y)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta;\phi,Y)}{\partial \theta^2}\right].$$

Recalling the expression of the log-likelihood from the general form:

$$\ell(\theta;\phi,y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi),$$

we obtain the score function:

$$S(\theta) = \frac{Y - b'(\theta)}{a(\phi)} \quad \Rightarrow \quad \frac{\partial S(\theta)}{\partial \theta} = -\frac{b''(\theta)}{a(\phi)},$$

where $b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$ and $b''(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta^2}$.

Since $\mathbb{E}[S(\theta)] = 0$ and the score is given by $S(\theta) = \frac{Y - b'(\theta)}{a(\phi)}$, taking expectations gives:

$$\mathbb{E}(Y) = b'(\theta)$$

$$\text{Var}(Y) = a(\phi)b''(\theta)$$

Thus, the variance of Y can be expressed as the product of two components: $b''(\theta)$, which depends only on the canonical parameter θ (and therefore on the mean μ) and is typically referred to as the variance function $V(\mu)$; and $a(\phi)$, which depends solely on the dispersion parameter ϕ.

4.3.2 Generalized Linear Models

Considering a random variable Y, referred to as the response variable or dependent variable, and a vector $\mathbf{x} = (x_1, \ldots, x_k)^T$, where k is the number of explanatory variables under study, also known as covariates or independent variables, which are assumed to explain part of the variability of the response variable Y. This response variable Y can be continuous, discrete, or dichotomous, and the covariates, whether deterministic or stochastic, can also be continuous, discrete, ordinal qualitative, or dichotomous.

Assuming we have data in the form $(y_i, \mathbf{x}_i)$, for $i = 1, \ldots, n$, resulting from realizations of $(Y, \mathbf{x})$ in n individuals, where the components Y_i of the random vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ are independent. The matrix representation of the data would be:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \text{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

As previously mentioned, generalized linear models are an extension of the classical linear model.

$$\mathbf{Y} = Z\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where Z is a matrix of dimension $n \times p$ specifying the model (usually the covariate matrix X with a first unitary vector), associated with a parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and $\boldsymbol{\varepsilon}$ is a vector of random errors with distribution

$$\mathcal{N}_n(0, \sigma^2 I),$$

that is, a multivariate normal distribution of dimension n, with zero mean vector and covariance matrix $\sigma^2 I$.

These assumptions imply that the expected value of the response variable is a linear function of the covariates, that is,

$$\boldsymbol{E}(\boldsymbol{Y} \mid Z) = \boldsymbol{\mu} \quad \text{with} \quad \boldsymbol{\mu} = Z\boldsymbol{\beta}.$$

To simplify the transition from Linear Models (LM) to GLM, the GLM can be specified using three components, which are defined as:

- Random component: the random variables Y_i are independent, with a distribution belonging to the exponential family, and with $E(Y_i) = \mu_i$.
- Systematic component: A linear predictor η_i is defined as a combination of the explanatory variables:

$$\eta_i = \mathbf{z}_i^T \boldsymbol{\beta}$$

- Link function between the random and systematic components:

$$g(\mu_i) = \eta_i$$

GLM are obtained by extending the assumptions underlying LM in two main directions:

- The distribution of Y_i can be any distribution belonging to the exponential family, as previously defined;
- It allows for link functions other than the identity to relate the linear predictor η_i to the mean μ_i, that is,

$$\eta_i = g(\mu_i)$$

  where $g(\cdot)$ is a monotonic and differentiable function called the link function.

When $\eta_i = \mu_i$, the function $g(\cdot)$ is referred to as the canonical link function (Dobson & Barnett, 2008).

One of the most commonly used distributions within the exponential family for modelling count data is the Poisson distribution.

Poisson distribution

The Poisson distribution, denoted by $Y \sim \text{Po}(\lambda)$, is used to model count data. Such data typically represent the number of occurrences of some event in a defined time period or space, when the probability of an event occurring in a very small time (or space) is low, and the events occur independently (Velosa & Pestana, 2008).

The probability mass function of the discrete random variable Y is

$$P(Y = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \quad \text{and} \quad \lambda > 0\,,$$

where λ is the expected number of occurrences (i.e., $\mathbb{E}[Y] = \lambda$), and $e^{-\lambda}$ ensures that the total probability over all possible values of k sums to 1.

The Poisson distribution can be written in the exponential family form as:

$$f(y;\eta) = \exp\left(y\eta - e^{\eta} - \log y!\right),$$

where $\eta = \log\lambda$ is the natural (canonical) parameter of the exponential family representation, and the sufficient statistic is Y, (its realization is denoted y), since Y appears as the coefficient of the natural parameter η in the exponent.

If a random variable follows a Poisson distribution, its expected value and variance are equal. However, data may exhibit overdispersion when the variance exceeds the mean, $\text{Var}(Y_i) > \mathbb{E}(Y_i)$, whereas for the Poisson distribution, $\text{Var}(Y_i) = \mathbb{E}(Y_i)$ (Dobson & Barnett, 2008; Murteira, 2024).

The Negative Binomial distribution provides an alternative model that accommodates overdispersion.

Negative Binomial distribution

The Negative Binomial distribution is a discrete probability distribution that models the number of trials Y needed to achieve a specified number of successes k, given a constant probability of success p in each trial (Dobson & Barnett, 2008; Velosa & Pestana, 2008).

However, one limitation of the Poisson distribution is that it assumes the mean and variance of the response variable are equal. In practice, count data often exhibit overdispersion, where the variance exceeds the mean.

In such cases, the Negative Binomial distribution provides a more flexible alternative.

Let $Y \sim \text{Negative Binomial}(k,p)$. The probability mass function of Y is given by:

$$P(Y=n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad \text{for } n = k, k+1, \ldots$$

and the expected value and variance of Y are given by:

$$\mathbb{E}(Y) = \frac{k(1-p)}{p}, \qquad \text{Var}(Y) = \frac{k(1-p)}{p^2}$$

The Negative Binomial distribution belongs to the exponential family (McCullagh & Nelder, 1989). To demonstrate this, its probability mass function must be expressed in the canonical form of the exponential family.

In the parameterization commonly used in Generalized Linear Models, the Negative Binomial distribution is defined by the mean μ and a dispersion parameter $r > 0$, which reflects the degree of variability in the data beyond that accounted for by a Poisson model. Specifically, r can be interpreted as a shape or dispersion parameter that controls the amount of overdispersion (Hilbe, 2007).

Using the mean μ and the natural (canonical) parameter θ, we have:

$$\mu = \mathbb{E}(Y) = \tfrac{r(1-p)}{p}, \quad \theta = \log\left(\tfrac{\mu}{\mu+r}\right)$$

The probability mass function can be rewritten as:

$$f(y;\mu,r) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y$$

The logarithm of the probability mass function expressed in terms of μ and r is:

$$\log f(y;\mu,r) = y \cdot \log\left(\frac{\mu}{r+\mu}\right) + r \cdot \log\left(\frac{r}{r+\mu}\right) + \log\left(\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\right)$$

This structure fits the exponential family form:

$$f(y;\theta,r) = \exp\left(y \cdot \theta - b(\theta) + c(y)\right)$$

Where:

$$\theta = \log\left(\frac{\mu}{\mu+r}\right) \quad \text{(Canonical parameter)}$$
$$b(\theta) = -r \cdot \log(1-e^{\theta}) \quad \text{(Cumulant function)}$$
$$c(y) = \log\left(\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\right) \quad \text{(Base measure)}$$

In the case of the negative binomial distribution, the canonical link function is:

$$g(\mu_i) = \log\left(\frac{\mu_i}{\mu_i+r}\right)$$

This formulation allows the negative binomial distribution to be used within the GLM framework, providing a flexible approach for modeling overdispersed count data.

4.3.3 Generalized Linear Mixed Model

When the data exhibit a hierarchical structure, repeated measures, or when dealing with longitudinal data, a GLMM is appropriate, as it allows for the inclusion of random effects. In this study, a GLMM with a Negative Binomial distribution was used to model the count of vacuoles, (Y), the dependent variable, given the presence of overdispersion and the repeated-measures structure of the data. This approach allows for the appropriate modelling of count data with extra variability (overdispersion) and incorporates random effects to capture intra-group dependence not explained by fixed covariates (Velosa & Pestana, 2008).

In this parameterization, the variance is given by

$$\text{Var}(Y) = \mu + \frac{\mu^2}{r},$$

which allows the model to account for overdispersion.

Understanding the model: The fixed coefficients (β_j) represent the effect of the predictor variables on the response variable, on the logarithm of the expected value of the response.

The basic model formula is:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + b_i$$

Where:

- μ_i = expected mean of the response variable for observation i
- β_0 = intercept

- β_j = coefficient associated with the explanatory variable x_{ji}
- b_i = random effect associated with observation or group i

A logarithmic link function was used because it assures the positivity of the predicted values and allows for a multiplicative interpretation of the effects. This approach is appropriate in situations where proportional changes are more meaningful than absolute ones, and where the response variable is strictly positive. The estimated coefficients represent the multiplicative effect of the covariates on the expected value of the response variable. That is, a one-unit increase in a given covariate multiplies the expected value of Y by $\exp(\beta_j)$, holding the other variables constant.

The main assumptions of the GLMM are:

- The response variable follows a distribution from the exponential family — in this case, a Negative Binomial distribution.
- The link function correctly specifies the relationship between the mean of the response and the linear predictor.
- The random effects are normally distributed: $b_i \sim \mathrm{N}(0, \sigma^2)$ independent of the residual variation.
- Observations are conditionally independent given the random effects.

Parameter estimation: The estimation of GLMM parameters is typically performed using Maximum Likelihood Estimation (MLE). However, due to the presence of random effects, the likelihood involves integrating over their distribution, which often lacks a closed-form solution. Therefore, numerical approximation methods such as the Laplace approximation or adaptive Gaussian quadrature are used (Velosa & Pestana, 2008; Zuur et al., 2009). The `glmmTMB()` function from the `glmmTMB` package in R implements these methods efficiently, allowing for flexible specification of fixed and random effects structures, and supports a range of distributions, including the Negative Binomial.

4.3.4 Model Comparison

To evaluate and compare different Generalized Linear Mixed Models (GLMM) with the same distribution family, negative binomial distribution, and the same random effects structure, the `anova()` function was used with the Likelihood Ratio Test (Likelihood Ratio Test (LRT)). This approach allows for the comparison of nested models, models where one is a simplified version of the other, and assesses whether including variables or interactions significantly improves model fit by comparing their likelihoods.

The anova() procedure relies on Maximum Likelihood estimation and calculates the difference between the log-likelihoods of the models. The significance of this difference is then evaluated based on the chi-squared distribution, which is appropriate for comparing fixed effects using the Likelihood Ratio Test (LRT) (Zuur et al., 2009).

$$\mathrm{LRT} = -2 \times (\log L_0 - \log L_1)$$

where:

- $\log L_0$: log-likelihood of the simpler model
- $\log L_1$: log-likelihood of the more complex model

Additionally, the anova() function provides information criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which evaluate the trade-off between model fit and complexity. Lower values of AIC or BIC indicate models that achieve a better fit with fewer parameters. While both criteria serve to prevent overfitting, AIC tends to favor more complex models, whereas BIC applies a stronger penalty for the number of parameters, often favoring simpler models by the principle of parsimony (Velosa & Pestana, 2008).

This analysis thus enables the selection of the most parsimonious model that maintains strong explanatory power, contributing to a robust statistical interpretation of the data while avoiding overfitting.

4.4 Bootstrap Method

Statistical methods allow us to draw inferences about a population and estimate unknown parameters based on a sample. Point estimation plays a key role by providing values that can support more advanced analyses.

Given a sample

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

drawn from an unknown population, we aim to estimate parameter θ by using statistics $S(\mathbf{x})$

$$\hat{\theta} = S(\mathbf{x})$$

from the observed data.

Resampling techniques, such as the bootstrap, generate new samples from the original data, typically with replacement, to approximate the sampling distribution of $\hat{\theta}$. As illustrated in Figure 4.1, these bootstrap samples form the basis of a non-parametric approach to statistical inference.

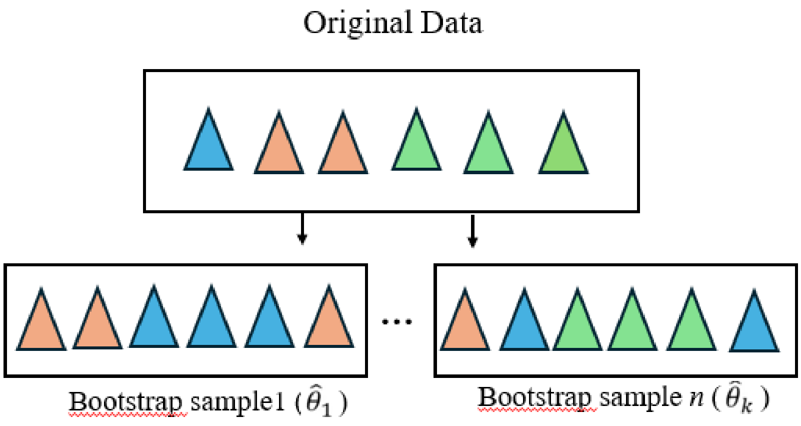

Original Data
...
Bootstrap sample1 ($\hat{\theta}_1$)
Bootstrap sample n ($\hat{\theta}_k$)

Figure 4.1: Illustration of the bootstrap resampling process.

Introduced by Bradley Efron in 1979 (Efron, 1979), the bootstrap is particularly valuable when classical methods struggle, such as with small sample sizes or non-normal data. It enables estimation of statistical accuracy, like standard errors or confidence intervals, without strong assumptions about the underlying population.

The bootstrap method typically follows four main steps:

1. **Resampling with replacement:** New samples of the same size are generated by randomly selecting observations from the original dataset, with replacement.

2. **Generation of bootstrap statistics:** For each $b = 1, 2, \ldots, B$, a bootstrap sample

   $$\mathbf{x}^{*b} = (x_1^*, x_2^*, \ldots, x_n^*);$$

   it is drawn, and the statistic of interest is computed:

   $$\hat{\theta}^{*b} = S(\mathbf{x}^{*b})$$

3. **Repetition:** The resampling and computation steps are repeated a large number of times. This process yields B bootstrap replications.

4. **Estimation and inference:** The bootstrap statistics are used to estimate the sampling distribution of the statistic. Confidence intervals can be constructed using the empirical percentiles—typically the 2.5th and 97.5th percentiles—yielding a bootstrap 95% confidence interval. This approach avoids assumptions about the underlying population distribution and is robust to skewness and small sample sizes.

Bootstrap distribution: The B statistics

$$\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$$

from the empirical distribution of the estimator under the bootstrap model, serving as an approximation of its sampling distribution.

The estimate of the standard error of the statistic can be estimated as:

$$\widehat{\mathrm{SE}}_{\hat{\theta}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{*b} - \bar{\theta}^{*}\right)^2}$$

where $\hat{\theta}^*$ is the average of the bootstrap estimates:

$$\hat{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b}$$

Percentile confidence interval: To construct a non-parametric $(1-\alpha)$ confidence interval, the ordered bootstrap estimates define the bounds:

$$CI_{(1-\alpha)} = \left[\hat{\theta^*}_{(\alpha/2)}, \hat{\theta^*}_{(1-\alpha/2)}\right]$$

where $\hat{\theta^*}_{(\alpha/2)}$ and $\hat{\theta^*}_{(1-\alpha/2)}$ are the empirical quantiles of probabilities $\alpha/2$ and $1-\alpha/2$ This method offers an intuitive, assumption-free way to assess the uncertainty of the estimator.

Chapter 5

Results and Discussion

5.1 Descriptive Analysis

Initially, a descriptive analysis was performed to characterize the sample and present the variables under study, using measures of central tendency, dispersion, and distribution shape. Graphical representations were also used when considered helpful to improve data visualization. Given the paired nature of the data, a summary of the key variables is provided below.

All statistical analyses were conducted using R software (version 4.4.1) in the RStudio environment. Several R packages were employed, including `ggplot2` and `ggpubr` for data visualization, `dplyr`, `tidyverse`, and `readxl` for data manipulation and import, `boot` for bootstrap analysis, `glmmTMB` and `lme4` for fitting generalized linear mixed models, `DHARMa` for residual diagnostics, `emmeans` and `MuMIn` for estimated marginal means and model selection, and `car` and `lmtest` for statistical tests. Additional packages such as `FactoMineR`, `corrplot`, `broom`, `psych`, and `FSA`, among others, were used as needed throughout the analysis.

5.1.1 Number of vacuoles for Weka and Saturation:

The box plots for the variable number of vacuoles, stratified by the outcome of each plugin that analyses the same images, reveal differences between the two methods, as outlined in Figure 5.1. The Weka plugin reaches higher values of the number of vacuoles (as seen by the range), compared to the Saturation plugin, but 50% upper values above the median are more dispersed, suggesting greater variability. However, the median values for both plugins are similar. For Saturation, the range spans from 13 to 20795. For this variable, the number of vacuoles ranges from a minimum of 5 to a maximum of 28076.

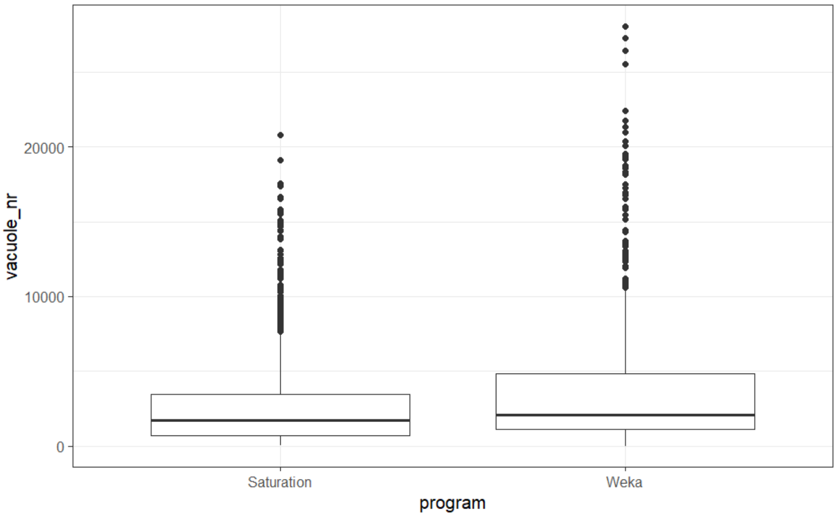
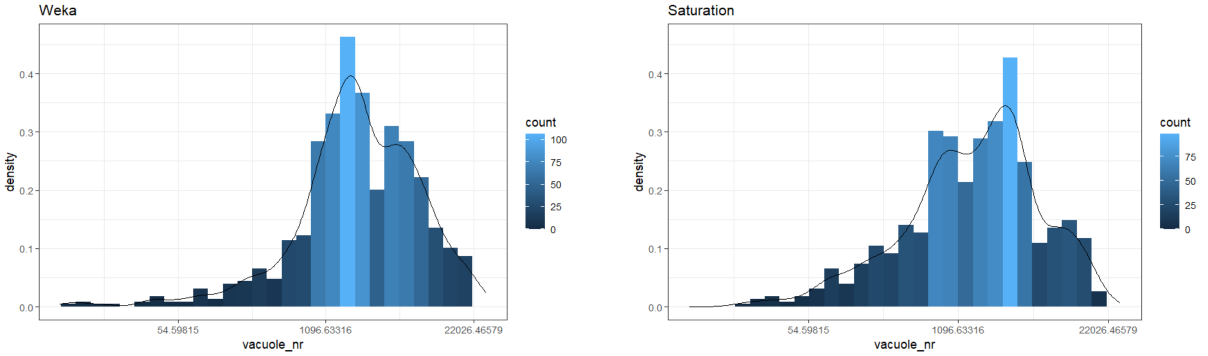


Figure 5.1: Distribution of the number of vacuoles identified by each program. The box plots compare the results obtained using the Saturation and Weka methods.

The distribution of the log-transformed number of vacuoles for both plugins is shown in Histogram 5.2. Given the strong right skewness observed in the raw values, the data have been log-transformed to improve visualization and interpretability. Without transformation, the high concentration of low counts and the presence of extreme outliers would obscure relevant distributional patterns. After transformation, both distributions appear more symmetric and reveal subtle differences between the two methods. While both distributions peak within a similar range, the Weka plugin shows a slightly narrower and more centralized distribution compared to Saturation, suggesting that Weka may produce more consistent estimates of vacuole count. In contrast, Saturation displays a broader spread with a longer right tail, which may indicate a higher variability in its measurements or sensitivity to larger values.

These differences highlight the importance of assessing not only central tendency but also dispersion when comparing automated image analysis tools.



(a) Histogram and density plot of the variable log number of vacuoles for the Weka.

(b) Histogram and density plot of the variable log number of vacuoles for the Saturation.

Figure 5.2: Comparison of Weka and Saturation for the variable number of vacuoles. The histogram represents the distribution of observations, while the overlaid density curve illustrates the estimated probability density function.

5.1.2 Area of Vacuoles for WEKA and Saturation:

The box plots for the variable area of vacuoles, stratified by the result of each plugin that analyses the same images, shows minimal differences between the two plugins, as shown in Figure 5.3. The Weka plugin detects a bigger area of vacuoles (as seen by the range), compared to the Saturation plugin, but 50% upper values above the median are more dispersed, suggesting greater variability. However, the median values for both plugins are close, but not identical. This could be attributed to Weka detecting a greater number of vacuoles than Saturation, as shown in Figure 5.1. As a result, it displays a larger vacuole area compared to Saturation. The slight difference in box sizes suggests variability discrepancies, with the Weka plugin exhibiting a wider spread in the third quartile relative to the Saturation plugin.

For the vacuole area, the values range from a minimum of 1.71×10^2 μm^2 to a maximum of 8.14×10^5 μm^2. The average vacuole area is 1.19×10^5 μm^2, with a standard deviation of 1.53×10^5 μm^2.

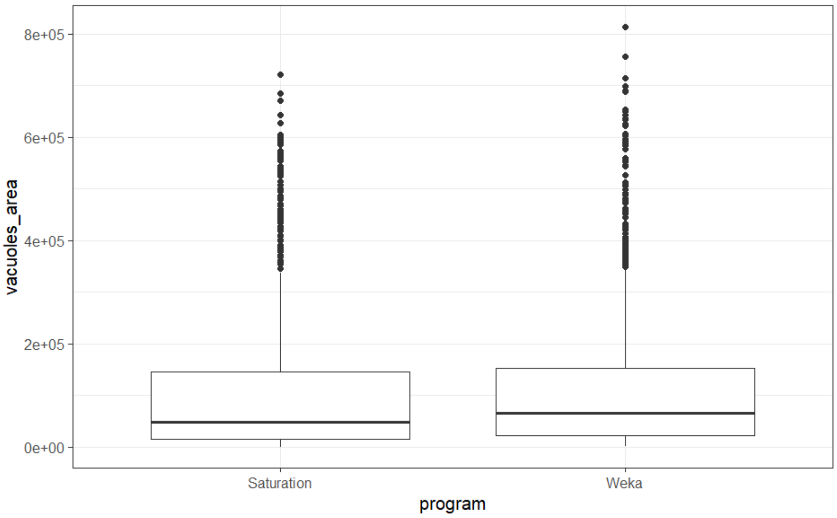
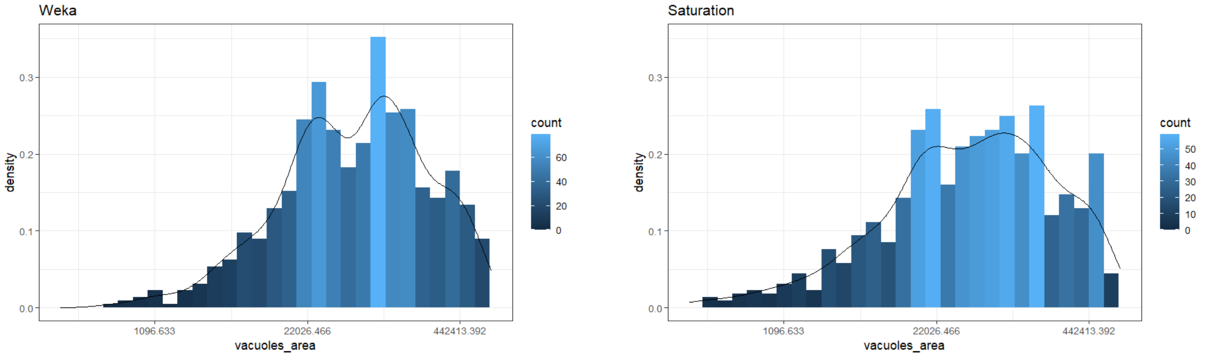


Figure 5.3: Distribution of the area of vacuoles identified by each program. The box plots compare the results obtained using the Saturation and Weka methods.



(a) Histogram and density plot of the variable log area of vacuoles for the Weka.

(b) Histogram and density plot of the variable log area of vacuoles for the Saturation.

Figure 5.4: Comparison of Weka and Saturation for the variable area of vacuoles. The histogram represents the distribution of observations, while the overlaid density curve illustrates the estimated probability density function.

Figure 5.4 illustrates the distribution patterns of log-transformed vacuole areas for both the Weka and Saturation plugins. The data were log-transformed to improve visualization and interpretability, particularly in the presence of skewed distributions. This transformation reduces the impact of extreme values and allows for a clearer comparison across values spanning multiple orders of magnitude, thereby facilitating more meaningful visual inspection and statistical analysis. In both plugins, the distributions exhibit positive (right) skewness, characterized by most observations being concentrated at lower vacuole areas, with a long tail extending toward higher values. This skewness likely reflects heterogeneity in vacuole size, possibly due to biological variability or experimental conditions, such as variations in steatosis severity among samples. Regarding specific differences, the Weka plugin shows a more pronounced bimodal pattern, with two identifiable peaks: one approximately at 22026 μm^2 and another slightly above. This suggests the possible existence of two distinct subpopulations of vacuoles. In contrast, the Saturation plugin also exhibits bimodality but with a broader and less sharply defined peak, indicating a more gradual accumulation of vacuole sizes across the mid-range values.

Consistent with previous observations regarding the number of detected vacuoles, Weka tends to identify a higher number of vacuoles, which leads to a cumulative increase in the measured area. This results in greater dispersion and higher maximum values, as evidenced by the heavier right tail of the distribution. Such variations must be considered when selecting or comparing image analysis methods for vacuole quantification. Further analyses will explore the area of vacuole distribution in relation to steatosis percentage, aiming to identify possible patterns associated with the progression of fatty liver disease. These findings are consistent with expectations in mice subjected to a high-calorie diet, which typically results in a fatty liver phenotype characterized by numerous vacuoles varying in size and number.

5.1.3 Percentage of steatosis for Weka and Saturation:

The box plots 5.5 illustrates the distribution of the percentage of steatosis as estimated by the Saturation and Weka plugins. Overall, both methods produce similar distributions, with comparable medians and interquartile ranges, indicating consistency in the ordinal classification of steatosis levels.

For the Saturation plugin, the percentage of steatosis ranges from a minimum of 0.0001 (< 1%) to a maximum of 0,48 (48%), with a median value of 0.18, and for the Weka plugin, values range from 0.02 to 0.51 (51%), with a slightly higher median of 0.22. The interquartile range is similar between methods, indicating consistent variability in the central portion of the data.

Both distributions show a similar range and variability, with some extreme values near the upper whiskers. These results indicate a consistent estimation of steatosis between the two image analysis methods, particularly in terms of central tendency and overall spread. The agreement supports the reliability of automated quantification for this variable, which is crucial in evaluating liver damage in metabolic disease models such as diet-induced steatosis.

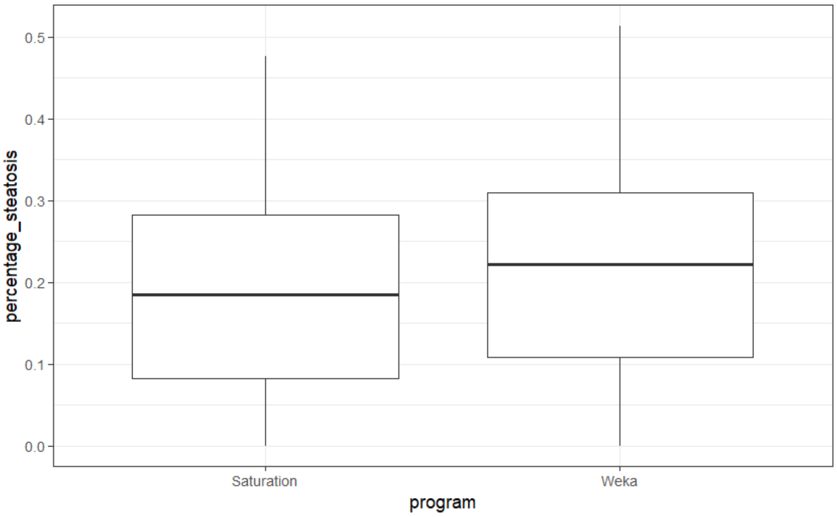


Figure 5.5: Distribution of the percentage of steatosis identified by each program. The box plots compare the results obtained using the Saturation and Weka methods.

5.2 Evaluation of the performance: Weka versus Saturation

The first objective is to investigate whether there are significant differences between the two plugins. Among the variables analysed, the number of vacuoles appears to be the most suitable for distinguishing between them. This choice was motivated by the variable's nature as a count of integers, which simplifies statistical handling and interpretation. Furthermore, it represents a direct measurement rather than a derived value based on ratios of other variables. This directness reduces potential sources of bias or compounded measurement error, providing a clearer and more reliable basis for analysis. It is important to note that we are working with replicates collected under varying conditions of magnification and resolution. Although the dataset includes 17 liver samples, there is a random number of images per liver (ranging from 4 to 7), analysed at different magnifications and resolutions for two distinct plugins. To account for the variability in image numbers among the set of images per liver, and to minimize potential biases, a new variable—the mean vacuole number—was created and used in further analyses in this work. It should be emphasized that the dataset does not include information to determine definitively which plugin performs better. Such an analysis will be an excellent direction for future research if this study is expanded. At this stage, our goal is to evaluate which plugin is more sensitive in detecting vacuoles. As demonstrated in Section 5.1 through graphical analysis, the distribution of the number of vacuoles variable deviates from normality. Accordingly, in all subsequent statistical analyses involving the mean and median, for the variable number of vacuoles, values of the 17 liver samples (statistical units) will be treated as non-normally distributed. Statistical tests were therefore conducted under the assumption of non-normality, thereby enhancing the validity and robustness of the resulting inferences by ensuring alignment between the analytical methods and the underlying data structure.

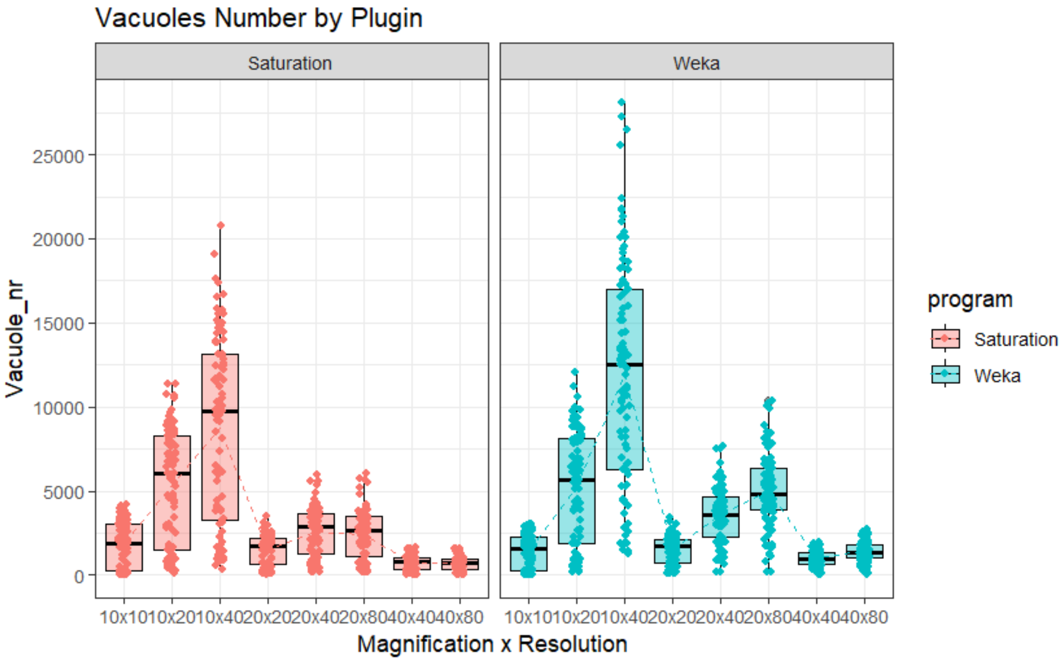


Figure 5.6: Distribution of the number of vacuoles detected by Saturation and Weka across different combinations of magnification and resolution. Each box plot represents the distribution of *Vacuole_nr* values for a given condition, with individual data points overlaid. The figure illustrates the variability and the differences in plugin performance in vacuole detection under varying imaging conditions.

The box plot 5.6 illustrates the distribution of the number of vacuoles across different Magnification × Resolution settings for both Weka and Saturation, across all measurements. The medians shown in the box plots indicate no substantial differences between the two plugins across most Magnification × Resolution combinations. However, notable differences are observed at the 10×40, 20×40, and 20×80 settings, where the Weka plugin detects a greater number of vacuoles compared to Saturation. This is evident from the higher medians in the box plots, as well as the broader distribution of values above the upper quartile (Q3) and extending into the whiskers, reflecting greater spread and variability. These patterns suggest that the Weka plugin consistently detects a larger number of vacuoles than Saturation, particularly under certain magnification and resolution conditions. Such findings emphasize the importance of selecting an appropriate Magnification × Resolution setting to obtain an accurate estimate of the percentage of steatosis. It is important to highlight that accurately assessing the percentage of steatosis is crucial for guiding treatment decisions, establishing prognostic evaluations, diagnosing disease severity, and evaluating liver suitability for transplantation.

5.2.1 Spearman's rank correlation

Accordingly to Triola, 2018, bivariate data analysis is often summarized visually through scatterplot and numerically through correlation coefficients. Prior to conducting the bivariate correlation analysis, the distribution of the variable number of vacuoles was assessed through graphical methods (histograms and Q-Q plots- see **Appendix B**). These evaluations indicated that the assumption of normality was not satisfied for the variables under study. Therefore, the Spearman rank correlation coefficient was chosen as the appropriate measure of association, given its robustness to violations of normality assumptions and its ability to capture monotonic, but not necessarily linear, relationships between variables (Field, 2013; Gauthier, 2001).

Given the violation of the assumption of normality, the Spearman rank correlation test was performed to assess the strength and direction of the monotonic relationship between the differences of the mean vacuole number detected by Weka and the Saturation plugins.

The hypotheses for the correlation analysis were defined as: Null hypothesis (H_0): There is no correlation between the mean vacuole number across both plugins ($\rho = 0$). Statement as:

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0$$

Table 5.1 summarizes the observed Spearman correlation coefficients (r_s) and the corresponding p-values, adjusted using the p.adjust() function, for each Magnification x Resolution (AR) group. This adjustment accounts for multiple comparisons to control the family-wise error rate. The Bonferroni correction method, known for its conservative nature, ensures that the overall Type I error rate remains below a chosen significance threshold (e.g., $\alpha = 0.05$). Analysing the Table 5.1, the Spearman correlation coefficients (r_s) ranged from moderate to high across most combinations of magnification and resolution between the two plugins. Particularly at lower magnifications (10×), all r_s values exceeded 0.90 and were highly significant (p < 0.00001), indicating strong monotonic agreement between methods under these settings. However, the correlation coefficients decreased as magnification increased. The 40×40 combination yielded a weaker and non-significant correlation (r_s = 0.5837; p = 0.1111), suggesting reduced reliability and higher variability between methods under this condition. In contrast, the 40×80 combination showed a moderately strong and statistically significant correlation (r_s = 0.7770; p = 0.003), partially recovering consistency. These findings underscore the importance of selecting appropriate imaging parameters, as both magnification and resolution can influence the agreement between automated quantification methods.

Table 5.1: Spearman correlation coefficients r_s and p-values by AR group

Magnification x Resolution	r_s	**p-value***
10 x 10	0.9289	< 0.00001
10 x 20	0.9314	< 0.00001
10 x 40	0.9338	< 0.00001
20 x 20	0.9069	< 0.00001
20 x 40	0.8284	0.00029
20 x 80	0.7377	0.00853
40 x 40	0.5837	0.11119
40 x 80	0.7770	0.00299

**Adjusted p-values using the Bonferroni correction method.*

The analysis yielded a Spearman's rank correlation coefficient (r_s), as shown in Figures 5.7, 5.8, and 5.9. A strong positive monotonic relationship was observed between the Weka and Saturation methods for the variable representing the differences in the mean number of vacuoles, with $r_s = 0.93$ and a significance level of p-value $< 10^{-6}$ across all combinations (10×10, 10×20, and 10×40). In the 10×10 and 10×20 combinations, the points are closely clustered around the red trend line, suggesting that both methods provide highly concordant results in measuring the vacuole count differences (i.e., as the Weka value increases, the Saturation value also increases consistently). However, in the 10×40 combination, the points are more dispersed around the red trend line, indicating greater variability for medium and high values of Weka and Saturation.

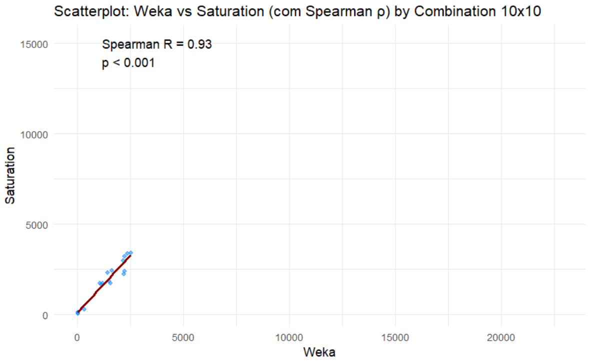


Figure 5.7: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 10×10)

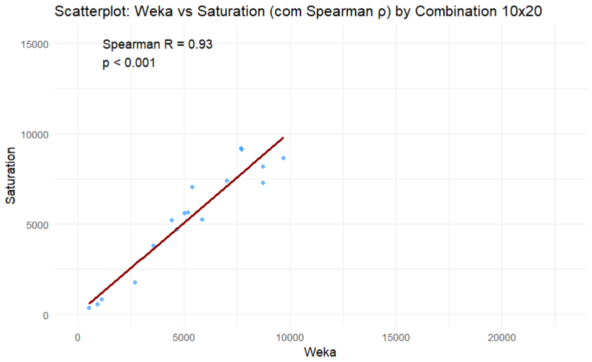


Figure 5.8: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 10×20)

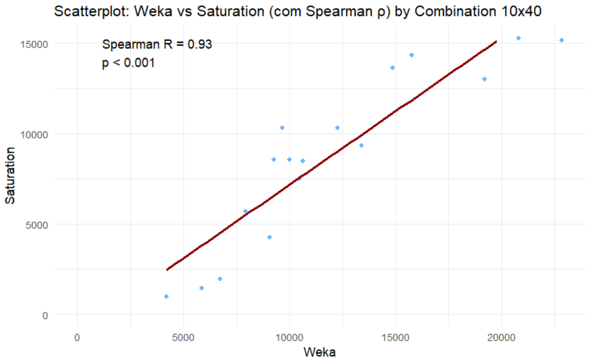


Figure 5.9: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 10×40)

For the combination 20×20, the analysis yielded a Spearman's rank correlation coefficient as shown in Figure 5.10. A strong positive monotonic relationship was observed between the Weka and Saturation methods for the differences in the mean number of vacuoles, with a Spearman's rho of $r_s = 0.91$ and a significance level of p-value $< 10^{-6}$. The points are closely aligned with the red trend line, indicating a high level of agreement and consistency between the two measurement methods. Similarly, for the combination 20×40, a strong positive monotonic relationship was observed, as shown in Figure 5.11, with a Spearman's rho of $r_s = 0.83$ and a p-value $< 10^{-6}$. Although the points remain relatively close to the red trend line, there is slightly greater dispersion compared to the combinations at 10× magnification, suggesting increased variability in measurements at intermediate values. For the combination 20×80, the analysis revealed a moderately strong positive monotonic relationship, as shown in Figure 5.12, with a Spearman's rho of $r_s = 0.74$ and a p-value of 0.001. While the correlation remains significant, there is noticeably greater dispersion of the points around the trend line, particularly for medium and high values, indicating higher variability and reduced consistency between the two measurement methods at larger combination sizes.

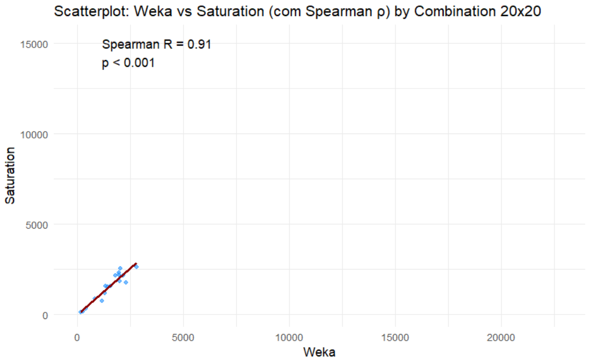


Figure 5.10: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 20×20)

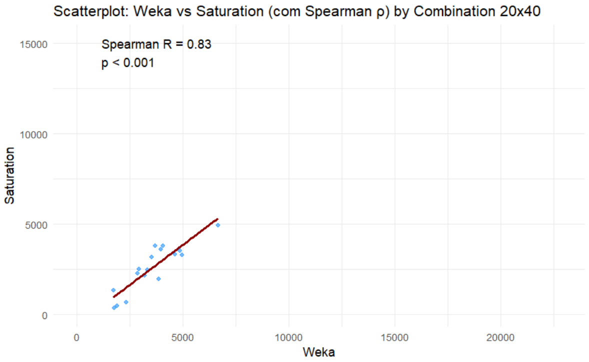


Figure 5.11: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 20×40)

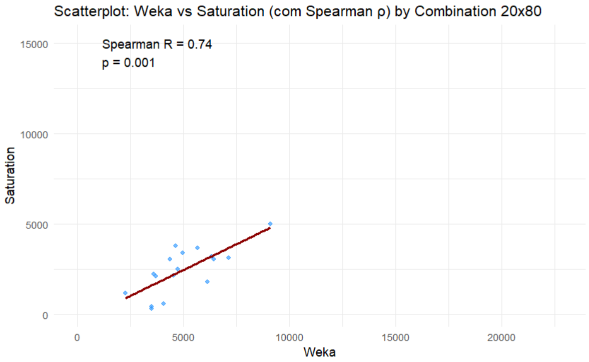


Figure 5.12: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 20×80)

For the combination 40×40, the analysis yielded a Spearman's rank correlation coefficient as shown in Figure 5.13. A moderate positive monotonic relationship was observed between Weka and Saturation for the differences in the mean number of vacuoles, with a Spearman's rho of $r_s = 0.58$ and a significance level of p-value = 0.014. Although the correlation is statistically significant, the points are widely dispersed around the red trend line, suggesting a lower degree of agreement between the methods. Moreover, the near-horizontal orientation of the trend line indicates limited consistency in the measurements for this combination size, with a weaker association compared to smaller combinations.

For the 40×80 combination, a strong positive monotonic relationship was observed between Weka and Saturation, with a Spearman's rho of $r_s = 0.78$ and a p-value $< 10^{-6}$. The points show a much better alignment with the red trend line compared to the 40×40 combination, indicating stronger consistency between the methods, although some degree of variability remains, particularly for intermediate values.

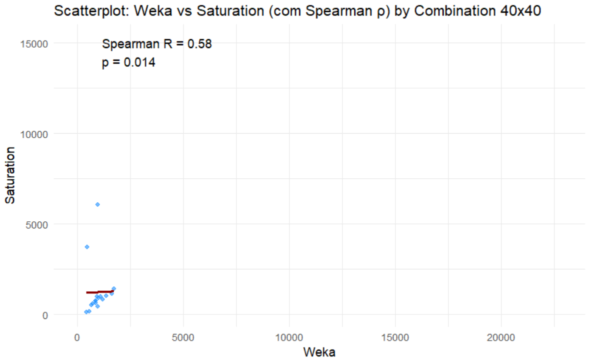


Figure 5.13: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 40×40)

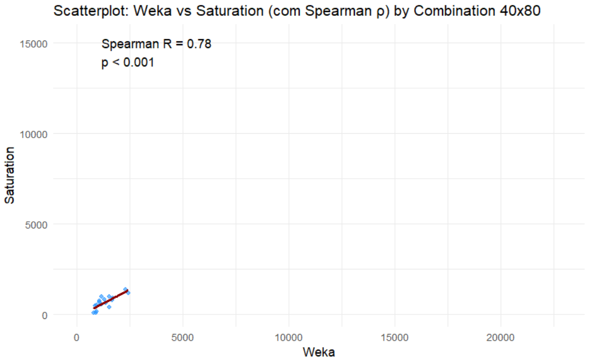


Figure 5.14: Spearman Correlation of the Differences in the Mean Number of Vacuoles between Weka and Saturation Plugins (Combination 40×80)

The analysis of Spearman's rank correlation coefficients across the eight combinations revealed consistently positive and statistically significant monotonic relationships between the Weka and Saturation methods for the mean number of vacuoles. For the combinations 10×10, 10×20, and 10×40 (at the same magnification level), a very strong correlation was observed, with points closely aligned to the trend lines, indicating a high degree of consistency and agreement between both Weka and Saturation. Although a slight increase in variability was noted for 10×40, where resolution increases, the overall relationship remained strong. In the intermediate combinations 20×20, 20×40, and 20×80 (Magnification 20×), greater dispersion around the trend lines was observed, particularly for 20×80, where the correlation strength decreased ($r_s = 0.74$), suggesting increased measurement variability with larger combination sizes. In the 40×40 and 40×80 combinations (magnification 40×), the correlation between methods remained strong, with better alignment of the points along the trend line at the higher resolution (40×80).

Overall, these findings suggest differences across combinations, indicating that smaller combinations provide greater consistency and agreement between Weka and Saturation measurements, while larger combinations tend to introduce greater variability and reduce the strength of the association. However, a robust and statistically significant correlation was consistently observed between the mean number of vacuoles detected by the two plugins.

5.2.2 Wilcoxon signed-rank test

The objective of this analysis is to evaluate whether there are significant differences between the two plugins (Weka vs. Saturation) while considering the combination of magnification and resolution as factors (10×10, 10×20, 10×40, 20×20, 20×40, 20×80, 40×40, and 40×80) for the variable mean of the number of vacuoles. Each combination includes a sample size of 17 pairs of values. Given the paired nature of the data, where the same images were analysed by both plugins, a Wilcoxon signed-rank test was employed to compare the two conditions 5.2. This test examines whether there is a significant difference in the median values between the two paired conditions. This non-parametric test was chosen because:

1. **Non-normal distribution:** The Shapiro-Wilk test was used to assess the normality of the paired differences between each combination of plugins. The test yielded a p-value of 7.42×10^{-9}, which is below for all usual significance levels. Therefore, the null hypothesis of normality was rejected, indicating that the differences do not follow a normal distribution.

2. **Small sample size:** Non-parametric methods are more robust for small samples and do not rely on distributional assumptions.

3. **Conservative approach:** The use of the Wilcoxon test ensures that any detected differences are less likely to be influenced by outliers or skewed distributions.

The p.adjust() function was used to adjust the p-values for multiple comparisons, in order to control the family-wise error rate. The Bonferroni correction method, known for its conservative nature, ensures that the overall Type I error rate remains below a chosen significance threshold (e.g., $\alpha = 0.05$). This correction was applied because multiple tests were conducted across different combinations of Magnification and Resolution within the same 17 statistical units. In this analysis, a significance level of $\alpha = 0.05$ was used to determine statistical significance.

Table 5.2: Wilcoxon Test for paired samples for comparing Weka vs Saturation

Magnification/Resolution	**10**	**20**	**40**	**80**
10	0.0004	1.0000	0.0002	—
20	—	1.0000	0.0002	0.0001
40	—	—	0.36	0.0001

Adjusted p-values using the Bonferroni correction method.

Analysing the results presented in Table 5.2, we observe that the mean of the number of vacuoles differs between several combinations of magnification and resolution. The effect of the plugin is not consistent, suggesting the presence of a heterogeneous effect between the plugin, magnification, and resolution.

Specifically, for the combinations 10×20, 20×20, and 40×40, both plugins yield comparable results, with no statistically significant differences. However, for all other combinations, the null hypothesis was rejected, indicating significant differences between the two plugins under these conditions.

The Wilcoxon test alone does not provide a definitive explanation for the observed differences. Several factors may account for the discrepancies between Weka and the Saturation method. Weka, a supervised classifier, extracts multiple features (colour, texture and shape) to identify vacuoles, whereas the Saturation method segments vacuoles by applying a fixed threshold to the image's saturation channel. Consequently, Weka can capture subtle variations in staining and morphology, while the threshold-based approach may fail to detect vacuoles with atypical appearance or to classify artifacts correctly when staining intensity varies. Nevertheless, the results highlight that Magnification and Resolution settings play a crucial role in image analysis, directly influencing vacuole count outcomes. This suggests that the observed differences may stem from how each plugin processes the images, rather than reflecting an inherent statistical superiority of one plugin over the other.

5.3 Generalized Linear Mixed Model

The primary objective of this analysis is to identify which program (Weka or Saturation) is more effective and to determine the best combination (Magnification and Resolution) for detecting the number of vacuoles. A GLMM with multiple explanatory variables will be constructed to assess how the response variable, the number of vacuoles, is affected and associated with other variables in the data set, namely program, magnification, and resolution. Given that both magnification and resolution have multiple levels, the model will allow for the evaluation of which specific levels contribute significantly to the vacuole count. During the initial exploration of the dataset, and given that the response variable, the number of vacuoles, represents count data (non-negative integers), we began the modeling process with a GLM assuming a Poisson distribution (McCullagh & Nelder, 1989). However, the Poisson model assumes equidispersion (i.e., the variance equals the mean), an assumption that was violated in our dataset due to overdispersion, where the variance exceeded the mean. This violation led to underestimated standard errors and inflated Type I error rates, thereby undermining the model's reliability (Cameron & Trivedi, 2013). To address this issue, we employed a Negative Binomial GLM, which introduces a dispersion parameter to accommodate overdispersion (Hilbe, 2011). Although this model provided a better fit to the variance structure, it did not account for the inherent correlation in the data, as multiple liver images originated from the same subject and were evaluated by both programs (Weka and Saturation), resulting in repeated measures. As a result, a Negative Binomial GLMM was employed, which incorporates random effects to model within-subject variability and is particularly suited for overdispersed count data with hierarchical or repeated measures designs (Bolker et al., 2009; Zuur et al., 2009). The final model included the number of vacuoles as the response variable, with *program*, *magnification*, and *resolution* as explanatory variables. This approach facilitated hypothesis testing regarding the associations between predictors and the response, providing estimates of their magnitude and direction, and allowed for a comprehensive understanding of the factors influencing vacuole detection in liver images.

To achieve these objectives, the following methodological approach was followed:

1. The response variable, the number of vacuoles, was analysed separately with each independent variable (Program, Magnification, and Resolution) to evaluate individual effects and identify

potential associations. This was performed using a Generalized Linear Mixed Model (GLMM) with a Negative Binomial distribution, as detailed in Section 5.3.1.

2. Next, the Variance Inflation Factor metric was used to measure multicollinearity among independent variables in a multivariable model. It quantifies how much the variance of an estimated regression coefficient is increased due to the correlation with other predictors in the model.

3. Finally, first, we fitted a negative-binomial GLMM including all independent variables in additive form and a random intercept to account for heterogeneity between specimens. Next, two models were fitted by adding interaction terms to capture potential combined effects 5.3.2.

5.3.1 GLMM Univariable Analysis

As an initial exploratory step, we fitted separate Negative Binomial GLMM for each independent variable - program, magnification, and resolution (see table 5.3). This approach allowed us to assess the individual association of each qualitative predictor with the mean number of vacuoles while already accounting for both overdispersion and within-subject correlation. The insights gained from these simpler models informed the construction of a final multivariable GLMM, including all relevant predictors and their potential interactions.

$$\log(\mathbb{E}[\text{vacuole\_nr}_i]) = \beta_0 + \beta_1 X_i + u_{\text{specimen}[i]}, \quad u_{\text{specimen}[i]} \sim \mathcal{N}(0, \sigma_u^2)$$

- vacuole_nr_i: response variable (number of vacuoles for unit i).
- β_0: fixed intercept (global mean on the log scale).
- X_i: covariate (*program*, *magnification*, *resolution*).
- $u_{g[i]}$: random intercept for group g, to which unit i belongs (*specimen_nr*).
- $\mathbb{E}[\text{vacuole\_nr}_i]$: expected value of the response variable, modeled using a logarithmic link function.
- $u_{g[i]} \sim \mathcal{N}(0, \sigma_u^2)$: random effects are assumed to follow a normal distribution with mean 0 and variance σ_u^2.

Table 5.3: Univariable GLMM (Negative Binomial) results for each qualitative predictor. Estimates are on the log scale; exponentiated estimates represent multiplicative effects on the expected number of vacuoles.

Variable	Term	(β)	$\exp(\beta)$	p-value
Program	*Program (Weka)*	0.401	1.493	5.33×10^{-15}
Magnification	*Magnification 20×*	-0.661	0.516	5.22×10^{-41}
	Magnification 40×	-1.764	0.171	3.91×10^{-230}
Resolution	*Resolution 20×*	0.772	2.164	7.84×10^{-20}
	Resolution 40×	1.189	3.283	5.38×10^{-49}
	Resolution 80×	0.603	1.827	3.27×10^{-12}

The model assessing the effect of the independent variable *program* indicates that using the Weka plugin is associated with a 49.3% increase in the expected number of vacuoles compared to the Saturation program ($\exp(0.401) = 1.493$). The coefficient estimate of 0.401, with a standard error of 0.0513, suggests a relatively precise estimate. The effect is statistically significant ($p < 0.001$), providing strong evidence of a positive association between Weka usage and vacuole count. Regarding the random intercept, it was modelled as following a normal distribution with mean 0 and estimated variance of 0.3151 (SD = 0.5613). This reflects the variability in the baseline vacuole counts between specimens, after accounting for the fixed effect of the *program*.

The model evaluating the effect of *magnification* shows that increasing magnification is associated with a significant decrease in the expected number of vacuoles. Specifically, a magnification of 20× is associated with a 48.4% reduction in vacuole count compared to the reference level ($\exp(-0.661) = 0.516$), while 40× magnification results in an 82.9% reduction ($\exp(-1.764) = 0.171$). Both effects are highly significant ($p < 0.001$), indicating a strong negative association between magnification and vacuole detection. This pattern is consistent with the fact that increasing magnification reduces the field of view, capturing a smaller sample area and consequently leading to fewer observed vacuoles. These results establish a clear inverse relationship between magnification level and vacuole count. Figure 5.15 reinforces that the vacuole count decreases as the magnification increases. At 10x magnification, the vacuole counts are highly variable, with many high outliers and a median above 4,000; but at 40x, the vacuole count is much more consistent and lower overall, with very few outliers and a median close to 1,000. As this model includes *magnification* as a fixed effect, the random intercepts for *specimen_nr* showed a variance of 0.2262 (SD = 0.4756), indicating moderate variability between specimens .

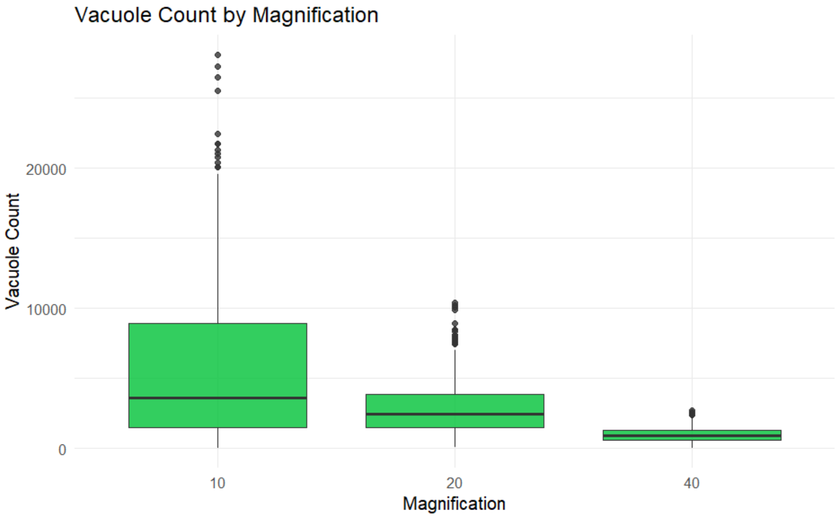


Figure 5.15: Box plots showing vacuole count distribution across different image magnification levels.

Regarding the variable *resolution*, the model indicates that higher resolution levels lead to a statistically significant increase in the expected number of vacuoles. Resolutions of 20, 40, and 80 pixels per unit are associated with increases of 116% ($\exp(0.772) = 2.16$), 228% ($\exp(1.189) = 3.28$), and 83% ($\exp(0.603) = 1.83$), respectively, compared to the reference resolution. All estimates are statistically significant ($p < 0.001$), suggesting that resolution enhancement improves vacuole detection, with the most pronounced effect observed at 40 pixels. Looking at Figure 5.16, higher resolutions are generally associated with increased vacuole detection, with the highest median and variability observed at 40x. However, 80x shows a decrease in median count, suggesting a potential image saturation or non-linear relationship. The 10x resolution exhibits a small median and less variability, as expected, due to the limitation in image detail. In this model, the random intercepts for *specimen_nr* were assumed to follow a normal distribution with a mean of 0, an estimated variance of 0.2959, and a standard deviation of 0.544.

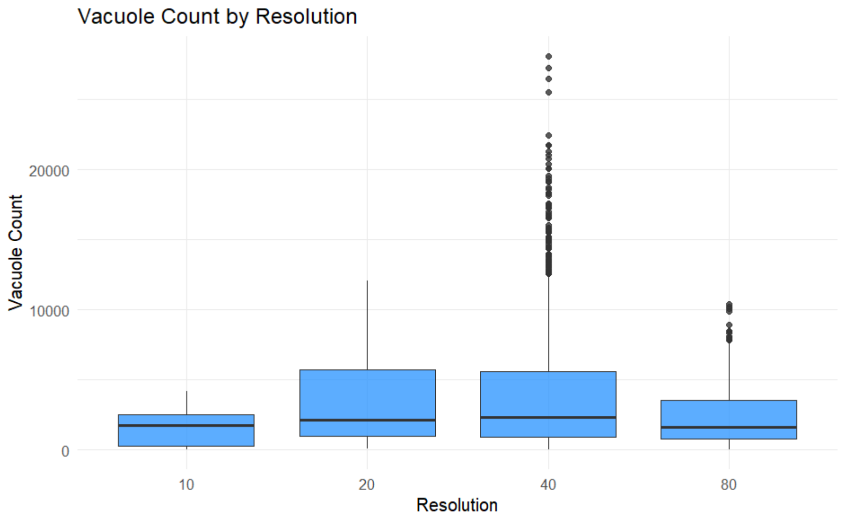


Figure 5.16: Boxplots showing the distribution of vacuole counts across different image resolution levels.

In summary, the models suggest that both program and imaging parameters (magnification and resolution) significantly influence vacuole detection, Figure 5.17. Using the Weka plugin and increasing image resolution enhances detection, while higher magnification leads to lower vacuole counts due to reduced field coverage- a result that aligns with biological expectations. These findings highlight the importance of optimizing imaging settings to ensure analytical accuracy in vacuole quantification.

To increase explanatory power, incorporating all predictors into a single multivariable model may improve model accuracy and provide a more comprehensive understanding of the factors influencing vacuole counts.

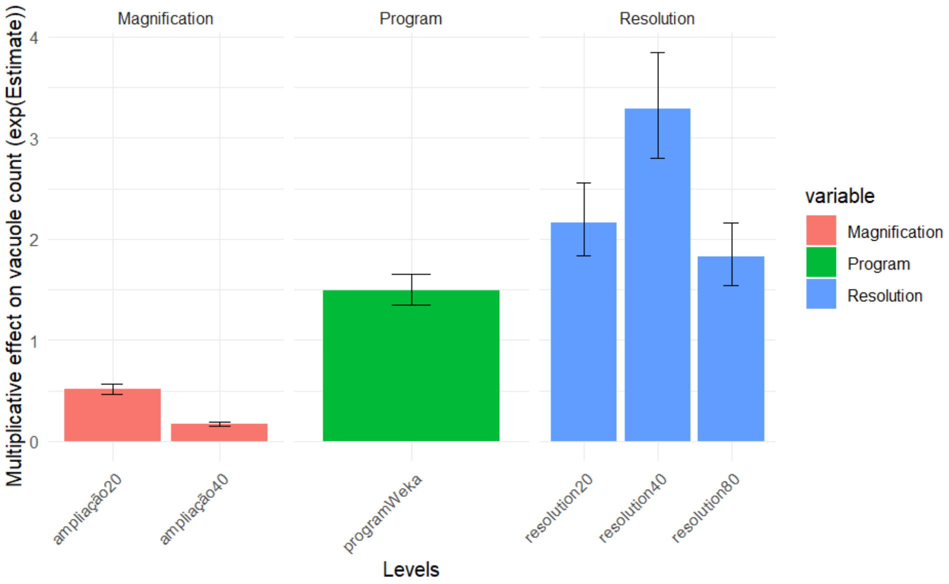


Figure 5.17: Multiplicative effects of the variables *program*, *magnification*, and *resolution* on the expected vacuole count, expressed as $\exp(\beta)$, with error bars representing confidence intervals.

5.3.2 GLMM Multivariable Analysis

The presence of overdispersion justifies the use of a Negative Binomial regression model. Overdispersion was assessed by initially fitting a Poisson regression model using a GLMM with a log-link function, and calculating the dispersion index as the ratio between the sum of squared Pearson residuals and the residual degrees of freedom:

$$\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{\text{Residual degrees of freedom}}.$$

where the Pearson residual was calculated as the standardized difference between the observed and expected values under the Poisson model, and the residual degrees of freedom, defined as the number of observations minus the number of estimated parameters.

A value of $\hat{\phi} > 1$ indicates overdispersion, suggesting that the Poisson model underestimates the variance, thus motivating the use of a Negative Binomial distribution. To assess the adequacy of the model, the dispersion parameter was calculated using the Pearson residuals. The resulting dispersion estimate was approximately 1209.6, indicating a substantial overdispersion in the data. This suggests that the variance of the response variable considerably exceeds the mean.

A negative binomial regression model with the nbinom2 parametrization was fitted using the `glmmTMB` package in R. This distribution allows the variance to increase quadratically with the mean, making it suitable for overdispersed count data. It is important to mention that the choice between the nbinom1 and nbinom2 was evaluated using both Akaike Information Criterion (AIC) and residual diagnostics based on the DHARMa package. Although the nbinom1 model exhibited slightly better residual behavior the nbinom2 model consistently yielded substantially lower AIC values. Since a lower AIC reflects a more optimal balance between model fit and parsimony, the nbinom2 distribution was chosen as the most appropriate specification.

An additive model will be used, which does not account for interactions between independent variables, assuming separate and independent effects. The model sums the effects of each variable:

$$\log\big(\mathbb{E}[\text{vacuole\_nr}_i]\big) = \beta_0 + \beta_1\,\text{program}_i + \beta_2\,\text{magnification}_i + \beta_3\,\text{resolution}_i + u_{\text{specimen}[i]},$$

$$u_{\text{specimen}[i]} \sim \mathcal{N}\big(0, \sigma_u^2\big).$$

- vacuole_nr$_i$: response variable (number of vacuoles for unit i).
- β_0: fixed intercept (global mean on the log scale).
- $\beta_1, \beta_2, \beta_3$: fixed coefficients (*program*, *magnification*, *resolution*).
- $u_{g[i]}$: random intercept for group g, to which unit i belongs (*specimen_nr*).
- $\mathbb{E}[\text{vacuole\_nr}_i]$: expected value of the response variable, modeled using a logarithmic link function.
- $u_{g[i]} \sim \mathcal{N}(0, \sigma_u^2)$: random effects are assumed to follow a normal distribution with mean 0 and variance σ_u^2.

In contrast, the interaction model allows us to analyse whether the effect of one independent variable depends on another in cases where a relationship between independent variables might exist. Instead of simply adding effects, it also considers the *combined effect* between variables:

$$\log\big(\mathbb{E}[\text{vacuole\_nr}_i]\big) = \beta_0 + \beta_1\,\text{program}_i + \beta_2\,\text{magnification}_i$$
$$+\,\beta_3\big(\text{program}_i \times \text{magnification}_i\big) + \beta_4\,\text{resolution}_i + u_{\text{specimen}[i]},$$
$$u_{\text{specimen}[i]} \sim \mathcal{N}(0, \sigma_u^2)$$

and,

$$\log\big(\mathbb{E}[\text{vacuole\_nr}_i]\big) = \beta_0 + \beta_1\,\text{program}_i + \beta_2\,\text{magnification}_i$$
$$+\,\beta_3\big(\text{magnification}_i \times \text{resolution}_i\big) + \beta_4\,\text{resolution}_i + u_{\text{specimen}[i]},$$
$$u_{\text{specimen}[i]} \sim \mathcal{N}(0, \sigma_u^2)$$

- vacuole_nr_i: response variable (number of vacuoles for unit i).
- β_0: fixed intercept (global mean on the log scale).
- $\beta_1, \beta_2, \beta_4$: fixed coefficients (*program*, *magnification*, *resolution*).
- β_3: fixed coefficient for the interactions: $\text{program}_i \times \text{magnification}_i$, and $\text{magnification}_i \times \text{resolution}_i$.
- $u_{g[i]}$: random intercept for group g, to which unit i belongs (*specimen_nr*).
- $\mathbb{E}[\text{vacuole\_nr}_i]$: expected value of the response variable, modeled using a logarithmic link function.
- $u_{g[i]} \sim \mathcal{N}(0, \sigma_u^2)$: random effects are assumed to follow a normal distribution with mean 0 and variance σ_u^2.

5.3.2.1 Variance Inflation Factor

Before fitting the model, we assessed multicollinearity among the independent variables by computing the variance inflation factor (Variance Inflation Factor (VIF)). Although multicollinearity does not invalidate GLMM as a whole, it can inflate the standard errors of the fixed-effect coefficients, render them unstable or of unexpected sign, and thus hinder the interpretation of individual effects (Zuur et al., 2009). If the independent variables are uncorrelated, then VIF = 1; the VIF value is expected to increase as the correlation among predictors increases. Following Montgomery et al., (2006), we consider VIF values above 5 as indicative of potentially problematic collinearity. Specifically, the VIF for *program* was 1.02, while *magnification* and *resolution* had values of 1.74 and 1.77, respectively. These results suggest a low to moderate correlation among predictors, well within acceptable limits, and no evidence of problematic multicollinearity.

5.3.2.2 Aditive Model

Considering the independent variables Program, Magnification, and Resolution, the following presents the analysis using an additive model, as shown in model (5.3.2).

Table 5.4: Multiplicative effects ($\exp(\beta)$) of predictors on mean vacuole count from the Negative Binomial (nbinom2) mixed model. All predictors are statistically significant ($p < 0.001$).

Variable	**exp(β)**	**95% CI**	**p-value**
Program (Weka)	1.48	[1.39, 1.58]	< 0.001
Magnification 20×	0.30	[0.27, 0.32]	< 0.001
Magnification 40×	0.08	[0.07, 0.09]	< 0.001
Resolution 20	3.24	[2.86, 3.66]	< 0.001
Resolution 40	6.92	[6.10, 7.85]	< 0.001
Resolution 80	8.58	[7.40, 9.93]	< 0.001

Analysing the random effect, where *specimen_nr* is treated as a random intercept, we observe a variance of 0.3363 across the 17 liver samples (SD = 0.5799). The estimated residual variance for this model is approximately 0.4049, so the random-effect variance represents 83.1% of the residual variance, indicating a large random-intercept effect and substantial between-specimen heterogeneity. This fully justifies the inclusion of *specimen_nr* as a random effect. Additionally, the dispersion parameter for the nbinom2 family is 2.47, confirming the presence of overdispersion in the vacuole count data. This supports the choice of a Negative Binomial distribution over the Poisson, which assumes equal mean and variance and would not be appropriate in this context. Examining the fixed effects of the predictor variables, and comparing them to the reference categories (Saturation, 10× Magnification, and 10× Resolution), the Program variable shows that utilizing the Weka plugin leads to a 48% increase in the expected vacuole count ($\exp(0.391) \approx 1.48$) compared to the Saturation method. This result aligns with findings from the simpler model, suggesting that Weka detects more vacuoles than Saturation, as shown in Chapter 5.3.1. For Magnification, the analysis reveals a significant reduction in vacuole count at higher levels. Specifically, compared to the reference level (10× magnification), using 20× magnification reduces the expected count by approximately 70% (rate ratio = ($\exp(-1.1915) \approx 0.30$), while 40× magnification results in a nearly 92% reduction (rate ratio = ($\exp(-2.5114) \approx 0.08$). This pattern is consistent with the simpler regression model containing only Magnification, supporting the hypothesis that higher magnification leads to fewer counted vacuoles, possibly due to improved resolution reducing overlapping or misidentified structures, as seen before in Chapter 5.3.1. In contrast, Resolution exhibits an increasing trend in vacuole count. Compared to 10× resolution, increasing the resolution to 20×, 40×, and 80× results in increases in vacuole count by factors of approximately 3.24, 6.92, and 8.58, respectively. These values correspond to the exponentiated model coefficients ($\exp(\beta)$), representing the multiplicative change in the expected number of vacuoles relative to the baseline resolution 10×. This suggests that higher resolutions allow for more precise vacuole detection, potentially counteracting the reduction caused by higher magnification, as noted in Section 5.3.1. All fixed effects included in the model (Program, Magnification, and Resolution) had very low p-values ($p < 0.001$), indicating strong statistical significance and suggesting that each variable contributes meaningfully to explaining the variability in *vacuole_nr*. The random effect associated with *specimen_nr* had a variance of 0.336, indicating moderate variability between specimens.

The figure 5.18 shows the multiplicative effects of the predictor estimates on the vacuole count. The Weka plugin increases the number of detected vacuoles by approximately 1.5 times compared to the Saturation plugin. Magnifications of 20× and 40× have a negative effect on vacuole counts when compared to the 10× reference level. In contrast, resolutions of 20×, 40×, and 80× are associated with an increase in vacuole count relative to the 10× resolution. The *resolution* predictors show confidence intervals well above 1, supporting a positive and statistically significant effect, although the interval for 80x is wider, suggesting greater uncertainty in its estimate.

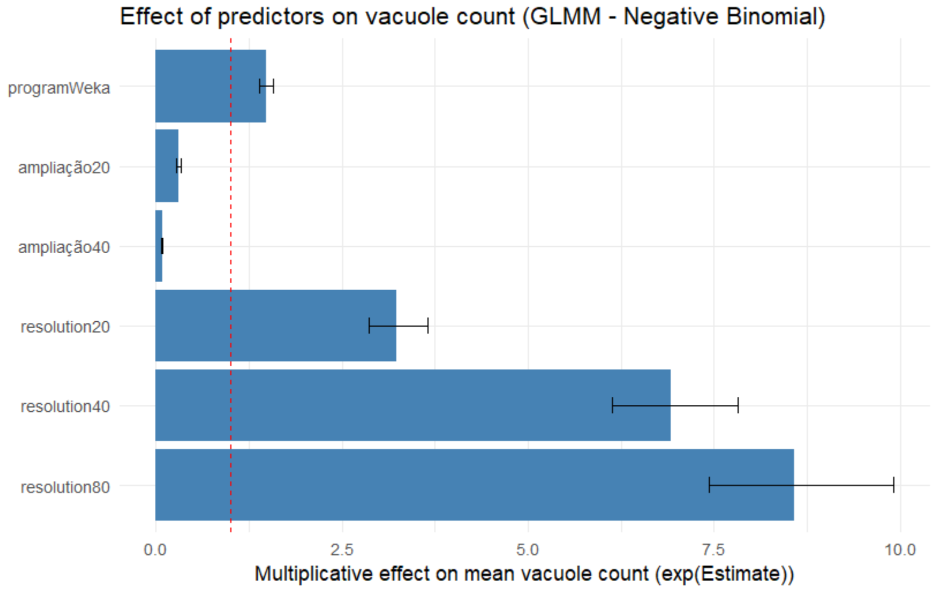


Figure 5.18: Multiplicative effects of each predictor on the expected vacuole count, estimated from a generalized additive linear mixed model with negative binomial distribution. The bars represent exponentiated coefficients (exp(Estimate)), which indicate how each variable influences the mean vacuole count relative to the reference levels (Saturation program, 10× magnification, and resolution 10). Error bars represent 95% confidence intervals.

5.3.2.3 Model with Interactions, Program x Magnification

Considering the independent variables Program, Magnification, and Resolution, the following presents the analysis using an iterative model, as shown in model 5.3.2.

Previously, an additive model was fitted on subchapter 5.3.2.2, which assumes that the predictor variables act independently, with no interaction effects. Under this approach, each variable contributes independently to the prediction of *vacuole_nr*, without being influenced by or influencing the effects of the others. To account for potential dependencies between predictors, the model was extended to include both the main effects and their two-way interaction terms. Specifically, this interaction model includes the main effects of *program*, *magnification*, and *resolution*, as well as their interactions: *program* × *magnification*, and *magnification* × *resolution*. By including these interaction terms, the model allows us to evaluate whether the effect of one variable depends on the level of another, capturing more complex relationships in the data. This approach provides a more comprehensive understanding of how predictors jointly influence the *vacuole_nr*.

To assess whether the effect of the image analysis *program* on vacuole detection varies with *magnification*, an interaction term between *program* and *magnification* was included in the model; see table 5.5. This extended model examines the combined influence of *program*, *magnification*, and *resolution*, using Saturation, 10× magnification, and resolution 10x as reference levels.

Table 5.5: Multiplicative effects ($\exp(\beta)$), 95% confidence intervals, and p-values from the Negative Binomial (nbinom2) mixed model with interaction Program x Magnification (`glmmTMB`).

Variable	**exp(β)**	**95% CI**	**p-value**
Program (Weka)	1.06	[0.95, 1.19]	0.308
Magnification 20×	0.24	[0.22, 0.26]	< 0.001
Magnification 40×	0.06	[0.05, 0.07]	< 0.001
Resolution 20	3.36	[2.97, 3.80]	< 0.001
Resolution 40	7.32	[6.45, 8.30]	< 0.001
Resolution 80	8.82	[7.57, 10.28]	< 0.001
Weka × Magn. 20×	1.57	[1.35, 1.83]	< 0.001
Weka × Magn. 40×	1.92	[1.63, 2.26]	< 0.001

Analysing the random effects, where *specimen_nr* is included as a random intercept, we observe a variance of 0.3502 across the 17 liver samples (SD = 0.5918). The estimated residual variance for the negative-binomial model is approximately 0.3891, so the random-effect variance represents 90.0% of the residual variance, well indicating a large random-intercept effect and substantial between-specimen heterogeneity. This objectively justifies the inclusion of *specimen_nr* as a random effect in the model. Additionally, the dispersion parameter for the nbinom2 family is estimated at 2.57, the same as in the additive model. The fixed effects of *magnification* and *resolution* were highly significant: higher magnifications (20× and 40×) were associated with a marked decrease in vacuole count, whereas higher resolutions (20, 40, and 80) led to substantial increases. Although the main effect of the Weka program was not significant on its own, its interaction with magnification was highly significant. Specifically, the interaction terms between Weka (program) and 20× magnification, and between Weka (program) and 40× magnification, indicate that Weka reduces the negative impact of higher zoom levels. For instance, under 40× magnification, Weka yielded a higher vacuole count compared to Saturation, suggesting that the program mitigates the loss of detail typically associated with increased zoom.

These findings demonstrate that the effect of magnification on vacuole detection depends on the program used. In practical terms, this implies that one program may perform better at lower magnifications while another may be more effective at higher levels. Therefore, program and magnification should be considered in combination when evaluating the performance of automated image analysis methods. This interaction highlights the importance of accounting for technical imaging parameters in quantitative microscopy workflows.

Figure 5.19 shows the predicted number of vacuoles (*vacuole_nr*) as a function of magnification levels, for the two plugins (Weka and Saturation). For both programs, the number of vacuoles consistently decreases as magnification increases, indicating an inverse relationship. This pattern confirms the findings observed across all fitted models. Weka appears to detect a higher number of vacuoles than Saturation, particularly at 10× magnification, although the difference becomes less pronounced at 40×.

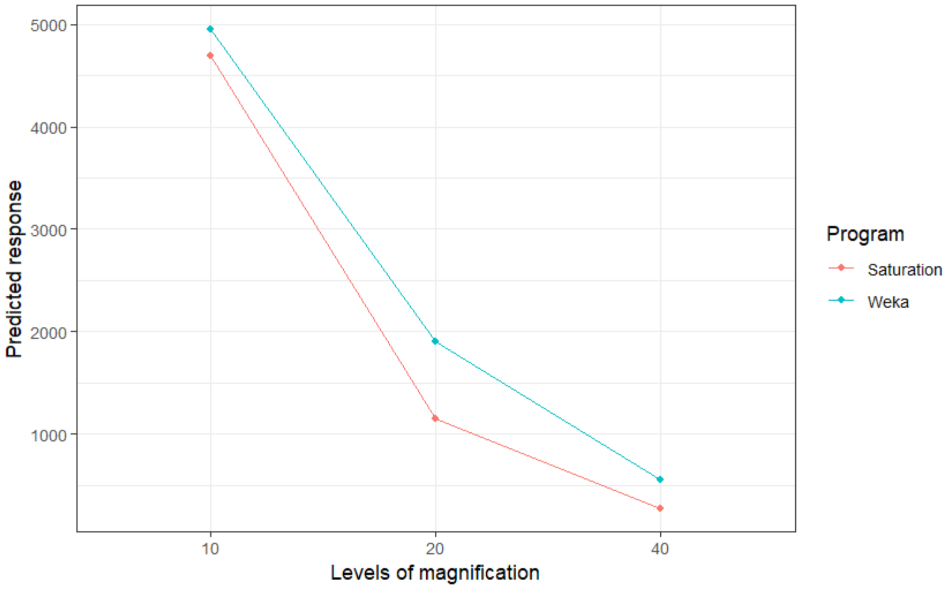


Figure 5.19: Predicted number of vacuoles across resolution levels (10, 20, 40, 80) for each magnification level (10×, 20×, 40×), based on a negative-binomial GLMM including the Program x Magnification interaction.

5.3.2.4 Model with Interactions, Magnification x Resolution

Considering the independent variables Program, Magnification, and Resolution, the following presents the analysis using an iterative model, as shown in model 5.3.2.

Analysing the impact of the predictors on *vacuole_nr*, distinct patterns were observed for *Program*, *Magnification*, *Resolution*, and the *Magnification* × *Resolution* interaction from Table 5.6.

Table 5.6: Multiplicative effects ($\exp(\beta)$), 95% confidence intervals, and p-values from the Negative Binomial (nbinom2) mixed model with `Magnification x Resolution` interaction (`glmmTMB`).

Variable	**exp(β)**	**95% CI**	**p-value**
Program (Weka)	1.48	[1.39, 1.58]	< 0.001
Magnification 20×	0.32	[0.27, 0.38]	< 0.001
Magnification 40×	0.08	[0.07, 0.09]	< 0.001
Resolution 20	3.29	[2.88, 3.76]	< 0.001
Resolution 40	6.81	[5.95, 7.79]	< 0.001
Resolution 80	8.33	[6.94, 10.00]	< 0.001
Magn. 20× × Res. 20	0.94	[0.75, 1.18]	0.55
Magn. 20× × Res. 40	0.99	[0.83, 1.19]	0.88

The random intercept for *specimen_nr* shows a variance of 0.3365 (SD = 0.58) across the 17 liver samples, indicating meaningful between-sample variability and justifying its inclusion as a random effect in the model. The results indicate that the use of the Weka plugin increases the expected vacuole count by a factor of approximately $\exp(0.39) \approx 1.48$, representing an increase of 48% compared to the saturation method. This finding is consistent with the simpler model that included only *Program* as a predictor, reinforcing Weka tendency to detect a higher number of vacuoles. As in previous models, *Magnification* shows a significant negative effect on vacuole counts. Compared to the 10× reference level:

- 20× magnification decreases the count by a factor of $\exp(-1.1547) \approx 0.315$, a 68% reduction;

- 40× magnification reduces it further to a factor of $\exp(-2.4889) \approx 0.083$, a 91% reduction.

These results support the findings from earlier models, confirming that increasing magnification leads to fewer vacuoles being counted, likely due to a smaller field of view. In contrast, *Resolution* exhibits the opposite trend, significantly increasing vacuole counts. Compared to 10× resolution:

- 20× resolution leads to an increase of approximately 228.5%,
- 40× resolution increases counts by 581%, and
- 80× resolution results in a 733% increase.

This suggests that higher resolution improves image clarity, enhancing vacuole detection. At lower resolutions, vacuoles may be missed due to blurring or pixelation, whereas higher resolutions allow for a more accurate count, a trend also observed in the simpler models.

Regarding the *Magnification* × *Resolution* interaction, some combinations were not statistically significant, such as *magnification20* × *resolution20* and *magnification20* × *resolution40*. This suggests that, in these cases, the magnification at the reference level (10×) may perform better. However, interpretation should be made cautiously, especially for combinations affected by data sparsity or collinearity.

Figure 5.20 shows the predicted number of vacuoles as a function of both resolution and magnification levels. For all levels of magnification, the number of vacuoles increases with resolution, confirming the findings previously observed across all models. The curve is steeper at 10× and progressively flattens at 40×. As magnification increases, the benefit of resolution becomes increasingly limited. These results reinforce the idea that there is an interaction between Resolution and Magnification, and that maximum detection efficiency can be achieved with low magnification and high resolution.

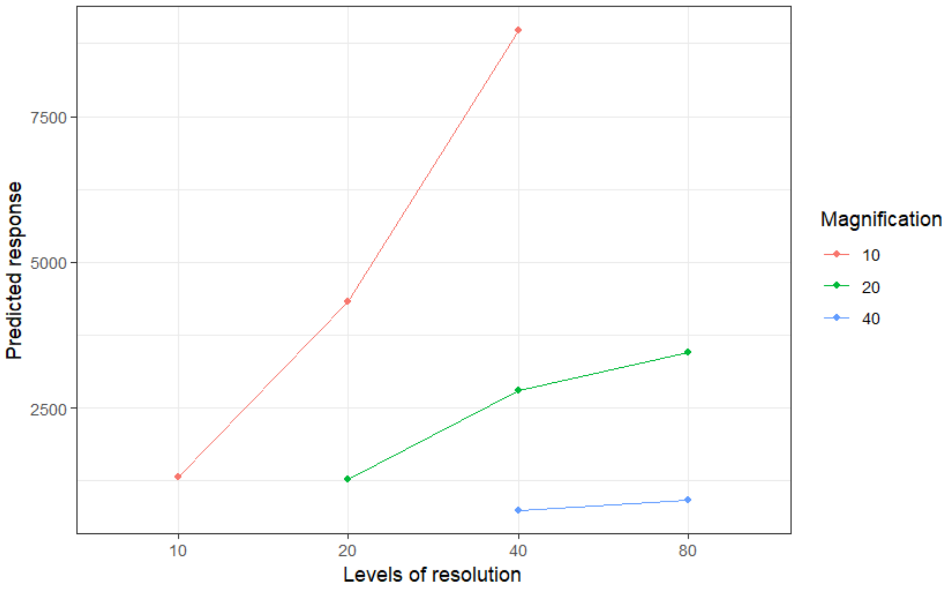


Figure 5.20: Predicted number of vacuoles across resolution levels (10, 20, 40, 80) for each magnification level (10×, 20×, 40×), based on a negative-binomial GLMM including the Program x Magnification interaction.

5.3.2.5 ANOVA

The Analysis of Variance (ANOVA), see Section 4.3.4, was performed to determine which model provides the best fit to the data. Table 5.7 shows a model comparison using likelihood ratio tests

(GLMM with a negative binomial distribution). The first comparison was made between the additive and interaction models, including the effect of the interaction between Program and Magnification.

Table 5.7: Model comparison using likelihood ratio tests (Additive GLMM and GLMM including interaction Program x Magnification)

Model	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
mod_nb2	9	26192	26241	-13087	–	–	–
mod_nb2_2	11	26129	26186	-13053	67.715	2	< 0.001

Based on the likelihood-ratio test performed by the `anova` function, comparing the additive model (without the interaction) to the model including the Program×Magnification interaction shows a highly significant improvement in fit, $X_0^2 = 67.715, p < 0.0001$. This supports the hypothesis that the effect of the image analysis program on vacuole detection varies depending on the magnification level. The AIC, where lower values indicate a better model fit, also supports this conclusion: the iterative model has the lowest AIC value (26129), indicating it is the best-fitting model among those tested. Similarly, the BIC is also lower for the interaction model (26187), and the log-likelihood is slightly higher (-13053) compared to the additive model (-13087). Based on these three criteria, the interaction model shows a better fit than the additive model.

Table 5.8 shows a model comparison using likelihood ratio tests (GLMM with a negative binomial distribution). This second comparison was made between the additive and interaction models, including the effect of the interaction between Magnification and Resolution.

Table 5.8: Model comparison using likelihood ratio tests (Additive GLMM and GLMM including interaction Magnification x Resolution)

Model	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
mod_nb2	9	26192	26241	-13087			
mod_nb2_3	11	26196	26255	-13087	0.4663	2	0.792

However, the interaction between *magnification* and *resolution* did not improve model fit compared to the additive model, $X_0^2 = 0.4663$, $p = 0.792$. This non-significant likelihood-ratio test suggests that these two variables contribute independently to predicting vacuole count in the current dataset. The AIC and BIC values for this model are slightly worse, indicating a minor increase in model complexity without corresponding gains in explanatory power. Additionally, the log-likelihood is identical to the previous model. Based on these three criteria, the interaction model provides no improvement when adding the interaction between Magnification and Resolution.

To conclude, the model incorporating the interaction between Program and Magnification demonstrates the superior fit; consequently, residual diagnostics were performed for this model (see **Appendix C**). Among the predictors, most show statistically significant associations ($p < 0.001$), except for the main effect of the Weka program, which is not significant. However, Weka significantly interacts with higher magnification levels, mitigating their otherwise strong negative effect on vacuole count. Consistent with previous findings, *magnification* has a strong negative effect: at both 20× and 40×, increasing magnification significantly reduces the vacuole count. In contrast, *resolution* shows a significant and progressive positive effect, with higher resolution levels (from 20× to 80×) leading to an increased vacuole count compared to the reference level of 10×. These results confirm that both Magnification and Resolution and the choice of image analysis program play critical roles in vacuole detection. They should therefore be jointly considered when developing and optimising automated image analysis workflows.

5.4 Number of Images per Liver

Based on the previous model results, Magnification was found to have a strong and consistent negative effect on vacuole count, with higher magnification levels (20× and 40×) significantly reducing the number of detected vacuoles. In contrast, Resolution exhibited a progressive positive effect, with higher resolution levels resulting in increased vacuole detection. Given these findings, a configuration combining 10× low magnification, which avoids loss of vacuole information — and 40× high resolution, which enhances detection capability, was considered representative and suitable for further analysis. While this choice does not exclude the possibility of more optimal combinations, this particular setup was selected as a practical and theoretically justified compromise to carry out the current analysis. This configuration was therefore applied uniformly across all 17 specimens to determine the minimum number of images required to obtain a reliable and stable estimate of hepatic steatosis percentage.

Table 5.9: Number of images per specimen at 10x magnification x 40x resolution

Specimen	**Number of Images**
1	6
2	4
3	7
4	6
5	5
6	5
7	5
8	6
9	5
10	6
11	6
12	6
13	5
14	5
15	5
16	6
17	5

This section is divided into two approaches: first, for each specimen and according to the number of images available, different combinations were tested by comparing the global median of all images with the median of each combination.The objective is to evaluate from which point the median begins to stabilize and to determine the minimum number of images required, considering that the number of images is directly related to the time needed for digitization and processing.

Accordingly, Velosa and Pestana (2008), when we are interested in selecting subsets of k elements from a total of n available elements, and the order of selection is not important, we are dealing with combinations of n elements taken k at a time ($n \geq k$). In this case, the selection is done without replacement:

$$^{n}C_{k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

This expression is valid when n and k are natural numbers.

Given that the available number of images in Table 5.9 varied between specimens, combinations were generated based on the total number of images per sample. Specifically, the following groupings were analysed: ${}^4C_2$–${}^4C_3$–${}^4C_4$; ${}^5C_2$–${}^5C_3$–${}^5C_4$–${}^5C_5$; ${}^6C_2$–${}^6C_3$–${}^6C_4$–${}^6C_5$–${}^6C_6$; and ${}^7C_2$–${}^7C_3$–${}^7C_4$–${}^7C_5$–${}^7C_6$–${}^7C_7$. These combinations simulate real-world situations where only a subset of images might be available for analysis due to time, cost, or operational constraints.

For each combination, the median percentage of steatosis per sample was calculated (represented in the plots by black circles). These were compared to the reference value, defined as the median of all images available for each specimen (highlighted as a red dashed line). The median was chosen instead of the mean because of its robustness against outliers, a common feature in biological measurements, thus providing a more stable central tendency indicator.

Secondly, based on the construction of replacement combinations, a new dataset was created using sets of 5, 6, and 7 images, sampled 3, 4, and 5 at a time, respectively. This analysis is limited to samples 3, 5, 6, 7, 9, 10, 14, and 17. These eight specimens, four classified as mild and four as severe, were selected to explore whether the degree of steatosis could influence the number of images required. The number of possible combinations with replacement was calculated using the formula:

$$ {}^n\mathrm{CR}_k = \binom{n+k-1}{k} $$

where n represents the total number of available images and k the size of the combination with replacement.

The resulting totals for each scenario were as follows:

- ${}^5C_3 = 35, \quad {}^5C_4 = 70, \quad {}^5C_5 = 126$
- ${}^6C_3 = 56, \quad {}^6C_4 = 126, \quad {}^6C_5 = 252$
- ${}^7C_3 = 84, \quad {}^7C_4 = 210, \quad {}^7C_5 = 462$

To further assess the reliability of these means, bootstrap resampling with 30, 50, 80 (5, 6, and 7 images, respectively) iterations for each of the 8 samples was conducted. In each bootstrap replicate, the mean steatosis percentage was calculated. These means were compared to the global mean obtained from all available images for each sample. This approach enabled us to evaluate the stability of the mean as a function of the number of images used. Additionally, empirical quartiles (2.5% and 97.5%) were obtained for each group of replicates to assess the estimated CI 95% associated with each combination size. Following the above analysis, 50 replicate bootstraps were run for each combination size (3, 4, and 5 images) per specimen. Using a fixed number of 50 replicates ensures consistent statistical analysis and allows for graphical comparisons across specimens with different numbers of available images, while avoiding bias introduced by differing total numbers of possible combinations. From these results, we calculated how often each combination size (3, 4, or 5 images) yielded the lowest variability and the closest estimate to the original mean. These frequencies are presented in the table below, providing insight into the stability of the mean and the potential trade-off between image quantity and processing effort.

It is important to note that sample 2 (which only had 4 available images), was exclude from the bootstrap analysis.. The results will provide a comprehensive overview of how the number of images affects the precision of steatosis quantification and can help guide the selection of an optimal imaging strategy that balances precision and efficiency.

5.4.1 Combinatorial Analysis of Image Subsets

The jitter plots are presented in ascending order based on the number of images available per specimen, starting from 4 up to 7 images per sample.

Analysis of Figure 5.21, corresponding to liver specimen 2 (with a total of 4 images), shows that the global median of steatosis percentage is 0.2082, represented by the red dashed line. Combinations ${}^4C_3$ and ${}^4C_4$ yield identical medians (0.208), which are essentially equal to the global value. Although combination ${}^4C_2$ shows a slightly higher median (0.213), the difference is minimal. These results suggest that using just 3 or 4 images may be sufficient to provide a stable estimate of steatosis percentage in this case. This may support a reduction in the number of images required for efficient analysis.

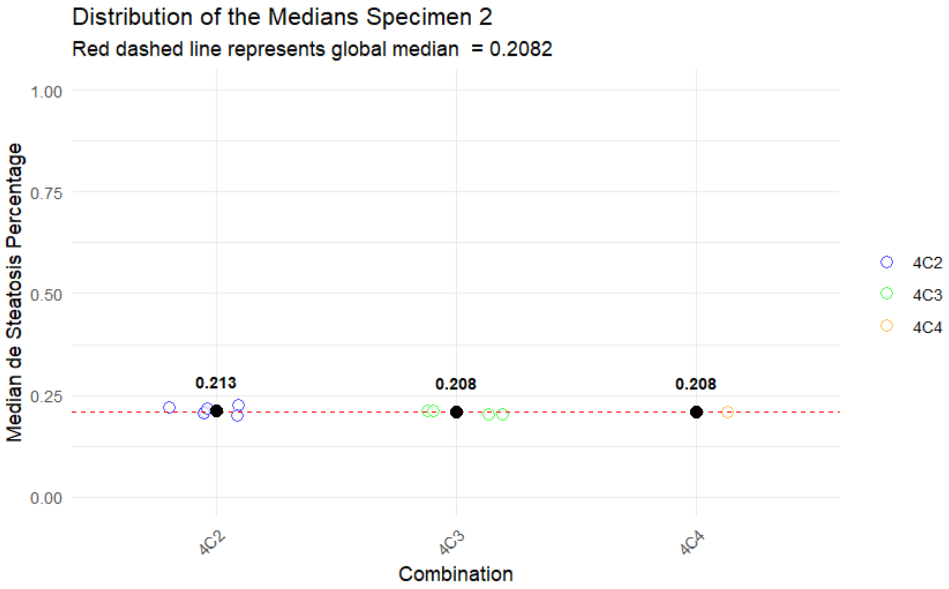
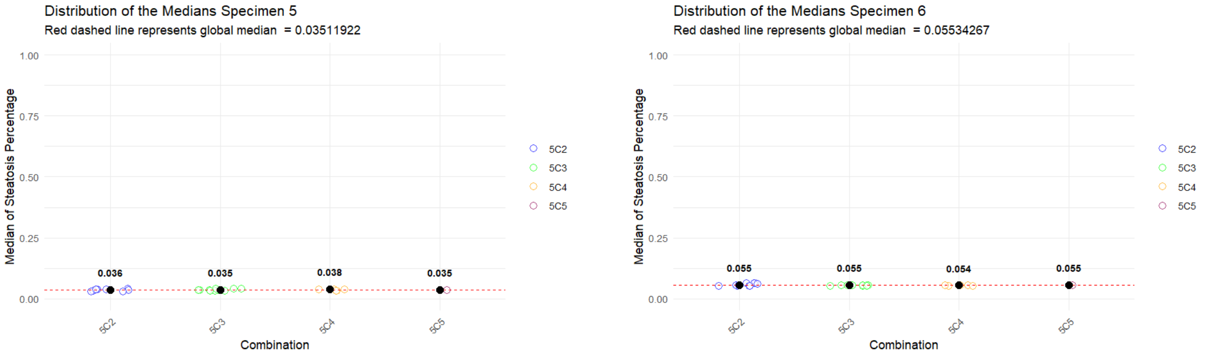


Figure 5.21: Distribution of the median steatosis percentages across different image combinations for specimen 2.

Analysis of Figure 5.22a, which refers to liver specimen 5 and the variable *steatosis percentage*, shows that the overall median is 0.0351 (represented by the red dashed line). All combinations (${}^5C_2$, ${}^5C_3$, ${}^5C_4$, and ${}^5C_5$) yield median values very close to this global median, suggesting high consistency across different image sets. The variation in median estimates is minimal, with values ranging from 0.035 to 0.038, indicating robust performance regardless of the number of images used. These results suggest that even reduced combinations, such as ${}^5C_3$ (three images), may be sufficient to produce reliable steatosis estimates in this specimen.

In Figure 5.22b, the coloured points represent the distribution of medians calculated from all possible image combinations of a given size, while the solid black circles indicate the median of each combination group. The red dashed line corresponds to the global median calculated from all available images for this specimen (0.0553). All combinations (${}^5C_2$, ${}^5C_3$, ${}^5C_4$, and ${}^5C_5$) yield median values that closely match the global median, ranging narrowly between 0.054 and 0.055. This tight clustering of values reflects high consistency and low variability across all image sets. These findings suggest that using a reduced number of images, such as three, may still provide reliable and accurate estimates for this specimen.



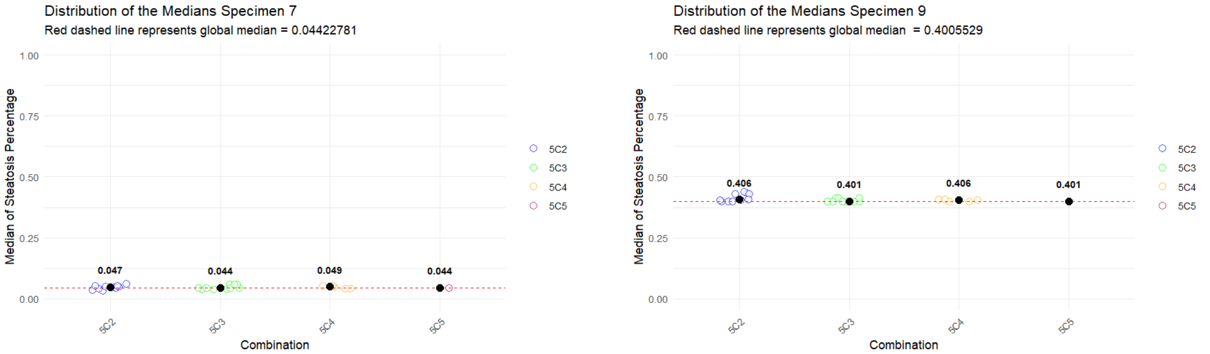
(a) Distribution of the median steatosis percentages across different image combinations for specimen 5.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 6.

Figure 5.22: Specimen 5 and 6 image analysis

Analysis of Figure 5.23a, which refers to liver specimen 7 and the percentage of steatosis, shows that the global median is 0.0442 (represented by the red dashed line). All combinations present tightly clustered median values, ranging from 0.044 to 0.049, indicating strong consistency across different image subsets. Notably, combinations ${}^5C_3$ and ${}^5C_5$ yield median values identical to the global median. These results support the idea that even reduced combinations, such as those using only 3 images, may be sufficient to obtain a reliable estimate of steatosis in this specimen.

Analysis of Figure 5.23b, which refers to liver specimen 9 and the percentage of steatosis, shows that the global median is 0.4006 (represented by the red dashed line). Combinations ${}^5C_3$ and ${}^5C_5$ yield median values that match the global median, while combinations ${}^5C_2$ and ${}^5C_4$ show slightly higher values around 0.406. Although the overall range of median values is narrow (0.401 to 0.406), the spread of individual estimates in combination ${}^5C_2$ appears larger, indicating higher variability in that group. In contrast, ${}^5C_3$ and ${}^5C_5$ demonstrate more concentrated values around the global median. These results suggest that, for this specimen, using at least 3 images can provide reliable estimates, particularly when the image subset is representative and balanced.



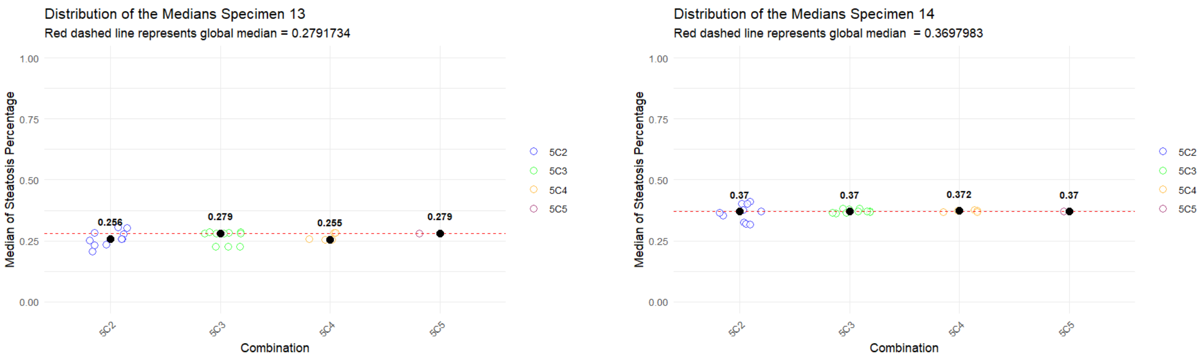
(a) Distribution of the median steatosis percentages across different image combinations for specimen 7.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 9.

Figure 5.23: Specimen 7 and 9 image analysis

Analysis of Figure 5.24a, which refers to liver specimen 13 and the percentage of steatosis, shows that the global median is 0.2791 (represented by the red dashed line). Combinations ${}^5C_3$ and ${}^5C_5$ yield median values that exactly match the global median, while combinations ${}^5C_2$ and ${}^5C_4$ exhibit slightly lower values of 0.256 and 0.255, respectively. Although combination ${}^5C_3$ appears to be a good indicator of the steatosis percentage, its median values are more dispersed compared to ${}^5C_4$. This suggests that 4 images may offer a more stable and consistent estimate in this case.

In Figure 5.24b, the dashed red line represents the global median obtained using all available images for this specimen (median =0.3698). All combinations exhibit median values that are very close to the global median, indicating strong consistency across subsets. Although combination ${}^5C_2$ shows slightly more dispersed values, its central estimate remains aligned with the overall median. Considering the time and resources required for image acquisition and processing, the use of 3 images appears to offer a good trade-off between analytical effort and estimation accuracy.



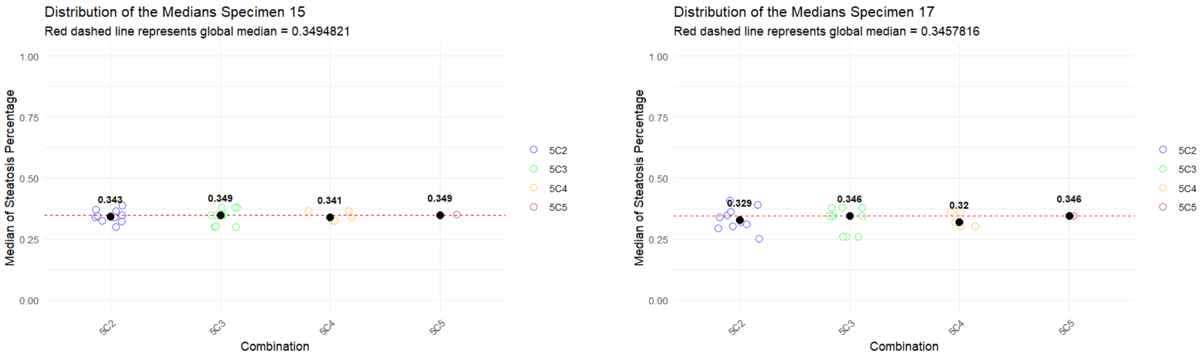
(a) Distribution of the median steatosis percentages across different image combinations for specimen 13.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 14.

Figure 5.24: Specimen 13 and 14 image analysis

Analysis of Figure 5.25a, which refers to liver specimen 15 and the percentage of steatosis, shows that the global median is 0.3494 (represented by the red dashed line). Combinations ${}^5C_3$ and ${}^5C_5$ yield median values that exactly match the global median, while combinations ${}^5C_2$ and ${}^5C_4$ produce values that are very close to it. Among these, combination ${}^5C_4$ appears to be a strong indicator of the steatosis percentage, suggesting that using 4 images may provide a reliable and efficient estimate in this case.

In Figure 5.25b, the dashed red line represents the global median obtained using all available images for this specimen (median =0.3458). Combinations ${}^5C_3$ and ${}^5C_5$ yield median values that closely match the global median, while ${}^5C_2$ and ${}^5C_4$ display slightly lower values (0.329 and 0.320, respectively). The dispersion of medians is similar across the different image combinations, although ${}^5C_5$ appears slightly more concentrated around the central value. These results suggest that, although three images can provide a reasonable estimate, using five images may improve consistency in this specimen.



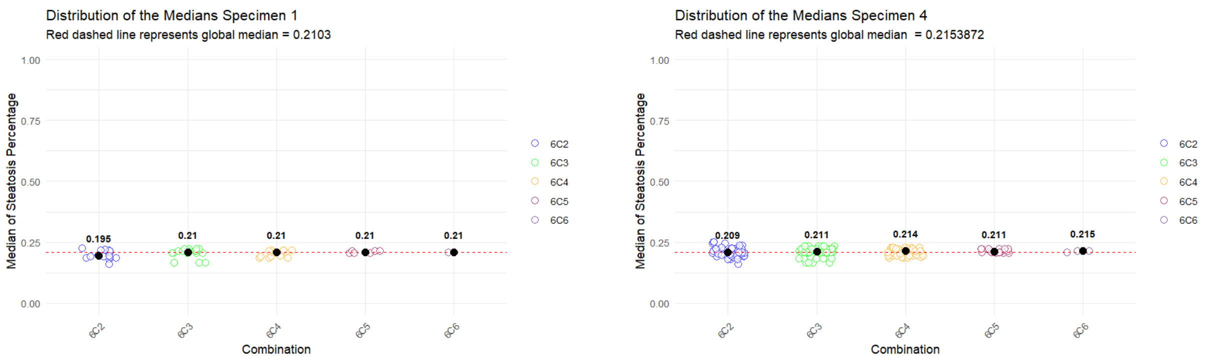
(a) Distribution of the median steatosis percentages across different image combinations for specimen 15.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 17.

Figure 5.25: Specimen 15 and 17 image analysis

In Figure 5.26a, the coloured points show the distribution of medians from all possible combinations of a given size, while the solid black circles indicate the overall median for each group. The dashed red line represents the global median obtained using all available images for this specimen (0.2103). Combination $^6C_2$ exhibits greater variability, with median values fluctuating more widely and a group median of 0.195, which is below the global median. This suggests that combinations using only two images are less stable and tend to underestimate the true value. From combination $^6C_4$ onwards, the group medians converge rapidly toward the global median (all equal to 0.210), indicating increased stability and accuracy with a larger number of images. Therefore, using three images appears to offer a good trade-off between accuracy and analytical effort.

Analysis of Figure 5.26b, corresponding to specimen 4, shows the median percentage of steatosis across different image combinations ($^6C_2$ to $^6C_6$). The red dashed line represents the global median (0.2154). All group medians are close to the global value, indicating a high degree of consistency. Among them, combination $^6C_6$ yields a median almost identical to the global one. However, combination $^6C_4$ also performs well and demonstrates low variability, suggesting that using 4 images may offer a good compromise between analytical accuracy and effort.



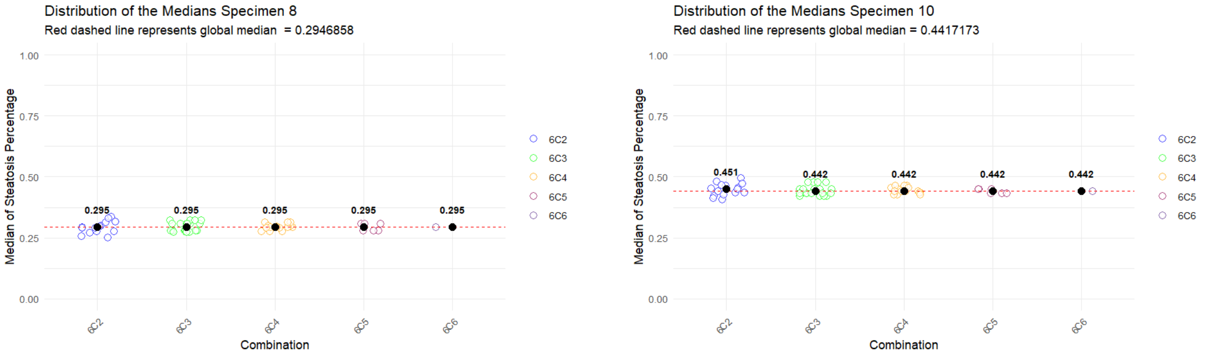
(a) Distribution of the median steatosis percentages across different image combinations for specimen 1.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 4.

Figure 5.26: Specimen 1 and 4 image analysis

In Figure 5.27a, the coloured points represent the distribution of medians from all possible combinations of a given size, while the solid black circles indicate the overall median for each combination group. The dashed red line corresponds to the global median calculated using all available images for this specimen (median = 0.2947). All combinations yield median values that closely match the global median, indicating strong consistency across image subsets. Notably, the $^6C_4$ group (combinations of 4 images) shows the most concentrated distribution around the central value, suggesting that this configuration may offer an optimal balance between accuracy and efficiency. These results confirm that the steatosis median estimates for specimen 8 are highly robust, with 4-image combinations standing out in terms of stability.

Analysis of Figure 5.27b, related to specimen 10, shows the median percentage of steatosis across different image combinations. The red dashed line represents the global median (0.4417). The median values for all combinations—except $^6C_2$ are nearly identical to the global median, indicating strong consistency. Combination $^6C_2$ presents a slightly higher group median (0.451) and greater variability. From combination $^6C_4$ onward, the estimates stabilize, suggesting that using 4 images already provides a reliable and efficient assessment of steatosis percentage.



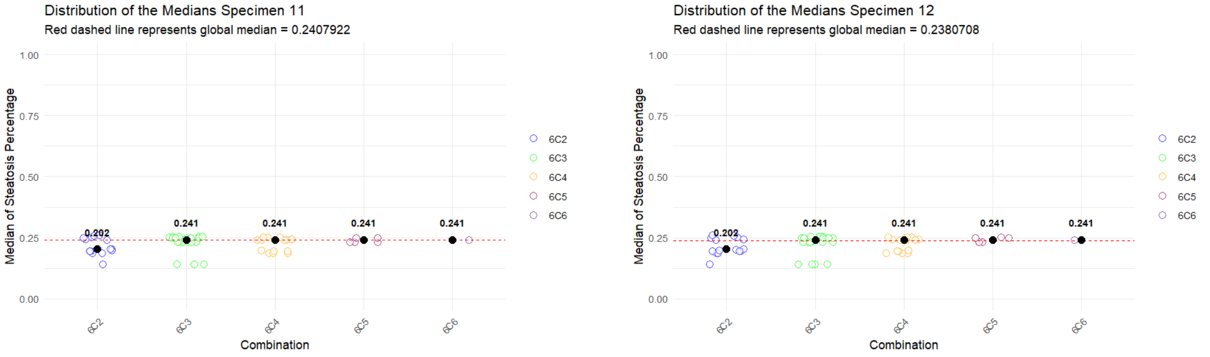
(a) Distribution of the median steatosis percentages across different image combinations for specimen 8.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 10.

Figure 5.27: Specimen 8 and 10 image analysis

Analysis of Figure 5.28a, related to specimen 11, shows the median percentage of steatosis across different image combinations. The red dashed line represents the global median (0.2407). The median values for all combinations, except ${}^6C_2$, are nearly identical to the global median, indicating strong consistency. Combination ${}^6C_2$ presents a slightly lower group. From combination ${}^6C_4$ onwards, the estimates stabilize, suggesting that using 4 images already provides a reliable and efficient assessment of steatosis percentage.

As observed previously in Figure 5.28a, Figure 5.28b, related to specimen 12, shows the median percentage of steatosis across different image combinations. The red dashed line represents the global median (0.238). The median values for all combinations, except ${}^6C_2$, are nearly identical to the global median, slightly above it, indicating strong consistency. From combination ${}^6C_4$ onward, the estimates stabilize, suggesting that using 4 images already provides a reliable and efficient assessment of steatosis percentage.



(a) Distribution of the median steatosis percentages across different image combinations for specimen 11.

(b) Distribution of the median steatosis percentages across different image combinations for specimen 12.

Figure 5.28: Specimen 11 and 12 image analysis

Figure 5.29, related to specimen 16, shows the median percentage of steatosis across different image combinations. The red dashed line represents the global median (0.3208). The median values for all combinations are identical, indicating a high degree of consistency. From combination ${}^6C_3$ onwards, the estimates remain stable, suggesting that using 3 images already provides a reliable and efficient assessment of steatosis percentage.

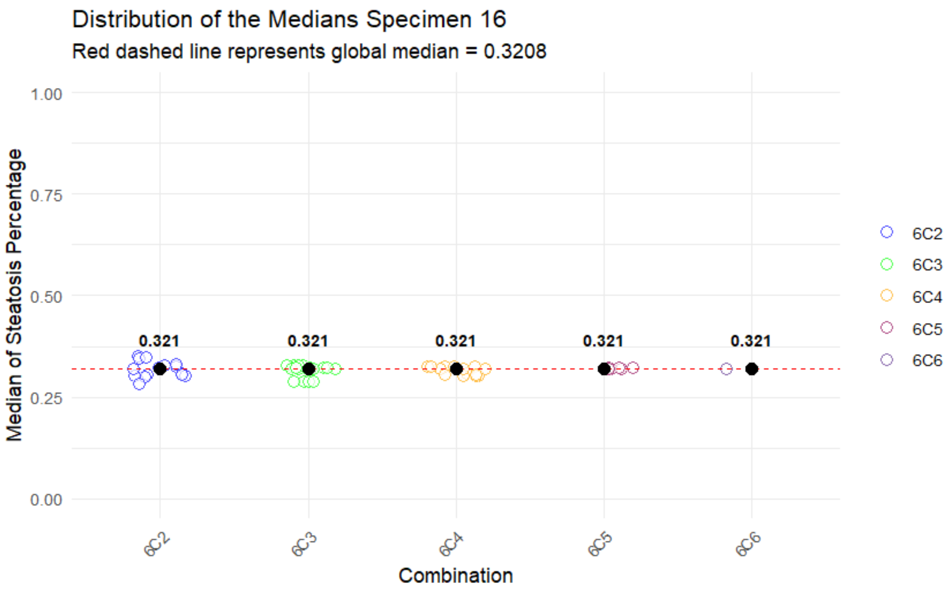


Figure 5.29: Distribution of the median steatosis percentages across different image combinations for specimen 16.

In Figure 5.30, the red dashed line represents the global median obtained for specimen 3 (0.0923). A total of six combinations were performed to evaluate the median steatosis percentage across different image selection strategies. Combinations ${}^7C_3$, ${}^7C_5$, and ${}^7C_7$ yielded median values that match or closely approximate the global median, suggesting that these subsets are representative of the overall sample. In contrast, combinations ${}^7C_4$ and ${}^7C_6$ resulted in lower medians, which may indicate a limited representation of the full variability in the specimen. Combination ${}^7C_2$ shows the greatest deviation and dispersion, implying that subsets with only two images may not adequately capture the overall distribution. These findings support the idea that using 4 images (${}^7C_4$) provides a good trade-off between accuracy and analytical efficiency.

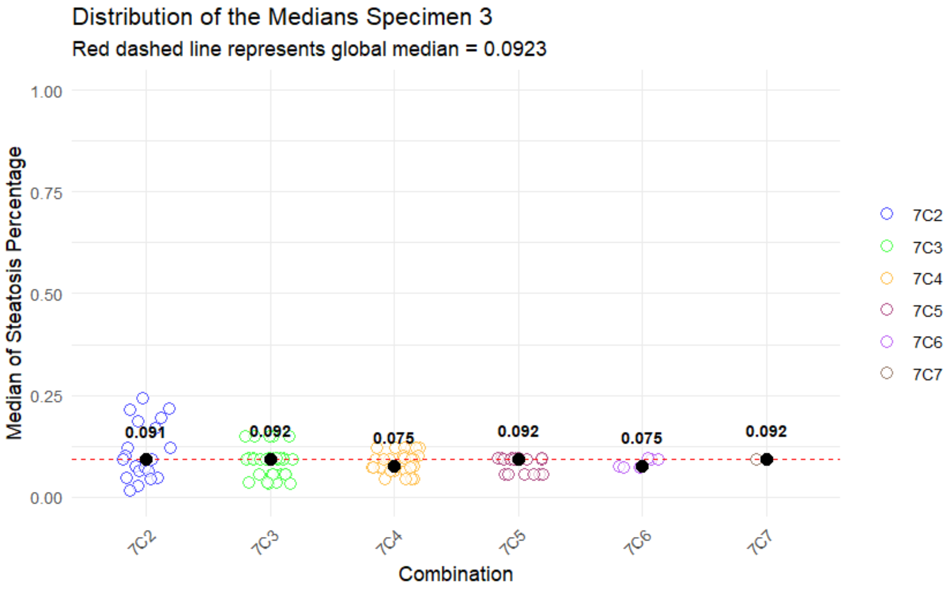


Figure 5.30: Distribution of the median steatosis percentages across different image combinations for specimen 3.

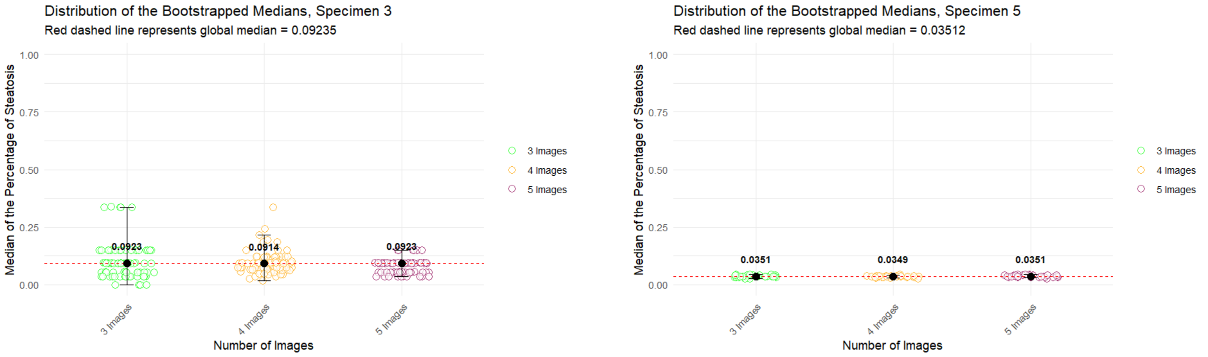
Across all 17 NAFLD specimens with 4, 5, 6, and 7 available images, the median steatosis percentage calculated from different image combinations reveals a consistent pattern. In most cases, subsets using only 3 images yielded median values very close to the global median derived from all images. Although combinations using more images (such as 4 or 5) occasionally resulted in slightly more stable or exact estimates, the added gain was often minimal. Interestingly, even in specimens with a larger number of images and therefore more potential combinations, the results from 3 image subsets remained robust. This supports the findings that selecting 3 representative images may offer a reliable and efficient approximation of the true steatosis percentage. From a practical perspective, this approach helps reduce the time and resources needed for image acquisition and analysis, while maintaining sufficient accuracy for biological interpretation and decision-making.

5.4.2 Bootstrap Analysis

The following analysis focuses on distinguishing two groups based on steatosis severity: mild and severe. To support this comparison, bootstrap resampling was applied to estimate the variability and stability of steatosis percentage across different image combinations. To ensure reproducibility of the bootstrap resampling, a fixed random seed (`set.seed (123)`) was used before sampling. The resulting graphs below illustrate how the number of images used influences the robustness of the estimates within each group.

Figure 5.31a displays the bootstrap results for specimen 3 (7 images), based on 80 replicates, showing the median percentages of steatosis calculated for combinations with replacement of 3, 4, and 5 images. The red dashed line represents the global median derived from all available images for this specimen (0.09235). Each black circle represents the median of the bootstrapped values corresponding to each combination size. As the number of images increases, the medians tend to converge towards the global median. For instance, with 5 images, the median is 0.0923, very close to the global median. The dispersion of individual medians is generally narrow across all groups; however, 4-image combinations show slightly more extreme values, while 5-image combinations show the least variability. These results support the idea that using more images improves the precision of the steatosis estimate. Thus, 5 images appear to offer the most robust and consistent assessment. Nevertheless, the decision must consider practical aspects, such as the time and cost associated with acquiring and processing more images, which may justify opting for fewer images in some scenarios.

Figure 5.31b presents the bootstrap results for specimen 5, using 30 replicates for combinations of 3, 4, and 5 images. The red dashed line shows the global median of all images (0.0351). As before, black circles mark the median per group. The median for 3 images is slightly below the global median (0.0351), while the medians for 4 and 5 images are slightly above it (0.0349 and 0.0351, respectively). Notably, the confidence interval is widest for 3 images, reflecting greater variability. In contrast, 5 images yield the narrowest confidence interval, indicating greater precision. These findings confirm the trend that increasing the number of images improves reliability. The difference between using 4 or 5 images is minimal, so 4 images may be a reasonable trade-off between accuracy and effort. Although 3 images still provide acceptable estimates, they come with higher variability, which might limit reliability in more sensitive analyses.



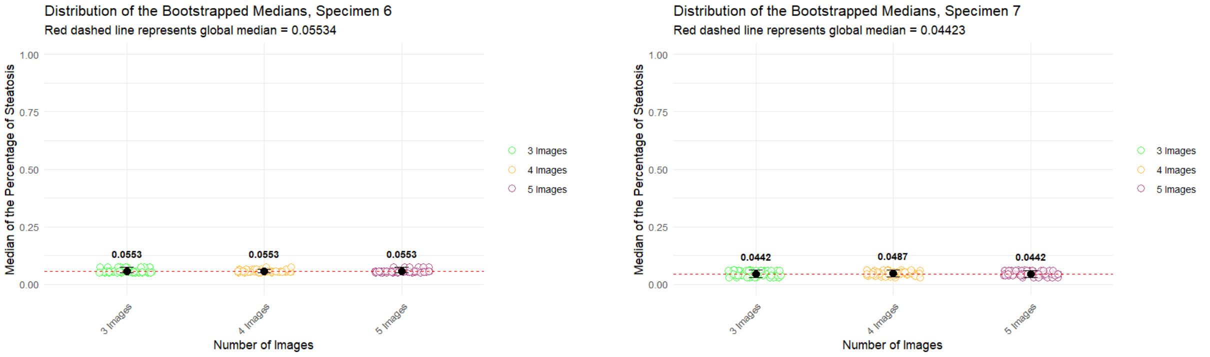
(a) Bootstrapped median values of the percentage steatosis for Specimen 3. Each point represents one bootstrap replicate (n = 80). Black dots indicate the meadian per group, with vertical lines showing the 95% confidence intervals.

(b) Bootstrapped median values of the percentage steatosis for Specimen 5. Each point represents one bootstrap replicate (n = 30). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

Figure 5.31: Bootstrapped median values of the percentage steatosis for specimens 3 and 5.

Figure 5.32a displays the bootstrap results for specimen 6, based on 30 replicates, showing the median percentage of steatosis obtained from combinations of 3, 4, and 5 images. The red dashed line represents the global median calculated from all available images for this specimen (0.05534). Each black circle indicates the median of the bootstrapped estimates for each combination size. The medians for both 3 and 5 images are very close to the global median, indicating good precision. All combinations exhibit similar confidence interval ranges, although the 4-image group shows a few more extreme values, suggesting slightly greater point-wise dispersion. These results suggest that using 3 images provides a reliable and efficient estimate of steatosis percentage. While 5 images offer slightly more consistency, the minimal differences do not justify the additional effort in many practical scenarios. Therefore, 3 images appear to be a reasonable trade-off between precision and resource use.

Figure 5.32b presents the bootstrap results for specimen 7, using 30 replicates for combinations of 3, 4, and 5 images. The red dashed line indicates the global median calculated from all available images (0.04423). As in previous cases, black circles represent the median of the bootstrap estimates for each image set. The medians for 3 (0.0442), 4 (0.0487), and 5 (0.0442) images are all close to the global median, indicating good overall precision. Among them, the 4-image group offers a slightly better balance between accuracy and confidence interval. However, confidence intervals remain relatively wide across all groups. Although small differences exist, the performance of the 3-image group is comparable to the others.



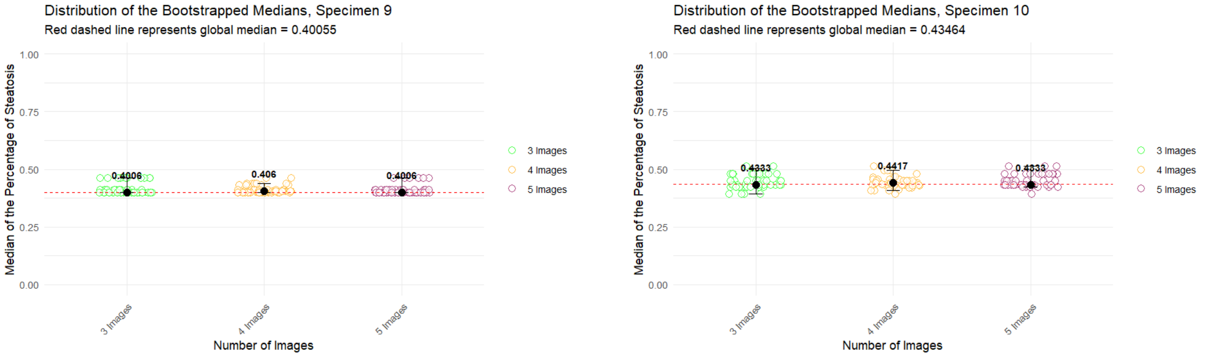
(a) Bootstrapped median values of the percentage steatosis for Specimen 6. Each point represents one bootstrap replicate (n = 30). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

(b) Bootstrapped median values of the percentage steatosis for Specimen 7. Each point represents one bootstrap replicate (n = 30). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

Figure 5.32: Bootstrapped median values of the percentage steatosis for specimens 6 and 7.

Figure 5.33a displays the bootstrap results for specimen 9 (5 images), based on 30 replicates, showing the median percentage of steatosis calculated from combinations of 3, 4, and 5 images. The red dashed line represents the global median derived from all available images for this specimen (0.40055). The median values for all image sets are close to the global median, indicating good central tendency alignment. However, the dispersion appears greater in the 4-image group, suggesting slightly less stability compared to the 3- and 5-image groups. While using 4 or 5 images may offer more robustness in some cases, the small difference in median between 3 and 5 images (both 0.4006) supports the use of only 3 images as a time-efficient alternative with acceptable accuracy.

Figure 5.33b presents the bootstrap results for specimen 10 (5 images), using 30 replicates for combinations of 3, 4, and 5 images. The red dashed line represents the global median obtained from all available images (0.43464). The median values for the combinations of 3 (0.4333), 4 (0.4417), and 5 images (0.4333) are all very close to the global median, indicating good precision across all scenarios. The dispersion is slightly greater in the 4-image group, as seen by the wider spread of values compared to the 3- and 5-image groups, which show more compact distributions. Despite this, the differences between the groups are minimal, suggesting that using 3 images is a reasonable and efficient option for estimating steatosis in this specimen.



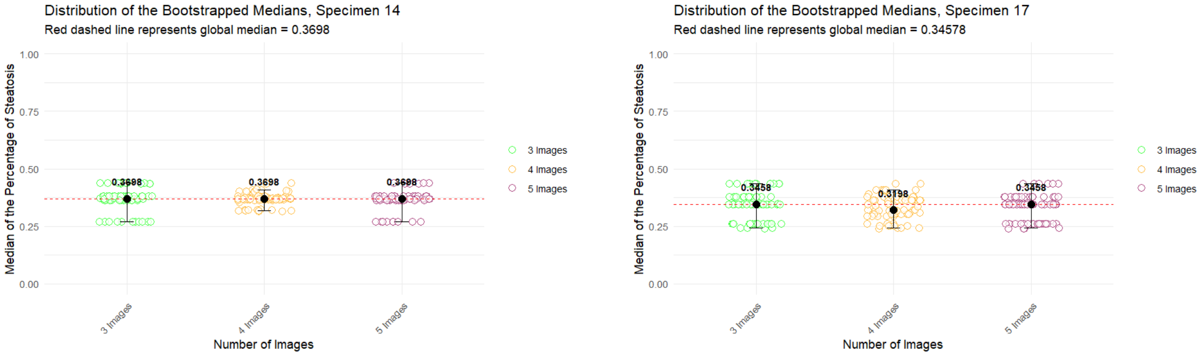
(a) Bootstrapped median values of the percentage steatosis for Specimen 9. Each point represents one bootstrap replicate (n = 30). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

(b) Bootstrapped median values of the percentage steatosis for Specimen 10. Each point represents one bootstrap replicate (n = 50). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

Figure 5.33: Bootstrapped median values of the percentage steatosis for specimens 9 and 10.

Figure 5.34a displays the bootstrap results for specimen 14, based on 30 replicates, showing the median percentage of steatosis calculated from combinations of 3, 4, and 5 images. The red dashed line represents the global median derived from all available images for this specimen (0.3698). The medians obtained for all image sets (3, 4, and 5) are identical or extremely close to the global median (0.3698), indicating excellent consistency and central tendency. While all groups display relatively low dispersion, the 5-image group appears slightly more compact, suggesting marginally lower variability and potentially greater robustness in the estimate.

Figure 5.34b presents the bootstrap results for specimen 17, using 30 replicates for combinations of 3, 4, and 5 images. The red dashed line indicates the global median calculated from all available images (0.34578). The results show that the medians for 3 (0.3456) and 5 images (0.3458) are nearly identical to the global median, while the 4-image group shows a slightly higher value (0.37198). Regarding variability, the 3-image group displays the widest dispersion, suggesting less stability in the estimate. The 5-image group has the narrowest spread, indicating improved precision with more images. Overall, while 4 and 5 images yield similarly robust results, the minimal difference between the 3- and 5-image medians supports the idea that using 3 images may still offer acceptable accuracy, though with slightly higher variability.



(a) Bootstrapped median values of the percentage steatosis for Specimen 14. Each point represents one bootstrap replicate (n = 30). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

(b) Bootstrapped median values of the percentage steatosis for Specimen 17. Each point represents one bootstrap replicate (n = 30). Black dots indicate the median per group, with vertical lines showing the 95% confidence intervals.

Figure 5.34: Bootstrapped median values of the percentage steatosis for specimens 14 and 17.

Following the previous findings and aiming to replicate them, a bootstrap procedure was repeated 50 times to assess the variability and mean accuracy using the same sets of images (3, 4, and 5). The table 5.10 presents the summarized results. Across all specimens, using 5 images most frequently resulted in lower variability, confirming earlier observations that increasing the number of images generally improves precision and accuracy. This is followed by 4 images, while 3 images tend to show higher variability. However, 3-image combinations often performed well in terms of mean estimation, indicating that although they are more variable, they can still yield accurate central estimates. When analysing the results by the degree of steatosis, specimens 3, 5, 6, and 7 (mild steatosis) suggest that using fewer images may be acceptable. In contrast, for specimens with severe steatosis (9, 10, 14, and 17), using at least 4 images is advisable, with 5 images providing the best precision and consistency.

Table 5.10: Frequency of Lowest Variability and Closest Median to the True Value Across Bootstrap 50 Replicates (k = 3, 4, 5). Variability inferred from Bootstrap 95% CI, and the true value corresponds to the global median.

Specimen	Criterion	k = 3	k = 4	k = 5
Specimen 3	Lowest variability	4	13	33
Specimen 3	Closest median	48	0	2
Specimen 5	Lowest variability	1	33	16
Specimen 5	Closest median	47	0	33
Specimen 6	Lowest variability	7	20	23
Specimen 6	Closest median	41	5	4
Specimen 7	Lowest variability	2	37	11
Specimen 7	Closest median	46	0	4
Specimen 9	Lowest variability	10	20	20
Specimen 9	Closest median	41	1	8
Specimen 10	Lowest variability	5	23	22
Specimen 10	Closest median	19	29	2
Specimen 14	Lowest variability	3	19	28
Specimen 14	Closest median	46	1	3
Specimen 17	Lowest variability	2	37	11
Specimen 17	Closest median	45	2	3

Table 5.11: Bootstrap results for all specimens: standard error (SE), and 95% confidence interval by number of images.

Specimen	k	SE	CI_inf	CI_sup
Specimen 3	3	0.00866	7.46×10^{-7}	0.33726
Specimen 3	4	0.00574	0.01675	0.21705
Specimen 3	5	0.00365	0.03437	0.14844
Specimen 5	3	0.00056	0.02717	0.04278
Specimen 5	4	0.00041	0.03016	0.04266
Specimen 5	5	0.00058	0.02717	0.04278
Specimen 6	3	0.00088	0.05038	0.07504
Specimen 6	4	0.00055	0.05155	0.06582
Specimen 6	5	0.00065	0.05038	0.07504
Specimen 7	3	0.00127	0.02862	0.06327
Specimen 7	4	0.00091	0.03301	0.06148
Specimen 7	5	0.00109	0.02862	0.05978
Specimen 9	3	0.00250	0.39937	0.46372
Specimen 9	4	0.00163	0.39942	0.43879
Specimen 9	5	0.00190	0.39937	0.46372
Specimen 10	3	0.00406	0.39201	0.50536
Specimen 10	4	0.00341	0.40734	0.49633
Specimen 10	5	0.00384	0.42266	0.51275
Specimen 14	3	0.00513	0.27025	0.43690
Specimen 14	4	0.00285	0.31694	0.40975
Specimen 14	5	0.00432	0.27025	0.43690
Specimen 17	3	0.00675	0.24147	0.43527
Specimen 17	4	0.00567	0.24147	0.40768
Specimen 17	5	0.00645	0.24147	0.43527

The bootstrap analysis above for all eight specimens reveals that the number of images used significantly influences the precision of steatosis percentage estimates, and this effect may be modulated by the degree of steatosis. In specimens with mild steatosis (Specimens 3, 5, 6, and 7), the use of five images most frequently results in the lowest variability and estimates closest to the global median. However, the difference is not always substantial, and in some specimens (e.g., 3 and 5), three images also yield acceptable performance. This suggests that a reduced number of images may be sufficient for mild cases, particularly when time or resource constraints are relevant. In specimens with more severe steatosis (Specimens 9, 10, 14, and 17), the benefits of using a higher number of images become clearer. Although estimates tend to remain close to the global median regardless of k, lower k values often result in more dispersed variability assignments across replications. Using at least four images consistently improves the stability of estimates, and five images generally offer the most robust results. These findings suggest that the degree of steatosis may influence the optimal number of images required. Nevertheless, using only three images appears to be a viable compromise in many contexts, especially for mild cases. This approach should be further validated in future analyses, ideally incorporating additional clinical or imaging variables from the dataset.

Analysing Table 5.11, we observe that for almost all specimens the standard error (SE) decreases as the number of images (k) increases. This indicates that including more images reduces the variability of the median estimate, leading to more stable and reliable results. Similarly, the amplitude of the 95% confidence interval also narrows with higher k values, suggesting increased precision in the estimation of steatosis percentage. There appears to be a distinction based on the degree of steatosis. In specimens with mild steatosis (specimens 5, 6, and 7), the SE values are already low with just three images, and the confidence intervals are narrow and consistent across k. This aligns with the results in Table 5.10, where three images were frequently sufficient to obtain accurate and precise estimates in mild cases. In contrast, specimens with more severe steatosis (specimens 9, 10, 14, and 17) show higher SE and wider confidence intervals when only three images are used. In these cases, using four or five images significantly reduces the uncertainty of the median estimate. Therefore, in severe cases, a higher number of images is recommended to ensure robust and reliable assessment of steatosis. These findings support the idea that the optimal number of images may depend on the degree of steatosis. Although three images may be sufficient in mild cases, at least four, preferably five, images should be used for more advanced steatosis to improve precision and reduce uncertainty of the estimation.

5.5 Quantitative Evaluation of Vacuole Size in Mild and Severe Hepatic Steatosis

The objective of this analysis is to determine whether the distribution of vacuole size follows a specific morphological pattern and whether the percentage of steatosis influences this distribution.

To investigate this, eight liver samples were selected—four from each category—based on the pathology-assigned steatosis scores: score 1 (mild steatosis) and score 4 (severe steatosis). The corresponding steatosis percentages, quantified using the Weka segmentation program at 10× magnification and 40× resolution, ranged between 0-15% for mild cases and 24-51% for severe cases.

This analysis was made possible through the use of quantitative image analysis tools, which provide detailed measurements, particularly the total area and number of vacuoles in each image.

To assess vacuole morphology, for each image of the selected samples, the ratio between the total vacuole area (sum of all individual vacuole areas) and the total number of vacuoles was calculated. This ratio serves as a proxy for the average vacuole size within each sample.

To hypothesize that liver samples with a higher degree of steatosis, reflected by both pathology score and steatosis percentage-will exhibit larger average vacuole sizes compared to those with lower steatosis levels.

To address this hypothesis, the mean of the ratio between vacuole area and vacuole number was calculated for each specimen.

Figure 5.35 shows the distribution of the ratio between vacuole area and vacuole number for all images within each specimen. It highlights that certain specimens, such as 3, 5, 6, and 14, exhibit low intra-specimen variation, suggesting that the measurements are consistent across different images from the same liver sample. In contrast, other specimens show greater variability, indicating that the vacuole area may differ between images of the same sample. This variation appears to be more prominent in specimens with higher degrees of steatosis (Severe), potentially reflecting morphological heterogeneity at more advanced stages of fat accumulation.

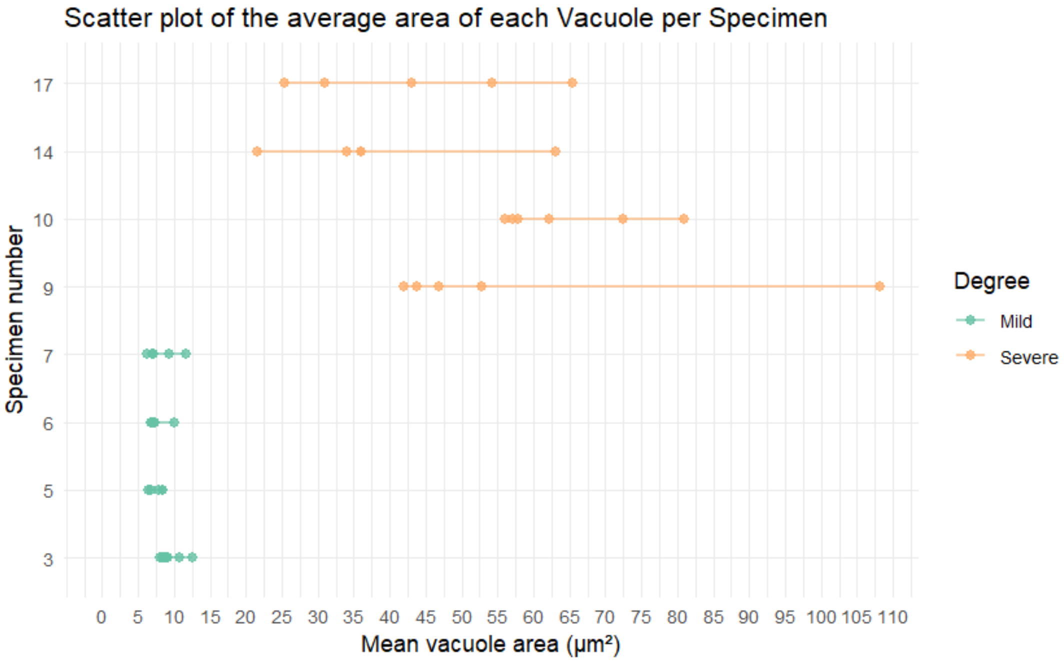


Figure 5.35: Scatter plot showing the average area of each vacuole (in µm²) across individual specimens. Each dot represents an image, coloured by steatosis severity (Ratio: Mild vs Severe). The horizontal distribution reflects vacuole size, and the vertical axis distinguishes specimens. Specimens with Severe steatosis consistently show higher vacuole area values compared to those with Mild steatosis.

To summarise, these findings reinforce the hypothesis that vacuole enlargement is associated with higher grades of steatosis, both in terms of pathology-assigned severity and steatosis percentage. Further analyses could include evaluating the distribution and variance of vacuole sizes across samples, which may provide additional insights into the progression and characteristics of steatosis at different severity levels.

Chapter 6

Conclusion

This work had three main objectives related to the evaluation of two free, open-source automated image analysis plugins (Weka and Saturation) for the quantification of hepatic steatosis in NAFLD using experimental mouse models.

The first objective was to determine whether there are differences between the plugins in terms of the number of lipid vacuoles detected per image. A strong positive monotonic correlation was found between the number of vacuoles identified by Weka and Saturation, indicating consistency. However, the Wilcoxon test revealed significant differences across combinations of magnification and resolution. To explore the differences between the plugins, three independent variables were analysed (Program, Magnification, and Resolution), against the dependent variable: the number of vacuoles. A GLMM with a Negative Binomial distribution, it was shown that both plugin and image settings significantly affect vacuole detection. The GLMM indicated that Weka tends to identify more vacuoles, particularly at higher magnification and resolution levels, though the interaction between plugin and image acquisition settings was heterogeneous.

The second objective was to evaluate the minimum number of images required per specimen for a reliable estimate of the percentage of steatosis. Based on the previous results, the combination of the Weka plugin, 10× magnification, and 40× resolution was selected as the reference setting. Although differences in accuracy were minimal, results indicated that using more than 3 images slightly improved accuracy. However, due to time and cost constraints, 3 images per specimen provide a reasonable trade-off between precision and efficiency.

The third objective aimed to explore whether a relationship exists between the degree of steatosis and the size of lipid vacuoles. To address the limitations of small sample size, bootstrap resampling was applied and repeated 50 times to generate confidence intervals for the vacuole area in mild vs. severe steatosis. The results suggest that steatosis severity may influence the optimal number of images required and that larger vacuoles are associated with more severe steatosis. This supports the hypothesis of a morphological progression pattern.

This work contributes to the validation and optimization of automated tools for steatosis quantification in preclinical NAFLD studies. Both Weka and Saturation are free, user-friendly plugins with extensive community support, integrated into FIJI (ImageJ). Standardized image acquisition conditions (Magnification and Resolution) and a consistent number of images per sample are critical factors in improving accuracy and reproducibility, reducing inter- and intra-observer variability typical of manual, pathologist-dependent assessments.

While this study focused on mice with different grades of NAFLD, future work could expand the sample to include healthy liver specimens, thereby enhancing external validity. Additionally, increasing the number of specimens would allow for the implementation of a holdout validation strategy, splitting the data into training and test sets, to adjust and evaluate model performance more accurately. Moreover, integrating clinical and demographic variables such as age, sex, and biochemical markers, currently absent from this dataset, would allow for a more comprehensive understanding of disease patterns and enable the use of survival analysis. Additionally, it may be valuable to develop a model that combines biochemical data with the percentage of hepatic steatosis to obtain more reliable and clinically meaningful predictions.

A limitation of this work was the inability to compare automated steatosis percentages with the semi-quantitative steatosis scores typically assigned by pathologists, due to missing annotations in the dataset. Despite the use of bootstrap techniques, the relatively small number of images per specimen limits generalizability and may introduce bias through repeated sampling.

This study builds on previous work, including Forlano et al., 2020, which reported high agreement (ICC = 0.97) between automated and manual assessments, and Munsterman et al., 2019, which developed a macro-based vacuole detection algorithm. It also aligns with findings from Homeyer et al., 2017, confirming that 10× magnification is both faster and sufficiently precise for digital pathology workflows. Moreover, while some previous studies noted that vacuole size may carry diagnostic value, this work took a step further by quantifying and linking vacuole size directly to disease grade, representing an innovative contribution to the field. It is important to highlight that, in recent years, there has been an increasing use of deep learning and machine learning tools in digital pathology. The quantification of metabolic dysfunction–associated steatotic liver disease (MASLD) has global importance, as the number of related publications grows steadily. For example, in their study, Farzi et al. (2025) introduced *Liver-Quant*, an open-source Python package designed to quantify fat (Steatosis Proportionate Area) and fibrosis (Collagen Proportionate Area) in whole-slide liver images.

In conclusion, this study validates the feasibility of using open-source automated tools to quantify hepatic steatosis in experimental NAFLD models. It proposes a statistically robust and reproducible pipeline that can reduce subjectivity and improve efficiency. Future research could explore the clinical applicability of these tools in human liver biopsies, potentially aiding early diagnosis, staging, and treatment decisions, and including transplant eligibility. These contributions are particularly relevant in the global effort to reduce the burden of NAFLD through scalable, objective, and cost-effective approaches. To promote reproducibility and transparency, the complete image analysis pipeline developed in this study, including preprocessing scripts and statistical models, will be made publicly available on GitHub.

References

Ali, A., Shaker, A.-L., Abd, Y., & Ghallab, A. (2019). Anatomy of the mouse liver [Figura extraída da tese de Aya Ali].

Angulo, P. (2002). Nonalcoholic fatty liver disease. *New England Journal of Medicine*, *346*(16), 1221–1231. https://doi.org/10.1056/NEJMra011775

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K. W., Schindelin, J., Cardona, A., & Seung, H. S. (2017). Trainable weka segmentation: A machine learning tool for microscopy pixel classification. *Bioinformatics*, *33*(15). https://doi.org/10.1093/bioinformatics/btx180

Bancroft, J. D., & Gamble, M. (2008). *Theory and practice of histological techniques* (6th ed.). Churchill Livingstone.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Brunt, E. M. (2010). Pathology of nonalcoholic fatty liver disease. *Nature Reviews Gastroenterology Hepatology*, *7*, 195–203. https://doi.org/10.1038/nrgastro.2010.21

Brunt, E. M., Janney, C. G., Di Bisceglie, A. M., Neuschwander-Tetri, B. A., & Bacon, B. R. (1999). Nonalcoholic steatohepatitis: A proposal for grading and staging the histological lesions. *American Journal of Gastroenterology*, *94*(9), 2467–2474. https://doi.org/10.1111/j.1572-0241.1999.01377.x

Brunt, E. M., & Tiniakos, D. G. (2010). Histopathology of nonalcoholic fatty liver disease. *World Journal of Gastroenterology*, *16*(42), 5286–5296. https://doi.org/10.3748/wjg.v16.i42.5286

Brunt, E. M. (2005). Pathology of nonalcoholic steatohepatitis. *Hepatology Research*, *33*(2), 68–71. https://doi.org/10.1016/j.hepres.2005.09.006

Cai, D., & Zhou, Y. (2016). 3d microscopy and imaging in life sciences using fiji. *Nature Protocols*, *11*(6), 998–1009. https://doi.org/10.1038/nprot.2016.070

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2nd ed.). Cambridge University Press.

Chalasani, N., Younossi, Z. M., Lavine, J. E., Charlton, M., Cusi, K., Rinella, M., Harrison, S. A., Brunt, E. M., & Sanyal, A. J. (2018). The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the american association for the study of liver diseases. *Hepatology*, *67*(1), 328–357. https://doi.org/10.1002/hep.29367

Chew, N., Pan, X., Chong, B., Chandramouli, C., Muthiah, M., & Lam, C. (2024). Type 2 diabetes mellitus and cardiometabolic outcomes in metabolic dysfunction-associated steatotic liver disease population. *Diabetes Research and Clinical Practice*, *211*, 111652. https://doi.org/10.1016/j.diabres.2024.111652

Cholongitas, E., Pavlopoulou, I., Papatheodoridi, M., Markakis, G. E., Bouras, E., Haidich, A. B., & Papatheodoridis, G. (2021). Epidemiology of nonalcoholic fatty liver disease in europe: A systematic review and meta-analysis. *Annals of Gastroenterology*, 404–414. https://doi.org/10.20524/aog.2021.0604

Clark, J. M., Brancati, F. L., & Diehl, A. M. (2002). Nonalcoholic fatty liver disease. *Gastroenterology*, *122*(6), 1649–1657. https://doi.org/10.1053/gast.2002.33573

Cobbina, E., & Akhlaghi, F. (2017). Non-alcoholic fatty liver disease (nafld) - pathogenesis, classification, and effect on drug metabolizing enzymes and transporters. *Drug Metabolism Reviews*, *49*(2), 197–211. https://doi.org/10.1080/03602532.2017.1293683

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). John Wiley & Sons.

Crte, C., Fintini, D., Giordano, U., Cappa, M., Brufano, C., Majo, F., Menini, C., & Nobili, V. (2012). Fatty liver and insulin resistance in children with hypobetalipoproteinemia: The importance of aetiology. *Clinical Endocrinology*, *79*(1). https://doi.org/10.1111/j.1365-2265.2012.04498.x

Daniel, W. W. (2005). *Biostatistics: A foundation for analysis in the health sciences*. Wiley.

de Graaf, W., van der Jagt, R. H., & van Gulik, T. M. (2011). Primary nonfunction of the liver allograft. *Transplantation Proceedings*, *43*(9), 3241–3243. https://doi.org/10.1016/j.transproceed.2011.08.045

Denk, H., Abuja, P. M., & Zatloukal, K. (2019). Animal models of NAFLD from the pathologist's point of view. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1865*(5), 929–942. https://doi.org/10.1016/j.bbadis.2018.04.024

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models* (3rd). Chapman & Hall/CRC.

Drew, L. (2017). Fighting the fatty liver. *Nature*, *550*(7675), S102–S103. https://doi.org/10.1038/550S102a

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Fahy, E., Cotter, D., Sud, M., & Subramaniam, S. (2011). Lipid classification, structures and tools. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, *1811*(11), 637–647. https://doi.org/10.1016/j.bbalip.2011.06.009

Fang, T., Wang, H., Pan, X., Little, P. J., Xu, S., & Weng, J. (2022). Mouse models of nonalcoholic fatty liver disease (NAFLD): Pathomechanisms and pharmacotherapies. *International Journal of Biological Sciences*, *18*(15), 5681–5697. https://doi.org/10.7150/ijbs.65044

Farzi, M., McGenity, C., Cratchley, A., Leplat, L., Bankhead, P., Wright, A., & Treanor, D. (2025). Liver-quant: Feature-based image analysis toolkit for automatic quantification of metabolic dysfunction-associated steatotic liver disease [Originally posted as a medRxiv preprint (2024.05.21.24305727, https://doi.org/10.1101/2024.05.21.24305727)]. *Computers in Biology and Medicine*. https://doi.org/10.1016/j.compbiomed.2025.110049

Fernando, D., Forbes, J. M., Angus, P. W., & Herath, C. B. (2019). Development and progression of non-alcoholic fatty liver disease: The role of advanced glycation end products. *International Journal of Molecular Sciences*, *20*(20), 5037. https://doi.org/10.3390/ijms20205037

Field, A. (2013). *Discovering statistics using ibm spss statistics* (4th). SAGE Publications.

Fiorini, R. N., Kirtz, J., Periyasamy, B., Evans, Z., Haines, J. K., Cheng, G., Polito, C., Rodwell, D., Shafizadeh, S. F., Zhou, X., Campbell, C., Birsner, J., Schmidt, M., Lewin, D., & Chavin, K. D. (2004). Development of an unbiased method for the estimation of liver steatosis. *Clinical Transplantation*, *18*(6), 700–706. https://doi.org/10.1111/j.1399-0012.2004.00282.x

Fong, D., Nehra, V., Lindor, K., & Buchman, A. (2000). Metabolic and nutritional considerations in nonalcoholic fatty liver. *Hepatology*, *32*(1), 3–10. https://doi.org/10.1053/jhep.2000.8978

Forlano, R., Mullish, B. H., Giannakeas, N., Maurice, J. B., Angkathunyakul, N., Lloyd, J., Tzallas, A. T., Tsipouras, M., Yee, M., Thursz, M. R., Goldin, R. D., & Manousou, P. (2020). High-throughput, machine learning-based quantification of steatosis, inflammation, ballooning, and fibrosis in biopsies from patients with nonalcoholic fatty liver disease [The official clinical practice journal of the American Gastroenterological Association]. *Clinical Gastroenterology and Hepatology*, *18*(9), 2081–2090.e9. https://doi.org/10.1016/j.cgh.2019.12.025

Gauthier, T. D. (2001). Detecting trends using spearman's rank correlation coefficient. *Environmental Forensics*, *2*(4), 359–362.

Gawrieh, S., Knoedler, D., Saeian, K., Wallace, J., & Komorowski, R. (2011). Effects of interventions on intra- and interobserver agreement on interpretation of nonalcoholic fatty liver disease histology. *Annals of Diagnostic Pathology*, *15*(1), 19–24. https://doi.org/10.1016/j.anndiagpath.2010.08.001

Gong, J., Tu, W., Liu, J., & Tian, D. (2023). Hepatocytes: A key role in liver inflammation. *Frontiers in Immunology*, *13*, 1083780. https://doi.org/10.3389/fimmu.2022.1083780

Hallou, A., Yevick, H. G., Dumitrascu, B., & Uhlmann, V. (2021). Deep learning for bioimage analysis in developmental biology. *Development*, *148*(18), dev199616. https://doi.org/10.1242/dev.199616

Hardy, T., Anstee, Q. M., & Day, C. P. (2016). Mechanisms of nafld development and therapeutic strategies. *Nature Reviews Gastroenterology Hepatology*, *13*(2), 88–98. https://doi.org/10.1038/nrgastro.2015.206

Hartig, F. (2024, October). *Dharma: Residual diagnostics for hierarchical (multi-level / mixed) regression models* [Vignette, CRAN]. Retrieved May 28, 2025, from https://cran.r-project.org/web/packages/DHARMa/vignettes/DHARMa.html#goodness-of-fit-tests-on-the-scaled-residuals

Hilbe, J. M. (2007). *Negative binomial regression* [Reimp. 2008]. Cambridge University Press.

Hilbe, J. M. (2011). *Negative binomial regression* (2nd ed.). Cambridge University Press.

Homeyer, A., Nasr, P., Engel, C., Kechagias, S., Lundberg, P., Ekstedt, M., Kost, H., Weiss, N., Palmer, T., Hahn, H. K., Treanor, D., & Lundström, C. (2017). Automated quantification of steatosis: Agreement with stereological point counting. *Diagnostic Pathology*, *12*(1), 80.

Huby, T., & Gautier, E. L. (2022). Immune cell-mediated features of non-alcoholic steatohepatitis. *Nature Reviews Immunology*, *22*(7), 429–443. https://doi.org/10.1038/s41577-022-00699-0

Huisman, M., & Schutte, J. (2017). Automation of image analysis workflows with fiji: A review of macro-based solutions. *Journal of Microscopy*, *267*(3), 268–280. https://doi.org/10.1111/jmi.12500

Kemmer, I., Keppler, A., Serrano-Solano, B., Rybina, A., Özdemir, B., Bischof, J., El Ghadraoui, A., Eriksson, J. E., & Mathur, A. (2023). Building a fair image data ecosystem for microscopy communities. *Histochemistry and Cell Biology*, *160*(3), 199–209. https://doi.org/10.1007/s00418-023-02203-7

Kleiner, D. E., Brunt, E. M., Van Natta, M., Behling, C., Contos, M. J., Cummings, O. W., Ferrell, L. D., Liu, Y. C., Torbenson, M. S., Unalp-Arida, A., Yeh, M., McCullough, A. J., & Sanyal, A. J. (2005). Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology*, *41*(6), 1313–1321. https://doi.org/10.1002/hep.20701

Koruk, M., Tayşi, Ş., Savaş, M. C., Yilmaz, O., Akçay, F., & Karakök, M. (2003). Serum levels of acute phase proteins in patients with nonalcoholic steatohepatitis. *The Turkish Journal of Gastroenterology*, *14*(1), 12–17.

Kumar, V., Abbas, A. K., & Aster, J. C. (2015). *Robbins and cotran pathologic basis of disease* (9th ed.). Elsevier/Saunders.

Lau, C., Kalantari, B., Batts, K., et al. (2021). The voronoi theory of the normal liver lobular architecture and its applicability in hepatic zonation. *Scientific Reports*, *11*(1), 9343. https://doi.org/10.1038/s41598-021-88699-2

Leitão, J., Sá, S., Cardoso, R., & et al. (2020). Prevalence of hepatic steatosis and metabolic associated fatty liver disease in a portuguese population-based sample. *Revista Portuguesa de Gastrenterologia*, *27*(3), 155–162. https://doi.org/10.1159/000507853

Leow, W.-Q., Chan, A. W.-H., Mendoza, P. G. L., Lo, R., Yap, K., & Kim, H. (2023). Non-alcoholic fatty liver disease: The pathologist's perspective. *Clinical and Molecular Hepatology*, *29*(Suppl), S302–S318. https://doi.org/10.3350/cmh.2022.0329

Li, Y., Xu, C., Yu, C., Xu, L., & Miao, M. (2010). High serum uric acid increases the risk for nonalcoholic fatty liver disease: A prospective observational study. *PLoS ONE*, *5*(7), e11578. https://doi.org/10.1371/journal.pone.0011578

Ludwig, J., Viggiano, T. R., McGill, D. B., & Oh, B. J. (1980). Nonalcoholic steatohepatitis: Mayo clinic experiences with a hitherto unnamed disease. *Mayo Clinic Proceedings*, *55*(7), 434–438.

Machado, M. V., & Diehl, A. M. (2016). Nonalcoholic fatty liver disease: Pathogenesis and the role of genetics. *Nature Reviews Gastroenterology Hepatology*, *13*(8), 412–424. https://doi.org/10.1038/nrgastro.2016.136

Manikat, R., & Nguyen, M. H. (2023). Nonalcoholic fatty liver disease and non-liver comorbidities. *Clinical and Molecular Hepatology*, *29*(Suppl), s86–s102. https://doi.org/10.3350/cmh.2022.0442

Marques, J. S. (2021). Vacuole_count_liver_steatosis [Computer software].

Martin-Grau, M., Marrachelli, V. G., & Monleon, D. (2022). Rodent models and metabolomics in non-alcoholic fatty liver disease: What can we learn? *World Journal of Hepatology*, *14*(2), 304–318. https://doi.org/10.4254/wjh.v14.i2.304

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman; Hall/CRC.

Mendler, M. H., Kanel, G., & Govindarajan, S. (2005). Proposal for a histological scoring and grading system for non-alcoholic fatty liver disease. *Liver International*, *25*(2), 294–304. https://doi.org/10.1111/j.1478-3231.2005.01052.x

Momose, Y., Ishii, M., & Kawai, T. (2011). Prevalence of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among a largely middle-aged population utilizing ultrasound and liver biopsy: A prospective study. *Gastroenterology*, *140*(1), 124–131. https://doi.org/10.1053/j.gastro.2010.09.033

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis* (4th). John Wiley & Sons.

Munsterman, I. D., van Erp, M., Weijers, G., Bronkhorst, C., de Korte, C. L., Drenth, J. P. H., van der Laak, J. A. W. M., & Tjwa, E. E. T. L. (2019). A novel automatic digital algorithm that accurately quantifies steatosis in nafld on histopathological whole-slide images. *Cytometry Part B: Clinical Cytometry*, *96*(6), 521–528. https://doi.org/10.1002/cyto.b.21790

Murteira, B. J. F. (2024). *Estatística: Inferência e decisão*. Sociedade Portuguesa de Estatística.

Nassir, F., Rector, R. S., Hammoud, G. M., & Ibdah, J. A. (2015). Pathogenesis and prevention of hepatic steatosis. *Gastroenterology and Hepatology*, *11*(3), 167–175.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Pietzsch, T., Preibisch, S., Tomancak, P., & Saalfeld, S. (2015). The imagej ecosystem: Open-source software for multidimensional image analysis. *Nature Methods*, *12*(12), 1000–1007.

Promrat, K., Lutchman, G., Uwaifo, G. I., Freedman, R. J., Soza, A., Heller, T., Doo, E., Ghany, M., Premkumar, A., Park, Y., Liang, T. J., Yanovski, J. A., Kleiner, D. E., & Hoofnagle, J. H. (2004). A pilot study of pioglitazone treatment for nonalcoholic steatohepatitis. *Hepatology*, *39*(1), 188–196. https://doi.org/10.1002/hep.20012

Qayyum, A., Kamar, N., & Rostaing, L. (2012). Primary nonfunction of the liver allograft. *Transplantation Proceedings*, *44*(9), 2653–2655. https://doi.org/10.1016/j.transproceed.2012.08.080

Rappaport, A. (1954). The microcirculatory acinar concept of normal and pathological hepatic structure. *The American Journal of Pathology*, *30*(3), 513–535.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, *2*(1), 21–33.

Sanyal, A. J., Brunt, E. M., Kleiner, D. E., Kowdley, K. V., Chalasani, N., Lavine, J. E., Ratziu, V., & McCullough, A. (2011). Endpoints and clinical trial design for nonalcoholic steatohepatitis. *Hepatology*, *54*(1), 344–353. https://doi.org/10.1002/hep.24376

Sayiner, M., Koenig, A., Henry, L., & Younossi, Z. M. (2016). Epidemiology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis in the united states and the rest of the world. *Clinics in Liver Disease*, *20*(2), 205–214. https://doi.org/10.1016/j.cld.2015.10.001

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: An open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682.

Schmid, B., & Mair, L. (2014). Image processing in biological research: A guide to the fundamentals and tools for analysis. *Bioinformatics*, *30*(4), 542–554.

Schwabe, R. F., Tabas, I., & Pajvani, U. B. (2012). Mechanisms of fibrosis development in nonalcoholic steatohepatitis. *Gastroenterology*, *142*(4), 711–725. https://doi.org/10.1053/j.gastro.2011.02.061

Tanaka, N., Kimura, T., Fujimori, N., Nagaya, T., Komatsu, M., & Tanaka, E. (2019). Current status, problems, and perspectives of non-alcoholic fatty liver disease research. *World Journal of Gastroenterology*, *25*(2), 163–177.

Trefts, E., Gannon, M., & Wasserman, D. H. (2017). The liver. *Current Biology*, *27*(21), R1147–R1151. https://doi.org/10.1016/j.cub.2017.09.019

Triola, M. F. (2018). *Elementary statistics* (13th). Pearson.

Vanderbeck, S., Benseler, V., & Fallon, M. B. (2013). Portal hypertension in nonalcoholic fatty liver disease: A review. *The American Journal of the Medical Sciences*, *346*(6), 486–491. https://doi.org/10.1097/MAJ.0b013e3182a5a4b1

Velosa, S. F., & Pestana, D. D. (2008, December). *Introdução à Probabilidade e à Estatística* (2nd ed.). Fundação Calouste Gulbenkian.

Wong, V. W.-S., Lazarus, J. V., Younossi, Z. M., Marchesini, G., Charatcharoenwitthaya, P., Abdelmalek, M. F., Bugianesi, E., George, J., & Fan, J.-G. (2019). Changing epidemiology, global trends and implications for outcomes of nafld. *Journal of Hepatology*, *79*(3), 842–852.

Younossi, Z. M., Gramlich, T., Matteoni, C. A., Boparai, N., & McCullough, A. J. (2004). Nonalcoholic fatty liver disease in patients with type 2 diabetes. *Clinical Gastroenterology and Hepatology*, *2*(3), 262–265. https://doi.org/10.1016/S1542-3565(04)00014-X

Younossi, Z. M., Koenig, A. B., Abdelatif, D., Fazel, Y., Henry, L., & Wymer, M. (2016). Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology*, *64*(1), 73–84. https://doi.org/10.1002/hep.28431

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with r*. Springer.

Appendices

Appendix A

Brunt system to grade NASH activity

Table A.1: Comparison of key histological features and scoring criteria among the Brunt, NASH CRN, and SAF systems for grading and staging NAFLD/NASH. Adapted from (Kleiner et al., 2005).

Numerical Grade or Stage	Brunt System	NASH CRN System	SAF System
Fibrosis Stage			
0	None	None	None
1	Zone 3 perisinusoidal fibrosis only	Perisinusoidal or periportal fibrosis; 3 substages defined	Perisinusoidal or periportal fibrosis
2	Zone 3 perisinusoidal fibrosis and periportal fibrosis	Perisinusoidal and periportal fibrosis	Perisinusoidal and periportal fibrosis
3	Bridging fibrosis	Bridging fibrosis	Bridging fibrosis
4	Cirrhosis	Cirrhosis	Cirrhosis
Ballooning Grade			
0	None	None	Only normal hepatocytes
1	Mild, zone 3	Few	Few: Clusters of hepatocytes with rounded shape and reticulated cytoplasm
2	Prominent, zone 3	Many	Many: Enlarged hepatocytes (≥2× normal)
3	Marked, zone 3		
Lobular Inflammation Grade			
0	No foci	No foci	No foci
1	1–2 foci per 20× field	<2 foci per 20× field	<2 foci per 20× field
2	2–4 foci per 20× field	2–4 foci per 20× field	>2 foci per 20× field
3	>4 foci per 20× field	>4 foci per 20× field	
Portal Inflammation Grade			
0	None	None	
1	Mild	Mild	
2	Moderate	More than mild	
3	Severe		
Steatosis Grade			
0	None	<5%	<5%
1	≤33%	5% to 33%	5% to 33%
2	33% to 66%	33% to 67%	33% to 67%
3	>66%	>67%	>67%

Appendix B

Quantile-Quantile Plots

In this appendix, some diagnostic plots are presented to assess the normality assumptions of the variables.

B.1 Quantile-Quantile plot of the variable Percentage of Steatosis

Analysing the Quantile-Quantile plot, we observe a tight cluster of values in the lower tail, reflecting a strong concentration of observations at or near zero. In the central region (roughly the 10th to 90th percentiles), the points lie close to the reference line, indicating that the bulk of the distribution is approximately normal. However, the clear departures in both tails reveal right-skewness and a violation of the normality assumption.

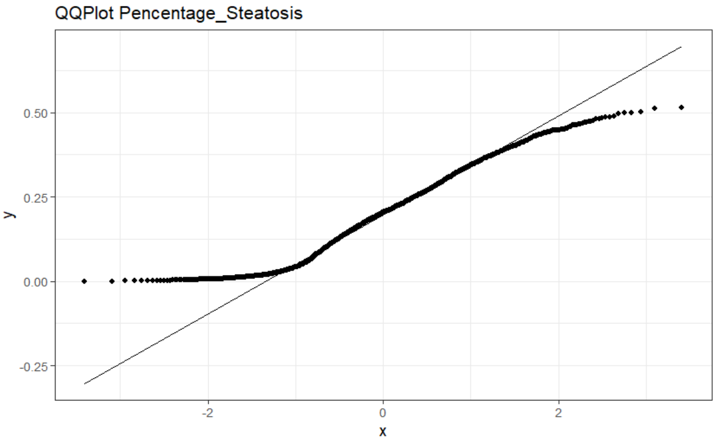


Figure B.1: Quantile-Quantile plot of the variable Percentage of steatosis.

B.2 Quantile-Quantile plot of the variable Number of vacuoles

Analysing the Quantile-Quantile plot, we observe a tight cluster of points near zero in the lower tail, reflecting very small counts. In the central region (roughly the 10th to 90th percentiles), the points fall close to the reference line, suggesting that mid-range counts are roughly normally distributed. However, the pronounced upward bend in the upper tail, where extreme values lie well above the line, reveals strong right-skewness, overdispersion, and a clear violation of the normality assumption.

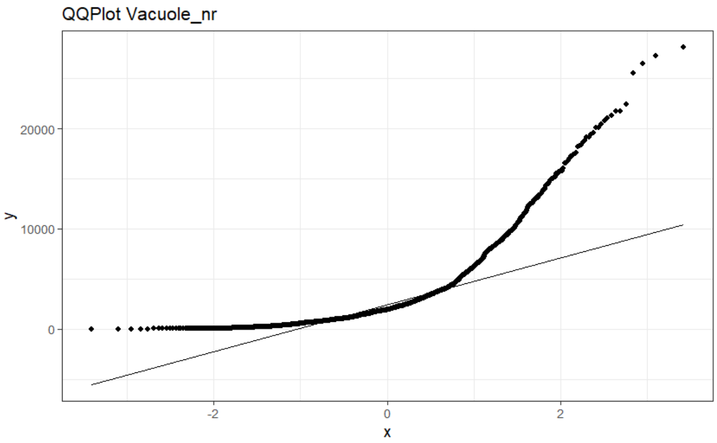


Figure B.2: Quantile-Quantile plot of the variable number of vacuoles.

B.3 Quantile-Quantile plot of the variable Area of vacuoles

Analysing the Quantile- Quantile plot, we observe a dense cluster of points near zero in the lower tail, reflecting many very small measurements. In the central region (approximately the 10th to 90th percentiles), the points lie close to the reference line, indicating that mid-range vacuole areas approximate a normal distribution. However, the pronounced upward bend in the upper tail, where the largest area values fall well above the line, reveals strong right-skewness, overdispersion, and a clear violation of the normality assumption.

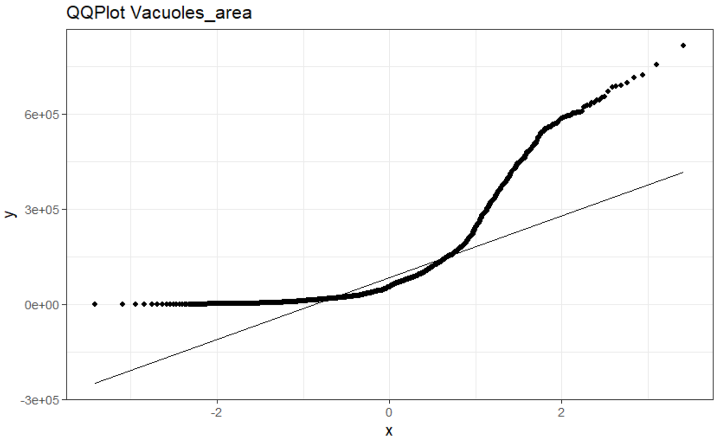


Figure B.3: Quantile-Quantile plot of the variable area of vacuoles.

Appendix C

Residual Analysis- Model with Interactions, Program x Magnification

The DHARMa (Diagnostics for Hierarchical Regression Models) package implements a simulation-based approach to generate readily interpretable scaled (quantile) residuals for fitted generalized linear (mixed) models. It is a comprehensive tool that provides several diagnostic plots and test functions to detect common model misspecifications, such as overdispersion, zero-inflation, and residual spatial autocorrelation. The resulting residuals are standardized to values between 0 and 1 and can be interpreted similarly to residuals from a linear regression model. Overdispersion is often a consequence of omitted predictors or an incorrectly specified model structure, yet standard residual plots often fail to capture such issues via residual correlations or patterns against predictors. Moreover, not all overdispersion is alike: in count data models, a negative binomial distribution produces different residual behaviour compared to a Poisson model with observation-level random effects. Additionally, heteroscedasticity—dispersion varying with predictors, is frequently present but seldom tested in GLMM, despite its influence on inference. Lastly, residual checks in GLMM are usually conducted conditionally on the estimated random effects, limiting diagnostics to the final level of the model hierarchy and neglecting potential issues with the full model structure (Hartig, 2024).

By analysing the Figure C.1:

The left panel: Quantile-Quantile plot residuals, tests whether the distribution of residuals matches the expected uniform distribution.

- KS test (Kolmogorov–Smirnov): $p = 0$
  The distribution of residuals significantly deviates from uniformity.

- Dispersion test: $p = 0.072$
  No statistically significant evidence of overdispersion or underdispersion at the 5% level.

- Outlier test: $p = 3 \times 10^{-5}$
  Strong evidence of outliers.

The observed "S-shaped" deviation in the Quantile-Quantile plot indicates a misfit in the distributional assumptions. Although the dispersion test is not significant, this pattern could be indicative of overfitting, zero-inflation, or non-independence in the data that allows the model to fit the training set too closely.

The right panel: Residuals vs. predicted values, the plot of DHARMa residuals against model predictions shows the empirical quantiles (0.25, 0.5, 0.75 in red) compared to their expected uniform distribution (horizontal dashed lines):

- The quantile curves deviate visibly from horizontal, indicating systematic patterns in the residuals.

- The combined quantile test is significant, reinforcing the evidence of poor residual fit.

This implies that the model does not fully capture the structure of the data across the range of predicted values.

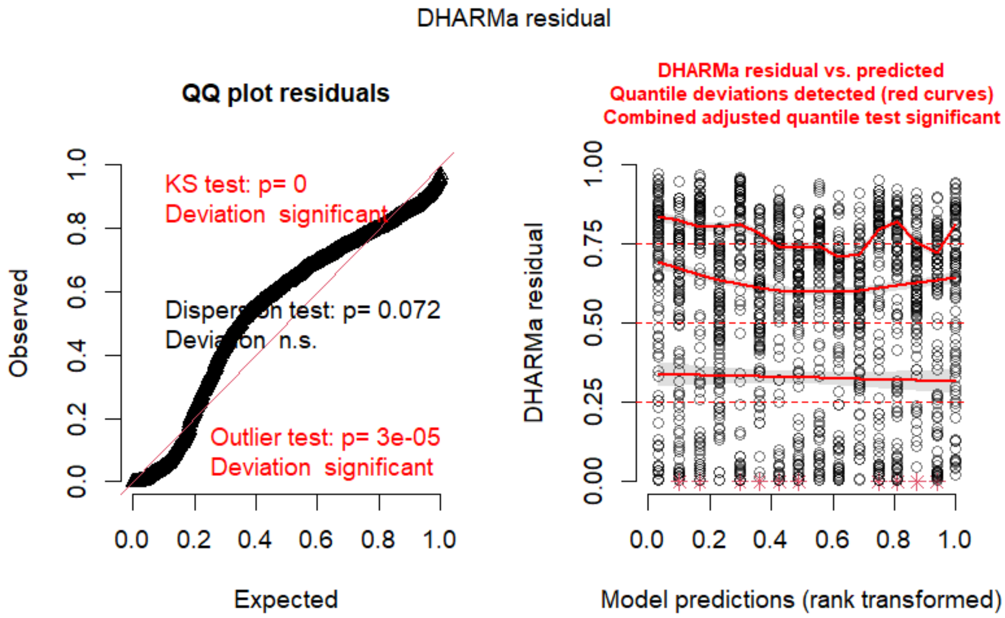


Figure C.1: DHARMa Residual Diagnostics: Quantile-Quantile Plot and Residuals vs. Predicted.

Additional test interpretations

- `testDispersion()` returns a dispersion statistic of 0.358 (ratio of observed to simulated SD), suggesting possible underdispersion. However, with p-value = 0.072, this deviation is not statistically significant. This suggests that while the model may slightly overestimate variability, the negative binomial distribution is adequate for modeling the overall dispersion in the data.

- `testZeroInflation()` returns p-value = 1, with a simulated median of zero zeros, meaning zero-inflation is not a concern for this model or dataset.

- `testUniformity()` results in a p-value $< 2.2 \times 10^{-16}$, strongly rejecting the null hypothesis of uniform residuals, confirming overall model misfit.

Despite these issues, we opted for a linear and interpretable GLMM structure that incorporates a random intercept for specimens, which accounts for between-specimen variability. Our primary goal is not prediction but understanding the effects of key variables on the response.

Nonetheless, the significant departures from residual uniformity suggest that model improvements, such as including additional covariates, revising the link function, or testing Nonlinear mixed models (NLMMs), could be explored in future work to better account for the observed heterogeneity.