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

Statistical Analysis of Longitudinal RBC Omics Data

Carolina Simões Gonçalves Silva

Mestrado em Bioestatística

Versão Provisória

Trabalho de Projeto orientado por:
Professora Marília Cristina De Sousa Antunes
Doutora Deborah Penque

2024

ACKNOWLEDGEMENTS

The work presented in this thesis was performed at the National Institute of Health Dr. Ricardo Jorge (Lisbon, Portugal), during the period between September 2023 and January 2025, under the supervision of Professor Marília Antunes and Doctor Deborah Penque.

I would like to acknowledge all the volunteers that participated in this study, supported by INSARJ, Toxomics, RNEM and FCT grants. Project was approved by the Ethical Commission of the National Institute of Health Dr. Ricardo Jorge and National Commission for Data Protection, Portugal.

To my supervisors Prof. Marília Antunes and Dr. Deborah Penque, for trusting me with the opportunity to develop this work. I am grateful for all their support, guidance, and knowledge during my involvement in this research project.

A special thanks to my colleges at the proteomics laboratory of DGH, Joana Saraiva, Sofia Neves, Fátima Vaz and Cristina Coelho, for their availability, patience, and advice.

My profound gratitude to my friends, Beatriz Antunes, Cristiana Pinheiro, and Mafalda Franco, for bringing me joy during more difficult times. To my partner, André Gonçalves, for accompanying me on this ride, full of ups and downs, and being a constant source of emotional support and motivation.

Last but not least, my sincere gratitude to my mother, for being unconditionally present to help, encourage and believe in me during my academic journey.

ABSTRACT

Red blood cells (RBCs) are emerging as important modulators of the immune system. Despite evidence that alterations in RBC functionality are associated with disease severity in COVID-19 patients, there is no information regarding the impact of RBC activity on the immune response to COVID-19 vaccination. This work aims to establish an adequate methodology for the statistical analysis of longitudinal RBC metabolomics data collected during COVID-19 vaccination (n=22, 5 time points) to identify metabolites with significant changes throughout the immunization process.

For the pre-treatment of the metabolomics data set, different pre-treatment methodologies comprised of imputation and normalization steps were compared to investigate which algorithm and application order was more adequate. Testing of these methods showed that normalization followed by kNN imputation using cosine distance was highlighted as the best-performing pre-treatment strategy. Following its application, generalized estimating equations (GEEs) created from the normalized data led to the identification of 30 metabolites with significant changes in concentration between different time points of COVID-19 vaccination.

Significant RBC metabolites were linked to major metabolic pathways in the cells, such as the metabolism of amino acids and purines, and the transport of small molecules through the cellular membrane. Some of these metabolites were discovered to have relevant functions in the development of an effective immune response against infections, like COVID-19. The connections between these metabolites and the defense mechanisms commonly used by cells to fight viral infections offer a strong clue for the immune functions that those metabolites may have in the human body, suggesting that the RBC metabolism could play a significant part in the generation of an immune response to COVID-19 vaccination.

Further work is in progress to integrate and correlate proteomic data retrieved from the same longitudinal experiment for a comprehensive depiction of the RBC function in the COVID-19 vaccine-induced immunization process.

Keywords: COVID-19 vaccination, Immune response, Longitudinal data, Metabolomic analysis, Red blood cells.

RESUMO

Os glóbulos vermelhos (também conhecidos como eritrócitos ou hemácias) são um tipo de células muito abundante no corpo humano cuja principal função é o fornecimento de oxigénio aos tecidos através do fluxo sanguíneo que percorre o sistema circulatório. Nos últimos anos, tem crescido o interesse por parte da comunidade científica pelo papel que os glóbulos vermelhos exercem no sistema imunológico, mediante o qual se ligam e eliminam citocinas e ácidos nucleicos a fim de modular a resposta imune. Curiosamente, tem vindo a descobrir-se que certas condições patológicas podem provocar mudanças no proteoma e metaboloma das hemácias e, consequentemente, no seu papel imunológico. Nesses estudos, os investigadores demonstraram que determinadas doenças (como a diabetes, o cancro e os distúrbios neurodegenerativos) afetam os aspetos morfofuncionais dos glóbulos vermelhos, por vezes alterando o seu metabolismo normal. Essas mudanças na funcionalidade das hemácias, por sua vez, podem ajudar a fornecer informações importantes sobre a gravidade e a progressão das doenças que as causaram. Em casos graves de COVID-19, especificamente, foi verificado que as alterações hematológicas induzidas pela infeção do vírus SARS-CoV-2 aumentavam com a progressão da gravidade da doença, sugerindo que estas modificações nos glóbulos vermelhos poderiam ser uma das causas da condição denominada por COVID longo.

Neste contexto, os estudos longitudinais de proteómica e metabolómica oferecem uma boa oportunidade para estudar a evolução do proteoma e metaboloma das células em resposta a estímulos internos ou externos. Graças aos avanços nas tecnologias ómicas de perfil longitudinal, a crescente disponibilidade de dados multivariados no âmbito da pesquisa biomédica tem exigido o desenvolvimento de métodos estatísticos apropriados que consigam descrever e modelar as relações complexas que existem entre as variáveis em estudo. Contudo, existem ainda vários problemas com a maioria dos métodos estatísticos usados atualmente para analisar dados ómicos com medições repetidas, como por exemplo a sua propensão a sobreajustes, o facto de não considerarem o desenho experimental na sua totalidade, não darem uso a todas as informações multivariadas intrínsecas aos dados, ou serem incapazes de estabelecer múltiplas associações entre conjuntos de dados ómicos distintos.

Enquanto estratégia primária de mitigação da COVID-19, a vacinação continua a ser a forma mais eficaz de proteção e prevenção contra esta doença. Por esse motivo, investigadores em todo o mundo têm-se esforçado para entender os processos moleculares subjacentes a este processo de vacinação que influenciam a resposta imunológica. Embora a proteómica e a metabolómica sejam frequentemente usadas para fornecer uma compreensão mais profunda das diversas funções celulares do sistema imune, esses dados são escassos no contexto da vacinação contra a COVID-19. Ainda assim, estudos recentes de metabolómica mostraram que a vacinação com mRNA contra a COVID-19 é responsável por induzir alterações metabólicas no plasma sanguíneo, sendo observados perfis metabólicos distintos em relação ao nível de resposta imune, o que permite destacar o papel de certos metabolitos como marcadores preditivos de resposta à vacinação. Apesar dessas descobertas, não há presentemente informações sobre o impacto da atividade das hemácias na resposta imune à vacinação contra a COVID-19.

Portanto, utilizando uma base de dados do metaboloma das hemácias obtido a partir de amostras de sangue coletadas durante a vacinação contra a COVID-19, este trabalho teve como objetivo o estabelecimento de uma metodologia adequada para a análise estatística desses dados metabolómicos longitudinais (n=22, 5 observações), levando em consideração o desenho experimental, a fim de identificar os metabolitos cuja concentração nas hemácias sofreu mudanças significativas ao longo do processo de imunização. Além disso, foi também investigada a existência de associações biológicas

e/ou estatísticas entre esses metabolitos, bem como o seu comportamento ao longo do tempo em resposta à vacinação contra COVID-19.

Tendo em conta o grande número de zeros presentes no conjunto de dados obtido da espetrometria de massa, os quais tanto poderiam representar metabolitos ausentes das amostras como abundâncias inferiores ao limite de deteção do espetrofotómetro, foi decidido que uma etapa de imputação seria necessária para fazer o pré-tratamento dos dados. Adicionalmente, foi desenvolvido um método de normalização para redimensionar os dados de cada indivíduo i e metabolito j através da expressão $(X_{ij} - baseline_{ij})/(max_{ij} - min_{ij})$, transformando-os em variações de concentração relativas à referência base da abundância natural dos metabolitos nos glóbulos vermelhos. De forma a investigar qual seria o algoritmo de imputação mais adequado aos dados ómicos deste trabalho, dois dos métodos mais comummente usados em metabolómica foram comparados, sendo estes o kNN (*k-nearest neighbors*) e o QRILC (*quantile regression imputation of left-censored data*). No caso do algoritmo kNN, foram testadas duas medidas de distância diferentes, a distancia de cosseno e a distância de Mahalanobis, ambas escolhidas pela sua invariância à escala dos dados. Ainda mais, foi também avaliada a ordem pela qual as etapas de imputação e normalização eram realizadas. O desempenho dos vários procedimentos de imputação e normalização foi aferido mediante a seleção aleatória de cinco pacientes e dez metabolitos para cada observação registada, e subsequente substituição desses valores de concentração por N/A. Após a aplicação dos diferentes métodos de pré-tratamento, cada conjunto de dados imputados e normalizados foi guardado para posterior análise. Este procedimento foi realizado 1000 vezes para garantir a sua reprodutibilidade. Por último, os valores obtidos em cada metodologia foram comparados com os dados originais normalizados (sem imputação) usando quatro medidas de erro, e o método com melhor desempenho foi empregado.

Após a conclusão do pré-tratamento, os dados metabolómicos foram modelados por meio de dois métodos distintos apropriados para estudos longitudinais. Começou-se por construir modelos lineares mistos (LMMs) para estudar as alterações metabólicas de cada paciente usando o "TimePoint" como variável independente/explicativa e incorporando o "PatientID" como efeito aleatório para introduzir alguma variabilidade no modelo derivada das diferenças inatas entre indivíduos. Alternativamente, foram criadas também equações de estimativa generalizada (GEEs) recorrendo a uma estrutura de correlação permutável para modelar a média populacional das tais alterações metabólicas. Ambas as abordagens foram desenvolvidas seguindo uma distribuição gaussiana e sempre assentes na suposição de homocedasticidade (variância constante entre diferentes observações). No final, cada modelo foi submetido a um diagnóstico como forma de testar as suas suposições e determinar a validade dos resultados alcançados.

A avaliação das diferentes metodologias de pré-tratamento de dados mostrou que o algoritmo de imputação kNN foi aquele que apresentou maior precisão na substituição dos valores desconhecidos com base no remanescente da amostra. Além disso, foi demonstrado que os métodos de pré-tratamento que incluíam o algoritmo kNN tinham desvios padrão mais baixos para cada medida de erro quando a normalização era realizada antes da imputação. Finalmente, o processo de normalização seguido de imputação kNN usando a distância do cosseno foi destacado como a estratégia de pré-tratamento com melhor desempenho. Depois da sua aplicação, todavia, os modelos LMM criados a partir dos dados normalizados foram diagnosticados com a presença de heterocedasticidade, quebrando assim uma de suas suposições fundadoras e, consequentemente, desqualificando dos seus resultados. O mesmo não foi observado com os modelos GEE que foram construídos, levando à identificação de 30 metabolitos que apresentavam diferenças significativas na sua concentração entre pelo menos duas das cinco observações medidas durante a vacinação contra a COVID-19. A maioria dessas alterações foi observada em t1 e t4, correspondendo aos efeitos de cada dose da vacina. Trajetórias de resposta

semelhantes foram encontradas entre grupos de até sete metabolitos, enfatizando potenciais associações biológicas dentro do metabolismo das hemácias.

A maioria dos metabolitos identificados como significativos pelos modelos GEE foram associados a vias metabólicas imprescindíveis nas células, tais como o metabolismo dos aminoácidos e das purinas e o transporte de pequenas moléculas através da membrana celular. Foi descoberto também que alguns dos metabolitos destacados têm funções relevantes no desenvolvimento de uma resposta imune eficaz contra infeções, nomeadamente contra a COVID-19. As suas principais funções parecem focar-se no aumento da produção de linfócitos, bem como a redução do stress oxidativo e diminuição da inflamação, ambos desencadeados pelo sistema imunológico inato em resposta a infeções. Coletivamente, as conexões encontradas entre os metabolitos dos glóbulos vermelhos e os mecanismos de defesa comummente usados pelas células para combater infeções virais oferecem uma forte pista para as funções imunológicas que esses metabolitos podem ter no corpo humano, sugerindo que o metabolismo das hemácias poderá desempenhar um papel significativo na geração de uma resposta imune à vacinação da COVID-19.

Como continuação deste estudo sobre a relevância dos metabolitos dos glóbulos vermelhos no desenvolvimento de respostas imunes, mais trabalhos estão em desenvolvimento para integrar e correlacionar dados proteómicos recuperados do mesmo ensaio longitudinal com o propósito de obter uma representação abrangente da função das hemácias no processo de imunização induzido pela vacina da COVID-19.

Palavras-chave: Vacinação COVID-19, Resposta imune, Dados longitudinais, Análise metabolómica, Glóbulos vermelhos.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS i
ABSTRACT ii
RESUMO iii
TABLE OF CONTENTS vi
LIST OF TABLES vii
LIST OF FIGURES ix
LIST OF ABBREVIATIONS xii
1. INTRODUCTION 1
1.1. Background and motivation 1
1.2. Objectives and work plan summary 3
2. STATISTICAL METHODOLOGY 5
2.1. State of the art 5
2.1.1. Data pre-processing 5
2.1.2. Estimation/imputation of missing values and zeros 6
2.1.3. Data normalization 7
2.1.4. Data pre-treatment 8
2.1.5. Objectives of longitudinal analysis 9
2.1.6. Sources of correlation in longitudinal data 9
2.1.7. Generalized linear models for longitudinal data 10
2.1.8. Generalized linear mixed effects models 11
2.1.9. Marginal models for longitudinal data 12
2.1.10. Estimation of marginal models through generalized estimating equations 12
2.2. Choosing the right approach 13
2.2.1. Imputation of zero values 13
2.2.2. Scaling of longitudinal data 14
2.2.3. Algorithm performance test 15
2.2.4. Development of linear mixed models 16
2.2.5. Establishing the generalized estimating equations 16
2.2.6. Pathway analysis with Metaboanalyst 18
3. RESULTS AND DISCUSSION 19
3.1. Data description 19
3.2. Comparison of pre-treatment methods 19
3.3. Data modeling through LMM vs GEE 21

3.4. Behavioral and biological connections 27
4. CONCLUSIONS AND FUTURE PERSPECTIVES 38
5. REFERENCES 39
6. ANNEXES 47

LIST OF TABLES

Table 3.1. List of GEE coefficients with statistical significance (p-value < 0.1) and respective confidence level for each of the 30 metabolites identified through the application of GEE models 25

Table 3.2. Pathway analysis conducted through MetaboAnalyst 6.0 using as input a list of compound names from the 30 metabolites found to have significant GEE model coefficients 32

Table 3.3. Enrichment analysis conducted through MetaboAnalyst 6.0 using as input a list of compound names from the 30 metabolites found to have significant GEE model coefficients 33

Table 3.4. Behavioral analysis of metabolites with similar response trajectories to COVID-19 vaccination 35

LIST OF FIGURES

Figure 3.1. Performance assessment of each imputation and normalization procedure using the Mean Absolute Error (MAE) 20

Figure 3.2. Performance assessment of each imputation and normalization procedure using the Relative Absolute Error (RAE) 20

Figure 3.3. Performance assessment of each imputation and normalization procedure using the Root Mean Squared Error (RMSE) 20

Figure 3.4. Performance assessment of each imputation and normalization procedure using the Root Relative Squared Error (RRSE) 20

Figure 3.5. Plot of the LMM model constructed from the normalized abundance changes observed in the metabolite (R)-lipoic acid over time during COVID-19 vaccination 22

Figure 3.6. Plot of the LMM model constructed from the normalized abundance changes observed in the metabolite 2-pyrrolidinone over time during COVID-19 vaccination 22

Figure 3.7. Plot of the LMM model constructed from the normalized abundance changes observed in the metabolite 5-methylcytidine over time during COVID-19 vaccination 22

Figure 3.8. Plot of the LMM model constructed from the normalized abundance changes observed in the metabolite adenosine 3',5'-diphosphate (PAP) over time during COVID-19 vaccination 22

Figure 3.9. Diagnostic plot of the LMM residuals of the metabolite 2-pyrrolidinone against the fitted values of its LMM model 23

Figure 3.10. Diagnostic plot of the LMM residuals of the metabolite adenosine 3',5'-diphosphate (PAP) against the fitted values of its LMM model 23

Figure 3.11. Diagnostic plots of the GEE residuals of the metabolite hypoxanthine against the fitted values of its GEE model 26

Figure 3.12. Diagnostic plots of the GEE residuals of the metabolite L-proline against the fitted values of its GEE model 26

Figure 3.13. Diagnostic plots of the GEE residuals of the metabolite (R)-lipoic acid against the fitted values of its GEE model 26

Figure 3.14. Diagnostic plots of the GEE residuals of the metabolite 5'-methylthioadenosine (MTA) against the fitted values of its GEE model 26

Figure 3.15. ANOVA results of the top ranked metabolites with the lowest FDR-adjusted p-values obtained for the comparison between the GEE models and their null counterparts 26

Figure 3.16. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite (R)-lipoic acid over time during COVID-19 vaccination 27

Figure 3.17. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite 2-ethyl-2-hydroxybutyric acid over time during COVID-19 vaccination 27

Figure 3.18. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite 2-hydroxyadenine over time during COVID-19 vaccination 27

Figure 3.19. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite 2-pyrrolidinone over time during COVID-19 vaccination 27

Figure 3.20. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite 5'-methylthioadenosine (MTA) over time during COVID-19 vaccination 28

Figure 3.21. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite 5-methylcytidine over time during COVID-19 vaccination 28

Figure 3.22. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite adenosine 3',5'-diphosphate (PAP) over time during COVID-19 vaccination 28

Figure 3.23. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite adenosine monophosphate (AMP) over time during COVID-19 vaccination 28

Figure 3.24. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite creatine over time during COVID-19 vaccination 28

Figure 3.25. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite creatinine over time during COVID-19 vaccination 28

Figure 3.26. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite D-arginine over time during COVID-19 vaccination 28

Figure 3.27. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite glucaric acid over time during COVID-19 vaccination 28

Figure 3.28. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite glycine over time during COVID-19 vaccination 29

Figure 3.29. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite guanosine 5'-diphosphate (GDP) over time during COVID-19 vaccination 29

Figure 3.30. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite guanosine monophosphate (GMP) over time during COVID-19 vaccination 29

Figure 3.31. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite hyodeoxycholic acid over time during COVID-19 vaccination 29

Figure 3.32. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite hypoxanthine over time during COVID-19 vaccination 29

Figure 3.33. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite isomaltose over time during COVID-19 vaccination 29

Figure 3.34. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite isopropyl alcohol over time during COVID-19 vaccination 29

Figure 3.35. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite L-carnitine over time during COVID-19 vaccination 29

Figure 3.36. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite L-methionine over time during COVID-19 vaccination 30

Figure 3.37. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite L-phenylalanine over time during COVID-19 vaccination 30

Figure 3.38. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite L-proline over time during COVID-19 vaccination 30

Figure 3.39. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite phenylacetaldehyde over time during COVID-19 vaccination 30

Figure 3.40. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite phosphoric acid over time during COVID-19 vaccination 30

Figure 3.41. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite pyridine over time during COVID-19 vaccination 30

Figure 3.42. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite sarcosine over time during COVID-19 vaccination 30

Figure 3.43. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite stearic acid over time during COVID-19 vaccination 30

Figure 3.44. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite trigonelline over time during COVID-19 vaccination 31

Figure 3.45. Plot of the GEE model constructed from the normalized abundance changes observed in the metabolite uric acid over time during COVID-19 vaccination 31

LIST OF ABBREVIATIONS

ADP – Adenosine 5'-diphosphate

AMP – Adenosine monophosphate

ANOVA – Analysis of variance

APRT – Adenine phosphoribosyltransferase deficiency

ATP – Adenosine triphosphate

BH – Benjamini-Hochberg

BMI – Body mass index

DHPD – Dihydropyrimidine dehydrogenase deficiency

FDR – False discovery rate

GC – Gas chromatography

GDP – Guanosine diphosphate

GEE – Generalized estimating equations

GLM – Generalized linear model

GLMM – Generalized linear mixed model

GMP – Guanosine monophosphate

HMDB – Human metabolome database

KEGG – Kyoto encyclopedia of genes and genomes

kNN – K-nearest neighbors

LC – Liquid chromatography

LMM – Linear mixed model

LOD – Limit of detection

MAE – Mean absolute error

MAR – Missing at random

MCAR – Missing completely at random

MNAR – Missing not at random

MS – Mass spectrometry

MTA – 5-Methylthioadenosine

N/A – Not available

NAD – Nicotinamide adenine dinucleotide

ORA – Over representation analysis

PAP – Adenosine 3',5'-diphosphate

QIC – Quasi information criterion

QICC – Corrected quasi information criterion

Q-Q – Quantile-quantile

QRILC – Quantile regression imputation of left-censored data

RAE – Relative absolute error

RaMP-DB – Relational database of metabolomic pathways

RBC – Red blood cell

RF – Random forest

RMSE – Root mean squared error

ROS – Reactive oxygen species

RRSE – Root relative squared error

SLC – Solute carrier

TCA – Tricarboxylic acid

1. INTRODUCTION

1.1. Background and motivation

The COVID-19 pandemic has had a profound impact on human history, changing lifestyles and enriching scientific knowledge on viral infections. During the acute phase of COVID-19 infection, excessive release of cytokines and other pro-inflammatory molecules can lead to multi-organ dysfunction or even death. Because of its concerning severity, the spread of this disease accelerated vaccine development and implementation worldwide, resulting in a significant reduction in death and hospitalization rates. However, four years later, the cellular mechanisms involved in the SARS-CoV-2 immunization response remain poorly exploited [1-3].

Systems biology approaches such as proteomics and metabolomics are important tools for understanding the molecular mechanisms of several diseases, including COVID-19. Metabolomics, in particular, refers to the large-scale study of small molecules, commonly known as metabolites, within cells, biofluids, tissues or organisms. Collectively, these small molecules and their interactions within a biological system are known as the metabolome. In this context, metabolomics can be employed to uncover metabolic pathways that are correlated with vaccine immunogenicity, and hence may have critical roles in immune response mechanisms. Because of its potential to provide deep insights into the biology of health and disease, the integration of metabolomics with clinical data is incredibly valuable for the discovery of biomarkers that could be used as new therapeutic targets for preventive treatments or vaccine development [3,4].

In several metabolomics studies of COVID-19 patients and SARS-CoV-2 vaccines, metabolic profiles of infected or immunized individuals exhibited dysregulated pathways, with the most common alterations being found in the tricarboxylic acid (TCA) cycle, lipid metabolism, purine metabolism, and amino acid biosynthesis. These metabolic pathways are essential host factors that influence the response to bacterial and viral infections, shaping the immune system. For instance, the regulation of the TCA cycle involves a mechanism of host defense against intracellular pathogens through hypoxia and reduction of microbial replication. In addition, purine metabolites such as adenosine can promote cell survival and proliferation by suppressing pro-inflammatory responses and promoting an anti-inflammatory response instead. Because of that, accumulation of adenosine due to deficiency in purine metabolite degradation increases the susceptibility to infection and autoimmunity. Interestingly, many of these metabolic changes were often linked to COVID-19 progression in infected patients, having a positive correlation with the severity of the disease [1,3-5].

One example of a metabolite altered by COVID-19 is the TCA intermediate alpha-ketoglutaric acid. When present in high levels, alpha-ketoglutaric acid can inhibit SARS-CoV-2 replication in human monocytes and is associated with anti-inflammatory and anti-thrombotic roles. However, during the progression of COVID-19, there is a decrease in the levels of this metabolite, which could be related to mitochondrial dysfunction, hypoxia, and tissue damage through the reduction of TCA cycle activation. Likewise, other TCA cycle intermediates such as pyruvate, citrate, malate, and lactate, were significantly decreased after CoronaVac (Sinovac) immunization and negatively associated with IgG levels, pointing to a quick activation of B-cells and subsequent induction of antibody responses, which play an essential

role in antiviral immunity. Thus, it is likely that some of these metabolites may have relevant roles in the underlying mechanisms for CoronaVac-induced immunity [1,6].

In a different study, researchers showed that glutamine, the most abundant amino acid in human plasma, was reduced in COVID-19 patients. This metabolite is normally used by immune cells as fuel for proliferation and growth, though it is also needed for SARS-CoV2 replication, meaning that the possible implementation of glutamine supplementation during infection to improve the patient's clinical outcome would require cautious deliberation. Additional metabolites related to the amino acid metabolism, like creatinine and phenylalanine, have also been found decreased in individuals with adverse symptoms following immunization, indicating that certain vaccine adverse events may be metabolically mediated [4,7].

As key components of cellular membranes and sources of energy, lipids play a significant role in modulating the immune response and inflammation by promoting viral entry into host cells, regulating viral translation/replication as well as immune activation, and inducing pro-inflammatory cytokine release. Hence, disturbances in the lipidomic profile of COVID-19 patients, particularly in the content of cholesterol compounds, suggest that these metabolites and their functions are probably connected to the immunization process of this disease [1,7,8].

Erythrocytes, or red blood cells (RBCs), make up about 83% of all cells in the human body. Though they are often overlooked as mere vessels for oxygen transport due to their lack of organelles like mitochondria, RBCs are metabolically active cells that play numerous roles beyond this function, maintaining homeostasis in many metabolic pathways. Recent studies have revealed the complexity of erythrocyte metabolism, including thousands of small-molecule analytes and hundreds of enzymes that can catalyze reactions. This new understanding has revamped interest in the erythrocyte and its role in biochemistry and physiology [5,9,10].

Some of the crucial mechanisms governed by RBCs include the control of gas transport and oxidative signaling; the regulation of bicarbonate, chloride, sodium/potassium, and calcium homeostasis; oxidation-reduction (redox) homeostasis, particularly glutathione, sulfur, and iron metabolism; release and uptake of metabolic mediators like lactic acid and carboxylates; and generation of metabolic signals impacting immune cell responses. RBCs can also respond to neurotransmitter stimulation and participate in circulating neurotransmitter metabolism, which is a function reminiscent of circulating neural cells. Additionally, they can serve as a vehicle for paracrine signaling by releasing microvesicles enriched with small molecules, bioactive lipids, and partially functional proteins. Furthermore, RBCs are involved in transamination reactions, as well as in the uptake and metabolism of circulating endogenous bioactive molecules and xenobiotics (phytochemicals or synthetic drugs). Last but not least, erythrocytes can communicate with several types of cells and tissues, namely immune cells, using actively or passively released material such as vesicles, proteins or metabolites, in order to influence their functioning through a cross-talk process [10].

Many erythrocyte dynamics are affected, however, by the inflammatory processes and oxidative stress manifested during SARS-CoV-2 infection. Significant changes in cell size and rigidity have been revealed to occur in RBCs due to COVID-19, with possible links to disease progression and severity. SARS-CoV-2 infection can induce structural damage in the plasma erythrocyte membrane resulting in shape loss, early lysis, excessive release of free iron ions, and elevation of serum ferritin. Because iron is essential for a great number of oxidation-reduction reactions, the dysregulated iron control mechanism of the erythrocyte creates an unfavorable redox status in cells that is commonly displayed by COVID-19 patients. This increase in reactive oxygen species (ROS) within RBCs leads to the harmful oxidation of biomolecules and is likely responsible for their abnormal morphology and aggregation in infected

individuals. Increased oxidation of RBC structural proteins, in particular, may contribute to the thromboembolic and coagulopathic complications seen in some critically ill patients [2,11,12].

Some COVID-19 survivors may also experience long-term changes in their RBCs, including altered hematological status and phenotypic changes. This persistence of abnormal RBC properties, with more heterogeneous sizes and deformation, could explain the symptomatology observed in Long COVID patients, which is characterized by a general state of malaise, headaches, concentration disorders, weakness, muscle fatigue, and short breath [11].

Despite the publication of a few metabolomics studies identifying RBC alterations that are induced by SARS-CoV-2 infection, research is still scarce on the impact of RBC activity on the immune cell response after COVID-19 vaccination, specifically regarding the underlying molecular processes and metabolic changes that affect such pivotal responses.

To shed some light on this matter, researchers from the Human Genetics Department of Instituto Nacional de Saúde Dr. Ricardo Jorge (Lisbon, Portugal) collected blood samples from 22 individuals at several time points during COVID-19 vaccination (Pfizer, Moderna, AstraZeneca or J&J) between April and September of 2021. From these samples, RBCs were isolated and its metabolites extracted for liquid chromatography coupled to tandem mass spectrometry (mass/charge intensity) analysis, also known as LC-MS/MS. The intensity of each molecule was then determined at five different times: t0 (before vaccination), t1 (24-72h after the first vaccine dose), t2 (before the second vaccine dose), t3 (24-72h after the second vaccine dose) and t4 (30 days after the last vaccine dose). Information on gender, age and BMI for every participant was also recorded.

The next step in the investigation was to retrieve relevant information regarding possible COVID-19-related immune activity from the RBC metabolome database obtained in this study. However, the type of repeated measures data required to evaluate the response of multiple individuals to COVID-19 vaccination, and therefore study their metabolic changes over time, meant that careful thought would be needed in the process of choosing the best statistical approach for the intended analysis. Thankfully, owing to the advancements in omics longitudinal profiling technologies, the increasing availability of multivariate data within biomedical research has pushed for the development of several statistical methods capable of describing and modeling the complex relationships between variables that occur in time-course experiments, such as this one.

1.2. Objectives and work plan summary

With this motivation in mind, the main goal of the present work was to establish an adequate methodology for the statistical analysis of the longitudinal RBC metabolomics data gathered at different time points during COVID-19 vaccination in order to identify metabolites with significant changes in concentration throughout the immunization process. Additionally, the existence of biologic and/or statistical associations between those metabolites, along with their joint behavior over time, would also be investigated.

The work plan developed to accomplish these objectives can be summarized into three parts:

(1) Pre-treatment of the database, seeking to establish the need for imputation, as well as which imputation method to apply, considering the nature of the data. Validation of the selected pre-treatment procedure.

(2) Construction of a statistical model for each metabolite appropriate for longitudinal data, building a ranking of corrected p-values (using, for example, False Discovery Rate) to determine which metabolites showed significant fluctuations in concentration over time (t0-t4) in response to the COVID-19 vaccination.

(3) Evaluation of the joint behavior of different metabolites over time, considering established biological or statistical associations.

2. STATISTICAL METHODOLOGY

2.1. State of the art

As previously stated, metabolomics is a research field that focuses on the identification and quantification of small molecules in biological samples with the intention of detecting the causes of their variability in response to certain experimental conditions. When conducting a metabolomics study, a snapshot of the metabolome is obtained reflecting the cellular state, or phenotype, of the analyzed biological samples under the experimental conditions tested. However, the variations observed within the metabolome during this experiment can derive from multiple sources, which are ultimately present in metabolomics data: (1) magnitude differences between metabolite concentrations resulting from their higher or lower natural abundances in cells; (2) fold change differences in metabolite concentration due to specific alterations in metabolic pathways induced through the experimental conditions of the study; (3) uninduced technical variations originated from sampling and analytical errors; (4) data heteroscedasticity caused by variables with non-constant standard deviations [13-16].

Following the emergence of metabolomics as a powerful tool for precision medicine, investigators quickly realized that metabolomics data was rather complicated to deal with due to some of the aforementioned properties, usually requiring additional processing steps before the statistical analysis could be performed [13-15].

Broadly speaking, data processing steps done prior to statistical analysis can be divided into several categories: data pre-processing (including baseline correction, peak alignment, and noise filtering), missing value estimation/imputation, data normalization, and finally data pre-treatment (which refers to centering, scaling and data transformation). Altogether, the goal of these procedures is to minimize the influence of disturbing factors by correcting for instrumental artifacts and irrelevant biological variability to enhance the signal-to-noise ratio, while also appropriately transforming the data into interpretable spectral profiles and reducing its dimensionality. Collectively, these tasks allow the separation of significant biological information regarding alterations in metabolite concentration from any other obscuring sources of variability introduced during the experimental studies, thereby emphasizing the largest differences between groups of observations [13,14,16].

Yet, most importantly, the choice for suitable processing methods must be determined by the biological question to be answered, the characteristics of the data set and the selected methods for statistical analysis. If not done properly, this method selection can greatly affect the outcome of the data analysis and, consequently, the ranking of relevant metabolites for the subject problem of the study, thus hindering the biological interpretation of the results obtained from the statistical analysis [13,14,16].

2.1.1. Data pre-processing

During the pre-processing stage of metabolomics data, various methods such as noise filtering, baseline correction, sample alignment, and also peak picking/merging, are implemented on the raw mass spectrometry data set to create a clean data matrix containing the relative abundances of metabolites in

each sample or subject under different conditions (e.g., disease vs treatment). This data matrix is then used for any further treatment still necessary before statistical analysis [13,16].

2.1.2. Estimation/imputation of missing values and zeros

Spread across MS-based metabolomics data, missing values and zeros are relatively common to find in the data matrix after the pre-processing steps. Caused by either biological factors or technical errors, these zero/missing values pose a big challenge in the data processing phase when encountered in high amounts since they can affect the correlation between variables and, subsequently, mislead the results of the downstream statistical analysis [13,15].

Generally, there are three kinds of missing values. Missing completely at random (MCAR) are missing values that result from random errors and stochastic fluctuations during the data acquisition process, namely from incomplete derivatization/ionization. Similarly, missing at random (MAR) are missing values caused by sub-optimal data pre-processing such as inaccurate peak detection and deconvolution of co-eluting compounds. Most times, however, MCAR and MAR data are very hard to distinguish. Lastly, missing not at random (MNAR) are missing values created from censored peak intensities that stayed below the limits of quantification of the mass spectrometer. Interestingly, an evaluation of the most prevalent types of missing values in different metabolomics data sets discovered that MCAR and MAR occur widely in GC/MS profiling data, while MNAR is more prominent in LC/MS targeted data [13,15].

Likewise, zero values can also be categorized into different types depending on their source. (1) Structural zeros, which refer to specific peaks that could not be found in the chromatogram for biological reasons. (2) Sampling zeros, meaning the peaks that were missed during peak picking despite being present in the samples. (3) Intensities or abundances that were below the limit of detection (LOD) of the mass spectrometer. (4) Negative values obtained from metabolomics measurements that are usually considered as spectral artifacts or noise, therefore being transformed automatically into zeros [13].

Noticeably, some types of missing and zero values share the same foundations, as is the case of MNAR and LOD-censored abundances. This is because different software used to analyze raw metabolomics data have their own particular default parameters for applying missing values or zeros, sometimes leading to missing values being replaced by zeros in the data set. Consequently, these changes make it more difficult to discern between, for example, sampling zeros that were originally MAR and true biological zeros that are absent from the samples. Even with the help of missingness algorithms, it can still be challenging to discriminate the several kinds of missing/zero values given this increased complexity of the data matrix. As a result, there is currently no general strategy for dealing with such values in metabolomics studies [13].

Nevertheless, a few practical methods to deal with missing/zero values have been proposed by researchers over the years. One of these approaches is to remove missing or zero values based on a certain threshold like the “80% rule”, where a metabolite is kept for data analysis if it has a non-zero or non-missing value for at least 80% of the samples, hence significantly reducing their numbers in the data set to facilitate statistical analysis. More options frequently employed by investigators include the imputation of missing values based on the mean or the minimum of non-missing values, replacing them with zero, or using a specific imputation algorithm to perform that substitution. The choice of imputation method, however, has been shown to influence the results and interpretation of subsequent statistical

analyzes, leaving scientists in disagreement on which imputation method is the optimal missing value estimation approach [13,15].

Surveys on several metabolomics studies have revealed that k-Nearest Neighbor (kNN), Random Forest (RF) and Quantile Regression Imputation of Left-Censored data (QRILC) are the most used imputation methods for mass spectrometry data sets. The RF algorithm, in particular, has shown great performance for MCAR/MAR data, whereas QRILC was demonstrated to be better suited for left-censored MNAR data. Seeing this selectiveness in missing data type, some investigators have started to recognize that each source of missing values should be dealt with in a specific way, thus believing the best strategy to be the application of missing data estimation algorithms to choose different methods to handle each type of missing values in metabolomics data. Aside from the utilization of imputation algorithms, most metabolomics studies tend to exclude metabolites that have a percentage of missingness above a determined threshold, after which they replace the remaining missing values by zero or by the minimum of non-missing values [13,15].

Regardless of these debates on the preferred approaches for imputing missing/zero values, one consensus among researchers is that caution needs to be taken when choosing an appropriate imputation method as employing an unsuitable one to estimate values may not only lead to undetected biological differences between groups but also introduce further bias in the results via the fabrication of significant divergences from non-representative peaks [13,15].

2.1.3. Data normalization

Following the pre-processing phase and the imputation of missing/zero values, there are two more categories of methods that are typically performed prior to the statistical analysis of metabolomics data. The first group of methods, called data normalization, is used to remove unwanted sample-to-sample variation, whereas the second group of methods, which include centering, scaling and data transformation, aids in adjusting the variance of different metabolites to reduce sample heteroscedasticity. Together, these procedures aim to improve data comparability, therefore contributing to an increased reliability and interpretability of the downstream statistical analysis [13].

Since data normalization is generally performed on rows (samples) while the other methods are applied on columns (metabolites), they can also be referred to in terms of the data matrix structure as row-wise normalization and column-wise normalization, respectively. Though this first category of approaches is debated here, the methods of centering, scaling and data transformation will be saved for discussion in the next chapter of data pre-treatment [13].

Data normalization is considered by most as a demanding yet essential task in metabolomics data processing due to several biological factors and technical reasons, such as the chemical diversity of metabolites resulting in many different responses to variations in the experimental conditions. The goal of sample-based normalization is to make samples comparable to each other by removing or minimizing any experimental variance and unwanted systematic errors/biases. More specifically, data normalization is used to reduce systematic variation or bias in the data related to instrumental/sampling problems (e.g., sample inhomogeneity, differences in sample preparation, and ion suppression), as well as to separate biological variability from other variations introduced in the experimental process [13,14,16].

Mathematically speaking, the normalized peak areas of the metabolites represent a fraction of their initial intensities over the summation of the integrated intensities for all spectral regions, thereby scaling

the spectra to the same overall concentration to account for different dilutions of samples. The most common sample-based normalization approaches include normalizing to the total spectral area (Constant Sum Normalization), normalizing to a reference sample (Probabilistic Quotient Normalization), and normalizing to a reference feature/metabolite (Internal Standard Normalization) [13,14,16].

2.1.4. Data pre-treatment

As briefly mentioned in the last chapter, data pre-treatment is a group of methods comprising centering, scaling and data transformation, which are meant to be applied to the metabolomics data set before statistical analysis in order to adjust the variance of each metabolite by lessening the impact of larger feature values, hence making all features more comparable and reducing sample heteroscedasticity [13].

Centering is a pre-treatment method for metabolomics data that subtracts the mean value from each measured metabolite and converts all metabolite concentrations into fluctuations around zero instead of around their mean value. By correcting for the differences between high and low abundant metabolites, centering allows the data analysis to focus on the relevant deviations in metabolite concentration between samples regardless of their individual biological abundances [13,16].

Scaling, on the other hand, is used to divide each variable by a scaling factor as a way to convert the data into variations of concentration relative to that scaling factor, ergo adjusting for different biological ranges that can vary widely between metabolites. With this in mind, the purposes of scaling metabolomics data are to adjust the scale differences that exist among metabolites and accommodate for heteroscedasticity while still reflecting the original intracellular metabolite concentrations [13,16,17].

Overall, scaling methods have different names according to the scaling factor used for each variable. Based on whether the scaling factor is a measure of data dispersion or size (e.g., standard deviation and mean, respectively), scaling methods can be divided into two subclasses: dispersion-based scaling (which includes autoscaling, Pareto scaling, range scaling, and vast scaling), and average-based scaling (such as level scaling). When comparing several scaling methods concerning their biological expectations, autoscaling and range scaling have been highlighted by some investigators as the best-performing methods for removing metabolite rank dependencies on the average metabolite concentration and the magnitude of fold changes [13,16].

Also known as unit variance scaling, autoscaling is a very useful standardization method for accentuating mean-level differences between groups by subtracting the mean of a particular feature from each observation and dividing the difference by the standard deviation of that feature. The result is a scale of values that have zero mean and unit variance, where a score of 0 represents an observation that is at the feature's mean level, and a score of 1 reflects an observation that is one standard deviation above the mean. Although this method makes all metabolites of equal importance, meaning they are equally weighted and have equal potential to influence statistical models, it is important to remember that it can also amplify analytical errors due to measurement dilution effects. The main advantages of unit variance scaling are (1) the improved performance of implemented algorithms, preventing certain features from dominating others; (2) its robustness to scale, decreasing the sensitivity of algorithms to scale and ensuring that the optimization process behaves consistently across different features; (3) the easier interpretability of scaled features due to them being centered around zero and having a standard deviation of one; (4) its better handling of outliers, whose impact is reduced by shrinking the range of values and making the distribution more Gaussian-like [13,16-18].

In contrast, range scaling (or mean normalization), uses the biological range of metabolites as its scaling factor, centering the data around the mean value and scaling it to a range of [-1, 1]. Similarly, min-max scaling is a popular normalization technique that transforms the data into a standard scale, typically between 0 and 1, by subtracting the minimum value (instead of the mean) of a specific feature and then dividing the difference by the range of values in that feature. The fundamental distinction between these methods and other scaling techniques is that column-wise normalization procedures (performed on metabolites) keep all values within a specific range, whereas scaling approaches only adjust the spread/variability of the data. Normalization methods are particularly useful when there is a desire to bring the scale of features that vary significantly to a comparable range. Their disadvantages, however, are the inflation of measurement errors and the higher sensitivity to outliers, since only two values are used to estimate the biological range (the minimum and maximum values), while all measurements are taken into account for calculating the mean or standard deviation used in other scaling techniques [16,17].

To finalize, data transformations are often necessary to adjust the variance of non-induced biological variation, whose correlation with the corresponding mean abundance of metabolites results in considerable heteroscedasticity in the data, which in turn has an impact on the subsequent data analysis. Therefore, the goal of data transformations is to correct for heteroscedasticity by converting multiplicative relations into additive relations and making skewed distributions more symmetric. Generally, the log and power transformations are the most used non-linear conversions of data in metabolomics to reduce heteroscedasticity [13,16].

Now, the metabolomics data is ultimately fit for statistical analysis.

2.1.5. Objectives of longitudinal analysis

In the field of health sciences, longitudinal studies play an important part in enhancing the understanding of disease development and persistence. The distinguishing feature of this type of study is that the participants are measured repeatedly on the same outcome throughout its duration, with two or more observations being taken at different time points, thereby allowing the assessment of within-individual changes in the response variable over time. Usually, the main interest of a longitudinal analysis is to uncover trends in these response trajectories and determine whether those changes are related to selected covariates (e.g., treatment group) [19-22].

2.1.6. Sources of correlation in longitudinal data

Despite the natural heterogeneity that exists among individuals in many extraneous variables (e.g., gender, socioeconomic status, and many genetic, environmental, social, and behavioral factors) that can influence the response to a certain stimuli, longitudinal studies are still capable of providing a very precise estimation of within-individual change. By performing multiple comparisons between the observations of each individual, every participant acts as their own control in a longitudinal analysis, therefore removing these extraneous, but unavoidable, sources of variability among individuals [19].

Nevertheless, some types of variability that can have an impact on the correlation between repeated measurements of the same individual: (1) between-individual heterogeneity, (2) within-individual biological variation, and (3) measurement error [19].

The first source of variability is between-subject heterogeneity, which reflects the natural variation in the propensity of individuals to respond, whether it be higher or lower than average. Since each individual's underlying tendency to respond is shared in all of the repeated measures obtained on that individual, the values of response gathered at one instance are expected to remain "high", "medium" or "low" at subsequent occasions, creating a positive correlation among repeated responses from the same individual, which are inevitably more similar than the responses across different individuals. This, in turn, results in a between-subject variance that does not remain constant over time due to the multiple diverging trajectories of response observed amongst the participants of a study [19-21].

The inherent biological variability of many health outcomes is another important source of variability that has an impact on the correlation among longitudinal responses. This within-individual biological variability is evident in almost all measured biological parameters, for example, serum cholesterol, blood pressure, and heart rate. Although some health-related variables have predictable cyclical rhythms (e.g., daily body temperature or monthly estrogen levels), most of them vary around an homeostatic set point in a random manner, and can fluctuate considerably over relatively short intervals of time. As a result, random deviations from an individual's response trajectory are likely to be more similar when measurements are obtained very close together in time. In other words, measurements taken closely together are typically more highly correlated than measurements that are further separated in time, thus introducing a serial correlation among repeated measures that decreases as the time separation between them increases [19].

The final source of variability in longitudinal data is the random measurement error, identified as a ubiquitous component of almost all studies, longitudinal or not. Because virtually all measurements are fallible, it is useful to quantify the relative magnitude of errors associated with a given measurement procedure, which is commonly done through calculating the variance of the measurement error. In general, the larger the variance of the measurement errors, the greater is the attenuation of the correlation among repeated measures in a longitudinal study, implicating that less reliable measurement procedures result in repeated measurements with smaller correlations compared to more reliable methods [19].

2.1.7. Generalized linear models for longitudinal data

Generalized linear models (GLMs) are a broad class of models widely used in biomedical research that are designed for the regression analysis of discrete or continuous responses with independent observations. Though their straightforward application to longitudinal data is not appropriate due to the existing correlation among observations obtained from the same individual, some extensions of this group of models are capable of handling longitudinal responses [19,22,23].

When considering the overall aim of a longitudinal analysis to study how subjects change over time and what characteristics influence those changes, it is important to clarify the specific concerns that the study intends to address in order to choose a suitable model for the analysis of the longitudinal data. If one is interested in describing the evolution of each subject separately, then generalized linear mixed-effects models are more adequate as they can introduce subject-specific parameters called random effects that explain the dependence between responses from the same individual. On the contrary, when the interest is in the population-averaged evolutions, then marginal models are the best choice. In summary, the fundamental difference between mixed-effects models and marginal models is their target of inference being either subject-specific or population-averaged estimates, respectively [21,23-26].

2.1.8. Generalized linear mixed effects models

Generalized linear mixed models (GLMMs) are an extension of GLMs that expands on the ordinary linear regression model by incorporating individual-specific random effects, meaning randomly varying intercepts and slopes, into the linear predictor function to provide a probability model that explains the correlations found in repeated measurements, thus accounting for both within-subject associations and between-subject variability [19,21,27-30].

The basic premise underlying the use of GLMMs for longitudinal data is the assumption of heterogeneity across individuals in the study population in a subset of the regression coefficients from a generalized linear model. By considering that some of the regression coefficients in the statistical model vary randomly across individuals according to a specific distribution while remaining constant within each subject, this random-effects approach is capable of reflecting the natural heterogeneity that exists among individuals due to unmeasured factors. Therefore, the response of a GLMM is modeled as a function of the expected value for the outcome variable through the development of a linear relationship with the explanatory variables. These mixed effects models are also typically referred to as conditional models since they allow the estimation of different parameters for each subject or cluster, resulting in parameter estimates that are conditional on that subject or group. This essentially means that the response is conditional on the random effects of the model [19,20,22-25,29].

Briefly, the development of a mixed effects model is divided into a two-step approach. In the first stage, a separate linear regression is specified for the observations of each subject as a function of selected covariates made on those observations (such as time of observation, in the case of longitudinal data), thus creating as many regressions as there are subjects in the study. Notably, each regression must use the same set of predictor variables in order to have the same vector of coefficients, though the regression coefficients themselves are allowed to vary over individuals. At the second stage, the regression coefficients modeled in stage one become the random outcome variables, hence the name random effects model. Ultimately, the term mixed effects stems from the combination of the two regressions into a single equation with two types of regression coefficients, named fixed effects and random effects, that model both kinds of within-subject (stage one) and between-subject (stage two) variability in the data [21,22].

The biggest advantage of GLMMs is their incredible flexibility. In addition to handling several types of data including clustered individuals and repeated measures, they can also deal with crossed or nested random factors, as well as manage both continuous and categorical time variables. On top of that, covariates can be measured just once per individual or repeatedly at each observation. Finally, and contrary to other multivariate models, GLMMs can be applied to unbalanced data that has missing outcomes for some individuals, something that is very common in longitudinal studies, without needing to drop those values from the model [20-22,28,29].

Despite being similar to ordinary multiple regression in many aspects, however, the process of fitting GLMMs is more computationally complex and intensive compared to most multivariate models, partially because of its acceptance of correlation between observations requiring additional work to specify models and assess goodness-of-fit. Still, the extra complexity involved in these models is compensated for by the additional flexibility they provide in model fitting [21,22,27].

Another important feature of GLMMs is their amenability for dealing with non-normal data, a problem often miss-handled by researchers. Many attempted shortcuts, such as transforming data to achieve normality, can end up ignoring the random effects or treating them as fixed factors. Even more, they

may violate statistical assumptions (e.g., homogeneity of variance across groups in non-parametric tests) or limit the scope of inference (one cannot extrapolate estimates of fixed effects to new groups). For this reason, GLMMs are considered the best tool for analyzing non-normal data involving random effects, with its only requirements being the specification of a distribution, link function and structure of the random effects. For mathematical and computational convenience, the random effects in GLMMs are frequently assumed to have a multivariate normal distribution. Furthermore, the linear mixed effects model is a special case of the generalized linear mixed effects model where the conditional mean, given the random effects, is related to the covariates via an identity link function, and hence the conditional distribution of the responses is assumed to be normal [19,27,31].

2.1.9. Marginal models for longitudinal data

The marginal model is another extension of the GLMs that, instead of considering the within-subject associations as random effects, incorporates them directly with the repeated measurements. These models are primarily used to make inferences about population means, reason why they are also known as population-averaged models. In general, marginal models are preferred over mixed-effects models when the overall population effect of a treatment/experimental condition is the primary interest of the study [19,28].

To estimate the regression parameters in a marginal model, some assumptions need to be made about the marginal distribution of the response at each occasion (more specifically about the mean and its dependence on the covariates), as well as the pairwise within-subject associations among the responses, thereby linking repeated observations from the same subject. Even though the joint distribution of the vector of responses is not fully specified through the marginal means, variances, and pairwise associations, these assumptions are sufficient to estimate and construct confidence intervals for the regression parameters. Moreover, by modeling the mean response separately from the covariance matrix of within-subject associations, marginal models are able to ensure that the interpretation of the regression coefficients does not rely on the assumed model for the covariance of repeated responses, thus allowing correct inferences to be made about changes in the average response of the population [19,24,30].

2.1.10. Estimation of marginal models through generalized estimating equations

Generalized estimating equations (GEE) are a statistical approach commonly employed to fit marginal models for clustered or repeated data analysis. Based on the concept of "estimating equations", the GEE model provides a very general and unified approach for analyzing correlated responses, whether they are discrete or continuous. The essential concept behind this method is the generalization and extension of the usual likelihood equations for a generalized linear model of a univariate response by incorporating the covariance matrix of the vector of responses [19,23,24,30].

As a type of marginal model, GEE follows a population-level approach based on a quasi-likelihood function to provide population-averaged estimates of the parameters in study. When estimating GEE models, the correct specification of the univariate marginal distributions requires the adoption of a “working” assumption about the association structure of the repeated measures. Through a combination of estimating equations for the regression parameters and moment-based estimation for the correlation

parameters entering the “working” assumptions, GEE models can estimate the coefficients associated with the expected value of an individual’s vector of responses, which are then used to estimate the average response of the population. Essentially, this means that the GEE model is averaging out any random effects to produce marginal estimates. It is worth noting that, even if the correlation structure is misspecified, the GEE model is still able to provide valid asymptotic confidence intervals for its population average estimate. On the other hand, GEE models can sometimes require large sample sizes in order to be sufficiently accurate, and are not very robust when handling non-randomly missing longitudinal data. Typically, GEE models are used for non-normal data [23-27,30].

Overall, the defining features of GEE models can be described as: (1) the variance-covariance matrix of responses is treated like nuisance parameters, making the model fitting process easier compared to mixed-effects models. (2) Under mild regularity conditions, the parameter estimates are consistent and asymptotically normally distributed even when the “working” correlation structure of responses is misspecified. (3) The distribution assumption is more relaxed and only requires the correct specification of the marginal mean and variance, as well as the link function that connects the covariates of interest to the marginal means [26,30].

2.2. Choosing the right approach

The following statistical analysis of the longitudinal metabolomics data set in study was conducted using the R Statistical Software (v4.3.2), with the exception of the pathway analysis done through the MetaboAnalyst Software (v6.0) [32,33].

2.2.1. Imputation of zero values

After receiving the clean metabolomics matrix, the first processing step was to determine the prevalence of missing/zero values in the data set. A breakdown of the different time points measured in the study revealed that two individuals had missing values for every metabolite in the t2 and t3 instances that correspond to the responses immediately before and after the second vaccine dose, which is consistent with the one-dose vaccination plan (J&J vaccine) that was administered to both of these patients. Because of this discrepancy, the two individuals with only three vaccination instances were ultimately withdrawn from the analysis so that the remaining cohort of patients with five measured time points could be more easily compared.

Along with the N/A cases, a high number of zero values was also discovered, comprising approximately 11% of all non-missing values. The initial answer to this problem was to discard any metabolites that presented a mean percentage of zero values superior to a chosen threshold of 50% in all samples and time points. However, this approach only led to the removal of 3 out of 85 metabolites from the data set, leaving still 9% zero values spread across 20 individuals and 5 time points.

Because there was a question on whether these zeros corresponded to the biological absence of metabolites in the samples or censored abundances below the limits of quantification of the mass spectrometer, it was important to continue the imputation process using other methodologies to further reduce the amount of zeros in the data with the hope of retaining only those who were biologically correct. For this purpose, two of the most commonly used imputation algorithms in metabolomics, the k-nearest neighbors and the quantile regression imputation of left-censored data, were compared to

investigate which method was more adequate to estimate values in the data set. Despite also being prominently used with mass spectrometry data, the random forest algorithm was not included in this comparison due to its more computationally demanding technique, which would likely delay the results of the analysis without bringing sufficient benefits to the value imputation process. Furthermore, two separate distance measures, cosine and Mahalanobis, were selected to test the kNN algorithm and determine if one could perform better than the other. Their invariance to scale was particularly important when choosing these distance measures as it is known that abundances can vary drastically between metabolites, leading to big changes in scale that significantly affect the estimates obtained from the kNN method.

The kNN imputation of zero values was accomplished through the development of a new R function, named “knnImputation”, that was capable of identifying any zero value as missing data needing imputation, while also treating all other zero values as non-missing data that could be taken into account for the estimation process, something that existing kNN algorithms were unable to achieve. To use this function, two lists of data frames were required as input: one with the concentration matrices of metabolites separated by time point and another with the matching distance matrices for each patient.

After removing from the data every N/A case related to non-vaccination instances, as well as the three metabolites with more than 50% zero values, the metabolomic data was divided into subsets of the different time points. Each distance matrix associated with a particular time point was obtained by calculating the cosine or Mahalanobis distances between the rows of the concentration matrix holding the metabolite abundances for that specific time point. Cosine distances were directly calculated from the metabolomic data using the “dist (method=cosine)” function available in the proxy R package (v0.4.27), whereas Mahalanobis distances were estimated as the Euclidean distance of the re-scaled loadings of a principal components analysis of the data with the help of the functions “prcomp (scale=TRUE)” and “dist (method=euclidean)” from the stats (v4.3.2) and proxy (v0.4.27) R packages, respectively [32,34]. Once all distance matrices had been attained, the “knnImputation” algorithm was employed to replace the zeroes in each data set with a mean concentration calculated from the k nearest neighbors of a specific patient. Whenever a zero value was encountered in a concentration matrix for a given time point, the function searched within the corresponding distance matrix for the five patients with the lowest distance measures from the “PatientID” where that zero was found, and used their abundances for that particular metabolite to calculate a mean concentration value, which would then become the new imputed value. For more details, the complete code for the “knnImputation” function is presented in Annex 6.1.

Conversely, the QRILC imputation of zero values was performed simply by using the function “impute.QRILC” from the imputeLCMD R package (v2.1) for each individual time point. Following both imputation methods, the different time point matrices were joined back together in order to return the data set to its original arrangement [32,35].

2.2.2. Scaling of longitudinal data

In the analysis of longitudinal data, standardization methods like autoscaling can lead to several complications, such as undesired changes in the distance between observations, that are related to the usage of different frames of reference (the original response scale, the intra-individual distribution, the inter-individual distribution within given time points, the inter-individual distribution across different time points, the variation within vs. between cohorts, and any combinations of these) [36]. Some of the

problems that often arise in longitudinal studies concerning the standardization within/across time points and individuals are the following:

(1) Standardizing repeated measures within individuals hinders the examination of mean-level differences between individuals since each individual's mean score becomes zero, and thus the standardized means do not inform whether the individuals differed in their original experiences;

(2) Standardization across individuals within measurement time points encumbers the observation of mean-level changes from one time point to another because all of the means become zero at every time point, whereas the raw-score means might have shown interesting alterations in the measured variable;

(3) Standardization across time points and individuals obfuscates the relative rank of an individual at a certain time point, resulting in confounded information about the time point-specific relative rank-orders and the mean-level changes [36].

To circumvent this issue, a normalization method similar to range scaling was developed based on the longitudinal nature of the data in order to allow for a straightforward illustration of the metabolites' evolution over time. Using the expression $(X_{ij} - baseline_{ij})/(max_{ij} - min_{ij})$ for each individual i and metabolite j, this new approach to normalizing data in longitudinal studies was able to rescale the data as individual's variations of concentration relative to the baseline of the metabolites' natural abundance in RBCs. Hence, the normalized value of abundance X could be determined by subtracting the t0 time point of metabolite j and individual i, representing their initial state (or baseline) when the experiment began, and then dividing the result by the biological range of that metabolite within that specific individual during the entire duration of the study, which could be calculated through the difference between the max and min values measured for the metabolite and individual in question. This allowed to center the data around the starting values of each metabolite and patient, scaling it to a range of [-1, 1] that represents how much the abundances increased or decreased over the course of the vaccination process relatively to their original concentration, therefore highlighting the prevailing metabolite fluctuations which are the primary focus of this work. More information on the "dataNormalization" function applied in this work can be found in Annex 6.2.

2.2.3. Algorithm performance test

To find the most suitable pre-treatment procedure for the metabolomics data, the developed methods of imputation and normalization were tested for best performance and ideal order of application. This performance testing was done by randomly selecting five patients (of the 20 individuals included in the analysis) and ten metabolites (from a subset of 59 metabolites with a mean percentage of zeros no higher than 10%) for each time point and replacing the corresponding metabolite concentration values with N/A. Then, the different pre-treatment methods were applied, and the resulting imputed and normalized data was saved. Here, some modifications were required to adapt the various functions used in this process to the new data set with N/A values. One of these changes was the use of the "pca" function from the R package SYNCSA (v1.3.4) in the first step of the Mahalanobis distance calculation (maintaining the necessary scaling of the data). Also mandatory, an updated version of the "knnImputation" function called "knnImputationForNA" was designed (Annex 6.3) [32,37].

For performance assessment, the pre-treatment method testing was repeated 1000 times. Finally, the obtained values were compared to the normalized original data using four error measures (MAE, RMSE, RAE, and RRSE), which were calculated using the Metrics R package (v0.1.4). Significant differences between the results of the pre-treatment methods were determined via Wilcoxon signed-rank test for two paired samples using the stats R package (v4.3.2). The method with the best performance was used to impute and normalize the data set in study [32,38].

2.2.4. Development of linear mixed models

In a first attempt to analyze the behavior of each metabolite through the COVID-19 vaccination, linear mixed models were created to represent the longitudinal data through the "lmer" function from the lmerTest R package (v3.1.3), using the "TimePoint" as the independent/explanatory variable and incorporating the "PatientID" as a random effect that introduces some variability into the model derived from innate differences between individuals. Here, the application of "lmer" was equivalent to using "glmer" with an identity link function since both assume that the response variable follows a Gaussian distribution, which is consistent with the [-1, 1] range of the data after normalization. For each metabolite, different time points were used as the reference level of the constructed model to obtain all possible coefficients, thus including not only intercept coefficients (t1-t0, t2-t0, t3-t0, and t4-t0) but also every other difference observed between time points (t2-t1, t3-t1, t4-t1, t3-t2, t4-t2, and t4-t3). P-values obtained from these models were adjusted using the Benjamini-Hochberg procedure, which implements the false discovery rate (FDR), and then used to assign the confidence level of each model coefficient. The metabolites with corrected p-values lower or equal to 0.1 were considered has having significant changes in concentration over time. Plots of metabolite abundance over time were constructed with the help of the R packages emmeans (v1.10.0) and ggplot2 (v3.5.0) [32,39-41].

For these significant metabolites, model assumptions of normality and homoscedasticity were evaluated through an analysis of the residuals, which was accomplished using the R packages stats (v4.3.2), graphics (v4.3.2), and car (v3.1.2). For this appraisal, normal Q-Q plots and histograms of residuals were used to determine if the residuals of each regression model were normally distributed, while "residuals vs fitted" plots were constructed to assess whether the residuals exhibited non-linear patterns indicative of data heteroscedasticity. Because this last criterion was not met, meaning that the variance of the residuals was not constant across time points, the application of LMMs for the analysis of the longitudinal metabolomic data was discontinued in favor of a modeling approach with less strict assumptions and, therefore, an easier estimation process [32,42].

2.2.5. Establishing the generalized estimating equations

Following the discovery of data heteroscedasticity in LMM models, a second effort was made to model the metabolites' concentrations through Generalized Estimating Equations (GEE). Taking advantage of the "geeglm" function from the R package geepack (v1.3.10), using "TimePoint" as the explanatory variable, "PatientID" as the sample id, and considering the Gaussian family as the approximate distribution of the population mean, QIC measures were calculated via geepack (v1.3.10) to establish the optimal correlation structure for the data among three of the most commonly chosen options by researchers: independent, exchangeable and unstructured. To explain their differences, an independence correlation structure means that the responses within subjects are not correlated, whereas a model with

an exchangeable correlation structure implies that all pairs of responses within a subject are equally correlated. Lastly, the unstructured correlation structure allows all correlations to vary freely, though this is often avoided unless there are strong reasons to believe that a simpler structure would not suffice. When comparing the different "working" correlations from the same metabolite, as well as the QIC and QICC parameters obtained, the exchangeable correlation structure was found to be the best compromise between having lower standard deviations than an independent correlation while also presenting an inferior QICC value than the unstructured matrix [32, 43-46].

Grounded on these results, the GEE equations were built with an exchangeable correlation structure through the same "geeglm" function from geepack (v1.3.10), taking into account the previously mentioned parameters. The same strategy used in LMM for obtaining every single coefficient was employed via selecting different time points as the reference level of each model. P-values for these coefficients were calculated through Wald tests and adjusted using the Benjamini-Hochberg procedure, with the FDR-adjusted p-values being used to assign the confidence level of each model coefficient. The metabolites with corrected p-values lower or equal to 0.1 were considered has having significant variations in concentration over time. For these significant metabolites, GEE effects plots were created using the "ggemmeans" function from the ggeffects R package (v1.5.1), as well as other functions from the ggplot2 R package (v3.5.0) [32,41,44-47].

Because GEE is a semi-parametric approach that does not make any particular assumption for the outcome variable distribution, only the mean of the data requires the specification of an appropriate link function, allowing for some flexibility in the functional form of the covariates, namely potential non-linear terms. Since the identity link chosen in the GEE models only serves to explain the population mean, it is possible that the complete variation of the data deviates from such distribution, reason why the models are not accompanied by any distributional assumptions nor there is a requirement to check them. Hence, for the GEE model validation, only the model assumption of homoscedasticity was judged for the significant metabolites via "residuals vs fitted" plots. In this case, the line across the center of the plots was roughly horizontal for every metabolite analyzed, indicating that the residuals followed a linear pattern (their variance was constant for all time points), therefore suggesting that there was no heteroscedasticity in the data [32,48].

As a further test of the relevance of each significant metabolite discovered by this methodology, their GEE models were compared to the corresponding null counterparts through an ANOVA test (glmtoolbox R package v0.1.10) to investigate which of them presented significant differences, ranking the resulting p-values corrected with FDR. In this analysis, it is important to note that some models can have coefficients with a significance level equal or lower than 0.1 and still not be distinguishable from the null model. In these cases, it is likely that those coefficients do not have enough strength in the model to compensate for the non-significant ones and subsequently pull the model away from its null equivalent [32,49].

Finally, the significant metabolites were grouped according to their response trajectories. To accomplish this classification, a concordance correlation coefficient was created based on the trajectory differences between time points. If the intercept coefficient (corresponding to the normalized abundance) of a metabolite increased from one time point to the next, then a positive score of 1 was attributed to that time difference. On the contrary, decreases in metabolite abundances were assigned a negative score of -1. The concordance coefficient was then calculated as the mean of these scores for a specific metabolite, which could result in a correlation coefficient between 0, indicating opposite metabolite trajectories, and 1, representative of two identical trajectories of metabolite abundance. Any other values obtained for this coefficient (0.25, 0.50, or 0.75) were not considered as relevant for the analysis since they suggested that there was no definitive correlation between the metabolite trajectories in question.

Moreover, given that the positive and negative associations uncovered between metabolites were only considering their general variations in concentration over time, an additional calculation of trajectory magnitudes was done for the correlated metabolites to better understand which of them shared the closest responses to COVID-19 vaccination. These differences in magnitude were determined by the norm of the vector containing the differences between the intercept coefficients of two metabolites for every time point [32].

2.2.6. Pathway analysis with Metaboanalyst

Using the MetaboAnalyst 6.0 software, the significant metabolites found by GEE were analyzed concerning their annotated features. To conduct this search, a list of the compound names was added as input, indicating the feature type as metabolites when required. For the pathway analysis, default parameter settings were employed regarding the visualization method (scatter plot), enrichment method (hypergeometric test), and topology measure (relative-betweeness centrality). Other available options for these parameters (Fisher's exact test and out-degree centrality) were also tested and produced the same results. The Homo sapiens KEGG database was selected as the pathway library and all metabolic pathways with at least one entry were considered. For the enrichment analysis, the RaMP-DB metabolite set library was adopted for the reference metabolome due to its integration of pathway databases from KEGG (via HMDB), Reactome, and WikiPathways. Only metabolite sets containing at least two entries were used [33,50].

3. RESULTS AND DISCUSSION

3.1. Data description

The metabolomic data set used in this work was composed of 20 individuals with a two-dose vaccination regimen, 82 metabolites with less than 50% zeros, and five time measurements for each patient and metabolite combination, comprising a total of 8200 values. The data was formatted with 84 columns of variables (PatientID, TimePoint, 82 Metabolites) and 100 rows of samples (20 patients, 5 time points for each patient). Each metabolite in the data set represents a different response variable that requires a unique model to be created separate from the others, though always using the TimePoint as the (categorical) explanatory variable with five different levels (t0, t1, t2, t3, and t4).

Of note, the information gathered for every participant of the study regarding their gender, age and BMI, as well as the type of vaccine administered, were omitted from the data set and not considered for any subsequent data analysis.

3.2. Comparison of pre-treatment methods

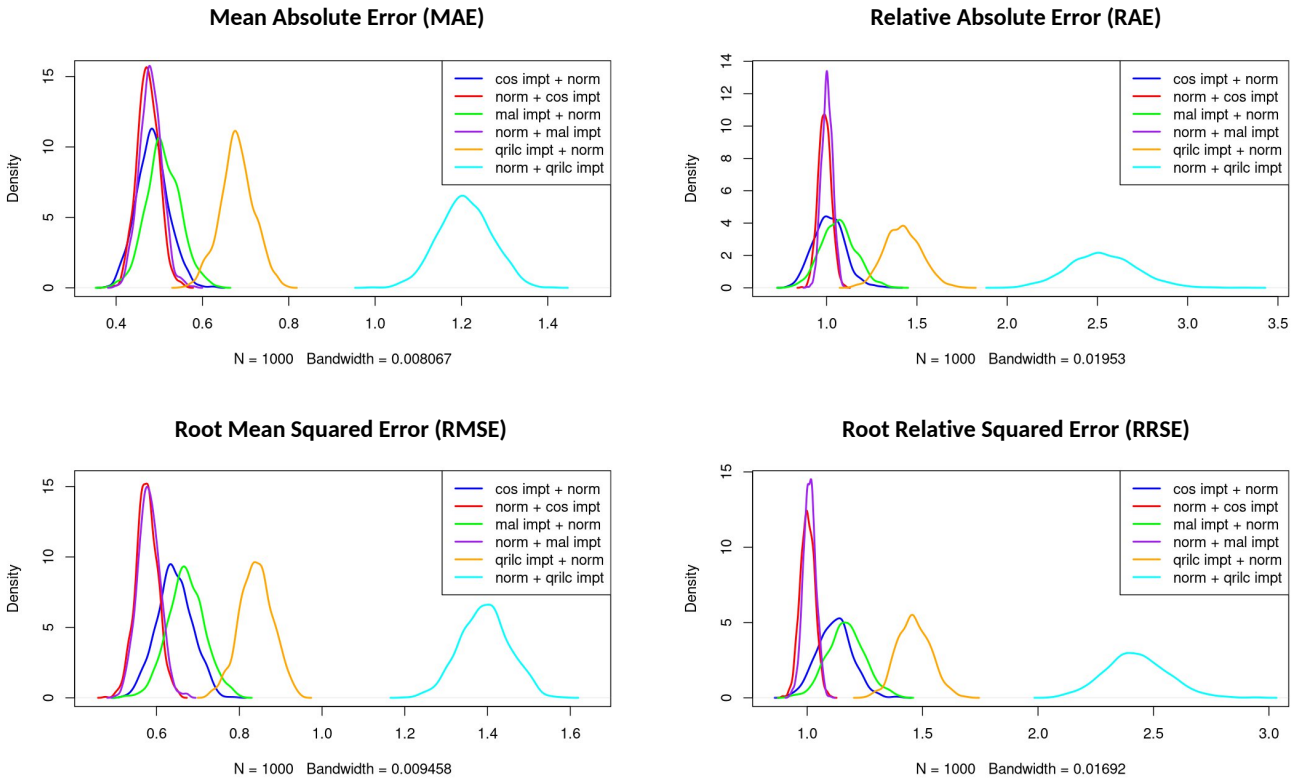
During the performance testing of the imputation and normalization procedures, four error metrics were used to compare the different methods: MAE, RMSE, RAE, and RRSE.

The Mean Absolute Error, or MAE, represents the average absolute difference between the predicted values of a model (in this case, the imputation algorithm, with prior or subsequent normalization of the data) and the actual observations obtained after normalization without any estimated values. In other words, MAE is a measurement of the average of the residuals in the data. Similarly, the Root Mean Squared Error, also called RMSE, determines the square root of the average squared difference between the predictions made by the model and the true observations in the data set, therefore giving the standard deviation of the residuals. Thus, when applied for model evaluation, lower values of MAE and RMSE imply a higher accuracy of the model predictions [51,52].

In turn, the Relative Absolute Error, known as RAE, provides the absolute error of the model's predictions relative to a trivial model that predicts the mean for every data point, while the Root Relative Squared Error, RRSE, calculates the squared error of the predictions relative to that same trivial model. Both of these metrics are expressed as a ratio that compares the model's forecast errors to the prediction errors of the simpler model, so that, when measuring the performance of different predictive models, lower error metrics translate to more accurate predictions made by the model of the real observations in the data. Because of that, a good forecasting model is one that produces a ratio closer to zero, whereas a poor model (one that is worse than the simple predictor) generates a ratio greater than one. Ultimately, a reasonable predictive model needs to have a ratio lower than one in order to deliver results that are better than the forecasts of a trivial model [53-55].

To compare the results of every method tested in this work, graphical distributions of the several 1000 value arrays obtained for each error measure were constructed (Figures 3.1 to 3.4), from which the

lowest mean and standard deviation were used to gauge the best-performing imputation and normalization technique.



Figures 3.1 to 3.4. Performance assessment of each imputation and normalization procedure using four error measures: Mean Absolute Error, Relative Absolute Error, Root Mean Squared Error, and Root Relative Squared Error. The evaluated methods were the following: kNN imputation with cosine distance + normalization (cos impt + norm); normalization + kNN imputation with cosine distance (norm + cos impt); kNN imputation with Mahalanobis distance + normalization (mal impt + norm); normalization + kNN imputation with Mahalanobis distance (norm + mal impt); QRILC imputation + normalization (qrilc impt + norm); normalization + QRILC imputation (norm + qrilc impt).

Of the two types of imputation evaluated, kNN and QIRLC, the kNN algorithm was shown to be the most capable of accurately replacing unknown values based on the remaining data sample, apparent by the lower mean errors achieved across different metrics. Additionally, data pre-treatment methods using the kNN algorithm demonstrated lower standard deviations for every error measure when normalization was performed before imputation, indicating that this should be the preferred order of method application.

Having this information, Wilcoxon signed-rank tests were conducted between the two kNN strategies with previous normalization that were differentiated only by the cosine or Mahalanobis distance measure employed in the algorithm. Paired samples were used in these comparisons to reflect the strong connection that exists among each couple of values from the two data groups because they belong to the same individual, metabolite and time point, making their independence impossible. From the outcome of these statistical tests, it was revealed that normalization followed by kNN imputation using cosine distance was the pre-treatment methodology that presented the lowest mean value for all error metrics calculated, meaning its results were the closest to the original data. It is worth mentioning, however, that the mean MAE and RMSE errors for this estimation procedure (0.4675 and 0.5694, respectively) are relatively high compared to the range [-1, 1] of values in the normalized data. The same can be said for

RAE and RRSE given that their mean values (0.9678 and 0.9827) are admittedly close to 1, suggesting that the chosen model has a very similar, albeit a little superior, predictive capacity compared to the average forecasts of a trivial model. Still, since a small improvement was indeed observed, the decision was made to proceed with the imputation of zero values in the data using the selected approach, instead of relying on a more simple and generic strategy such as the average of non-missing values.

Given the superiority of kNN over QIRLC that was observed in the imputation of the metabolomic data set, it is plausible to assume that the majority of zero values present in the data were not a result of LOD-censored abundances, for which the QIRLC algorithm would have had a better estimation performance. Instead, the results suggest that those values were more likely driven by the biological absence of metabolites in the blood samples or peaks that were accidentally missed during the pre-processing phase of the LC-MS/MS data. To validate this assumption, an additional comparison using the RF imputation algorithm could be carried out in future research to confirm the predominant types of missing values in the data since it is a popular method with a reliable estimation capability for MCAR and MAR, the same missing data that is thought be the origin of the zero values in this work. If these suspicions are corroborated, then a new assessment should be made on whether the RF technique has a higher estimation accuracy than the kNN process chosen for this study.

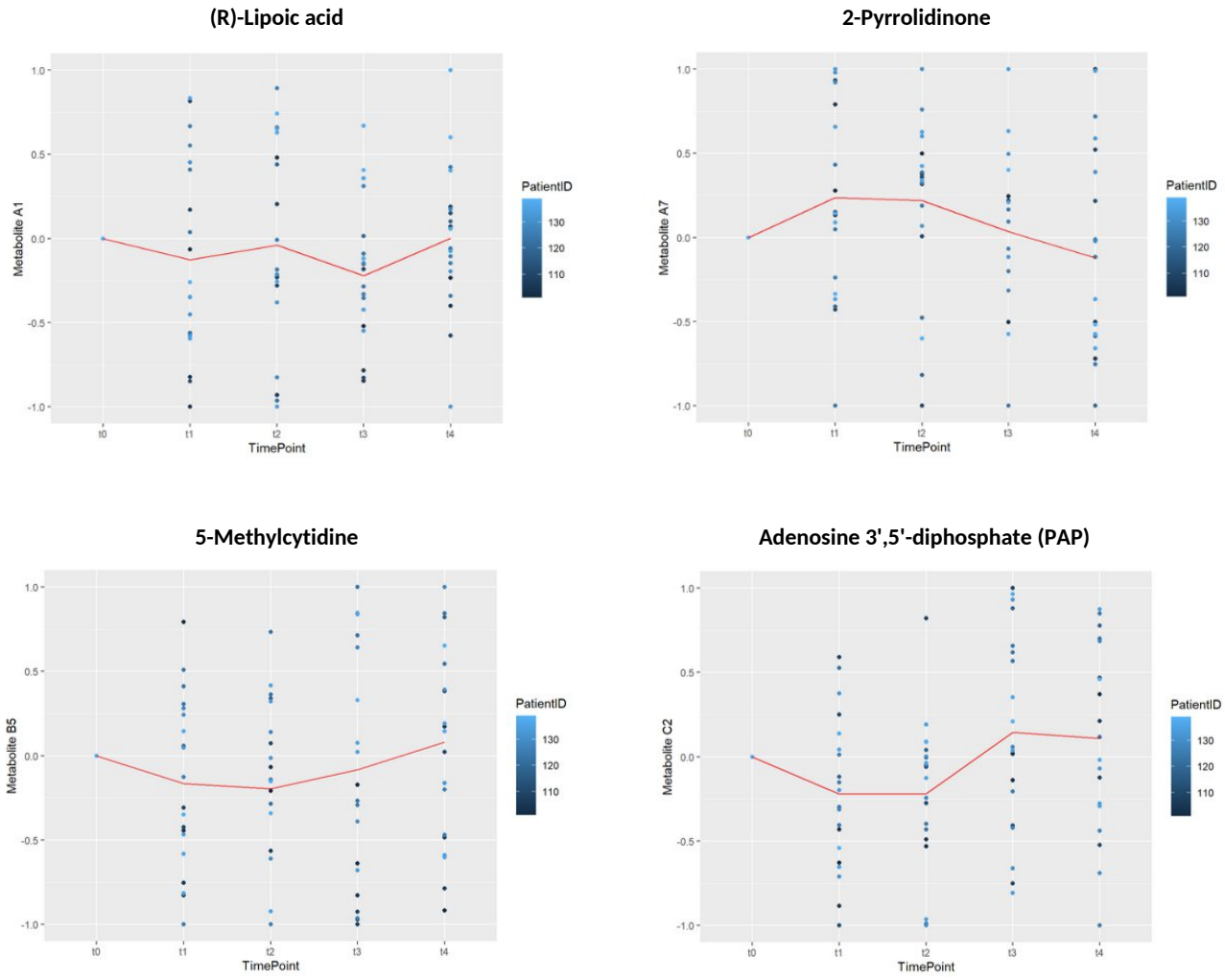
Another option that could be worth discussing is the use of the normalized data for statistical analysis without any imputation. Though this alternative was not attempted in the current study, it presents an interesting path to explore in future studies. By comparing different statistical models with and without data estimation, it would be possible to investigate how various imputation methods influence the modeling process and its statistically significant results. Nevertheless, because this type of exploratory analysis was not the focus of this work, the statistical analysis of the new normalized and imputed data set was continued.

3.3. Data modeling through LMM vs GEE

Linear mixed models created for each metabolite were designed following the expression [56]:

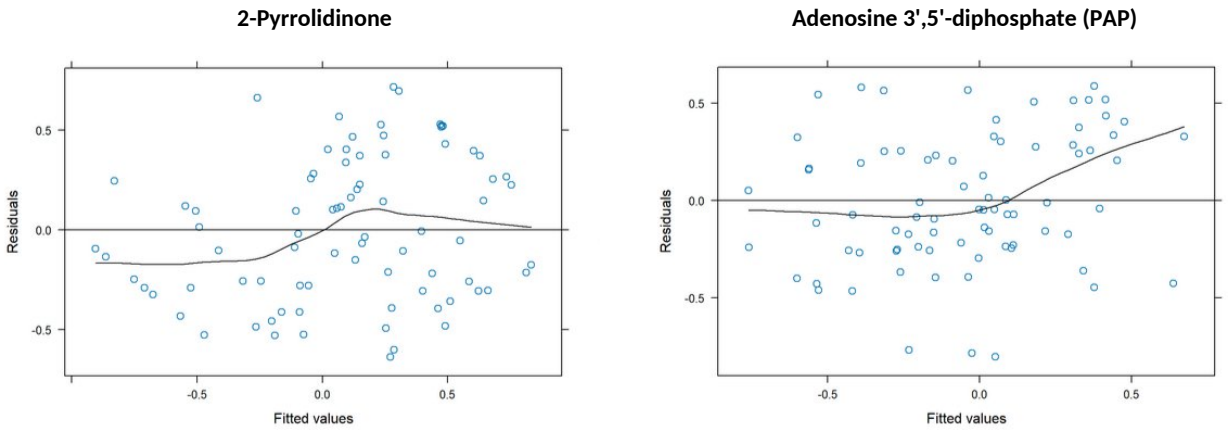
$$Y_i = X_i\beta + U_i\gamma_i + \epsilon_i \tag{3.1}$$

Where Y_i is the observed response, meaning the normalized abundance of the metabolite over time, $X_i\beta$ depicts the fixed effects of each time point on the concentration values, $U_i\gamma_i$ represents the random effects introduced by the different patients that participated in the study, and ϵ_i is the random error of the model predictions [56]. Some of the resulting models are presented in Figures 3.5 to 3.8.



Figures 3.5 to 3.8. Plots of the LMM models constructed from the normalized abundance changes observed in four different metabolites over time during COVID-19 vaccination. Present in these plots, the blue dots within each time point represent the measurements obtained from different patients, while the red line depicts the predictions of the LMM model.

From these LMM models, a strong data dispersion was detected within the different time points, indicating that patient measurements for the same metabolite varied widely from one another. Yet, despite such divergences, 24 metabolites were positively identified as having as least one significant coefficient in their LMM models. Unfortunately, diagnostic plots of those models (examples shown in Figures 3.9 and 3.10) demonstrated the existence of non-linear patterns in the variance of the residuals, highlighting the presence of heteroscedasticity in the data. Because LMM models are constructed on the assumption of data homoscedasticity, this regulation infringement meant that the LMM predictions could have a speculative nature, and hence no further statistical analysis should be conducted based on such outcomes.



Figures 3.9 and 3.10. Diagnostic plots of the LMM residuals of two metabolites against the fitted values of their respective LMM models. In each plot, the curved line represents the general pattern in which the residuals are distributed.

Taking their place in the modeling process, generalized estimating equations were used to study the behavior of each metabolite over time and determine which of them displayed significant alterations in concentration during the vaccination process.

As described by Wang (2014), the GEE notation is based on a set of longitudinal data consisting of K subjects with a total of N observations. For each subject $i(i = 1,2,\ldots,K)$, there are n_i observations where Y_{ij} denotes the j^{th} response $(j = 1,\ldots,n_i)$ and X_{ij} denotes a $p \times 1$ vector of covariates, with p being the covariates dimensionality. Moreover, $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})'$ is the response vector for the i^{th} subject with a mean vector $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ where μ_{ij} corresponds to the j^{th} mean. These responses are assumed independent across subjects but correlated within each subject. Finally, the marginal model is used to specify the relationship between μ_{ij} and the covariates X_{ij} through the equation [30]:

$$g\left(\mu_{ij}\right) = X_{ij}'\beta \tag{3.2}$$

Where g is a known link function and β is an unknown $p \times 1$ vector of regression coefficients with the true value as β_0. The conditional variance of Y_{ij} given X_{ij} is specified as $Var\left(Y_{ij}\middle|X_{ij}\right) = v\left(\mu_{ij}\right)\phi$, where v is a known variance function of μ_{ij} and ϕ is a scale parameter that often needs to be estimated. Mostly, v and ϕ depend on the distributions of outcomes. For instance, if Y_{ij} is a continuous response, then $v\left(\mu_{ij}\right)$ is specified as 1 and ϕ represents the error variance. However, when Y_{ij} is characterized as a discrete variable, $v\left(\mu_{ij}\right) = \mu_{ij}$ and ϕ is equal to 1. The variance-covariance matrix for Y_i is depicted as $Vi = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2}$, where $A_i = Diag\{v(\mu_{i1}),\ldots,v(\mu_{in_i})\}$ and $R_i(\alpha)$ is the so-called "working" correlation structure that defines the pattern of measures within subject, which is of size $n_i \times n_i$ and depends on a vector of association parameters denoted by α. In this work, the exchangeable correlation structure applied in the construction of the GEE models can be described by [30]:

$$Corr\left(Y_{ij}, Y_{ik}\right) = \begin{cases} 1 j = k \\ \alpha j \neq k \end{cases} \tag{3.3}$$

$$Sample \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix} \tag{3.4}$$

With the following estimators [30]:

$$\hat{\alpha} = \frac{1}{\left(N' - p\right)\phi} \sum_{i=1}^{K} \sum_{j \neq k} e_{ij} e_{ik} \tag{3.5}$$

$$N' = \sum_{i=1}^{K} n_i (n_i - 1) \tag{3.6}$$

$$\hat{\phi} = \frac{1}{N - p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} e_{ij}^2 \tag{3.7}$$

$$N = \sum_{i=1}^{K} n_i \tag{3.8}$$

$$e_{ij} = \left(y_{ij} - \mu_{ij}\right) \Big/ \sqrt{v(\mu_{ij})} \tag{3.9}$$

Using an identity link function and an exchangeable correlation structure, the GEE models revealed that 30 out of 82 metabolites had significant alterations in concentration along the different time points of COVID-19 vaccination, suggesting that they may have some involvement in the immunization process. The statistically significant coefficients obtained for each metabolite and corresponding confidence levels are presented in Table 3.1.

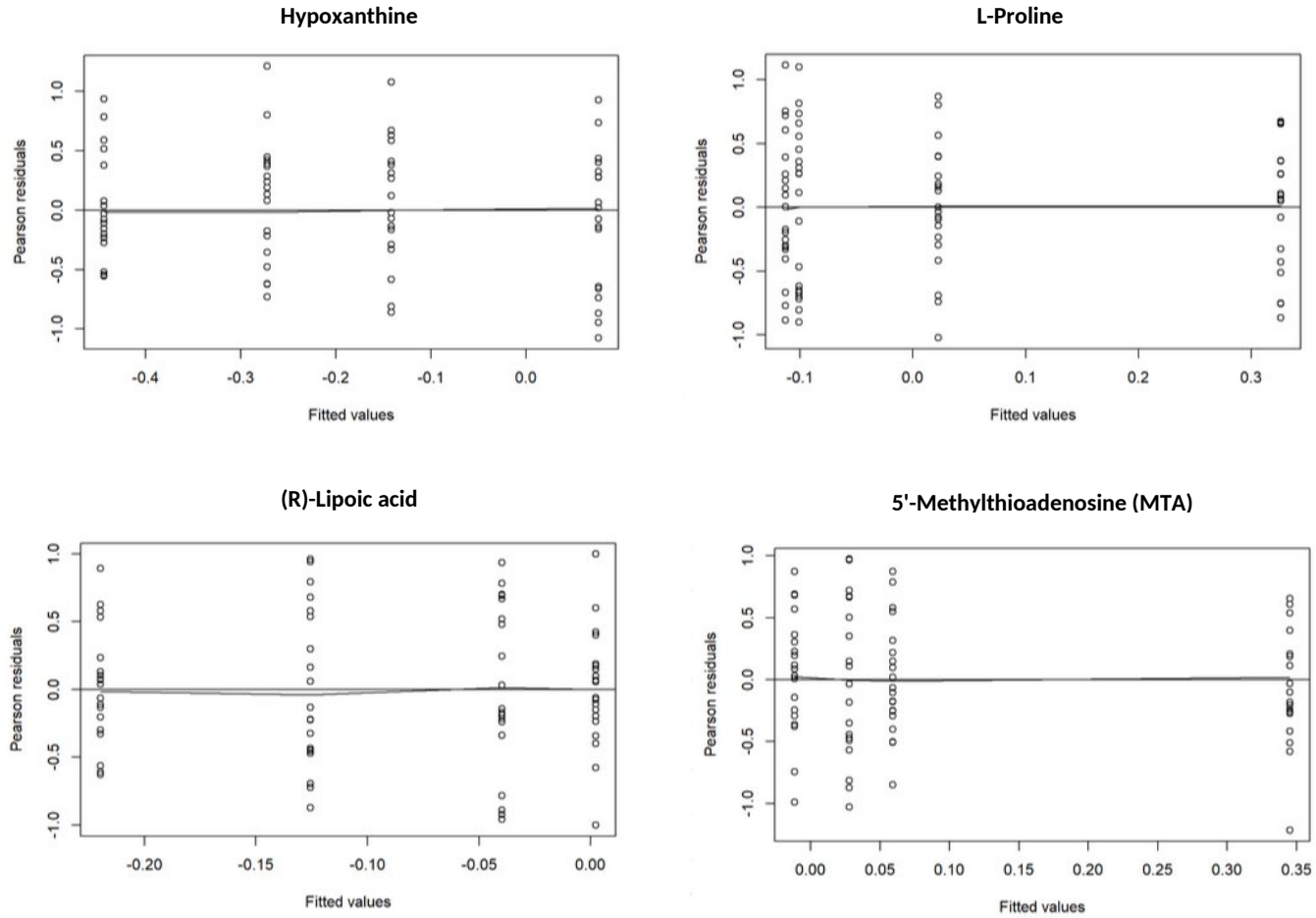
Table 3.1. List of GEE coefficients with statistical significance (p-value < 0.1) and respective confidence level for each of the 30 metabolites identified through the application of GEE models. All coefficients with a significance level of 0.1 or lower are shown.

Metabolite	Coefficient	Confidence Level
(R)-Lipoic acid	t3 - t0	0.05
	t4 - t3	0.05
2-Ethyl-2-hydroxybutyric acid	t4 - t0	0.05
2-Hydroxyadenine	t4 - t2	0.1
2-Pyrrolidinone	t4 - t1	0.01
	t4 - t2	0.1
5'-Methylthioadenosine (MTA)	t4 - t0	0.01
	t4 - t1	0.1
	t4 - t2	0.01
	t4 - t3	0.01
5-Methylcytidine	t4 - t1	0.05
	t4 - t2	0.1
Adenosine 3',5'-diphosphate (PAP)	t1 - t0	0.05
	t2 - t0	0.05
	t3 - t1	0.01
	t3 - t2	0.01
	t4 - t1	0.05
	t4 - t2	0.01
Adenosine monophosphate (AMP)	t1 - t0	0.01
	t2 - t1	0.01
	t3 - t1	0.1
	t4 - t0	0.1
	t4 - t2	0.05
	t4 - t3	0.1
Creatine	t4 - t1	0.01
	t4 - t2	0.05
Creatinine	t1 - t0	0.1
	t3 - t1	0.01
	t4 - t1	0.01
D-Arginine	t2 - t0	0.05
	t4 - t2	0.1
Glucaric acid	t4 - t0	0.01
	t4 - t1	0.05
	t4 - t2	0.01
	t4 - t3	0.05
Glycine	t4 - t2	0.1
Guanosine 5'-diphosphate (GDP)	t3 - t1	0.05
Guanosine monophosphate (GMP)	t1 - t0	0.05
	t3 - t1	0.1
Hyodeoxycholic acid	t4 - t0	0.05
Hypoxanthine	t1 - t0	0.05
	t2 - t1	0.01
	t3 - t2	0.1
	t4 - t0	0.01
	t4 - t2	0.01
	t4 - t3	0.05
Isomaltose	t2 - t1	0.05
Isopropyl alcohol	t3 - t2	0.05
	t4 - t3	0.1
L-Carnitine	t3 - t0	0.1
L-Methionine	t2 - t0	0.1
L-Phenylalanine	t2 - t0	0.05
	t4 - t0	0.1
L-Proline	t4 - t0	0.05
	t4 - t1	0.01
	t4 - t2	0.01
	t4 - t3	0.05
Phenylacetaldehyde	t3 - t0	0.05
	t3 - t1	0.05
	t4 - t3	0.05
Phosphoric acid	t4 - t0	0.05
	t4 - t1	0.1
	t4 - t2	0.05
Pyridine	t1 - t0	0.01
	t2 - t0	0.05
	t4 - t1	0.01
Sarcosine	t1 - t0	0.1
Stearic acid	t1 - t0	0.01
	t4 - t0	0.1
Trigonelline	t4 - t1	0.1
Uric acid	t1 - t0	0.1

Abbreviations: t0, before vaccination; t1, 24-72h after the first vaccine dose; t2, before the second vaccine dose; t3, 24-72h after the second vaccine dose; t4, 30 days after the last vaccine dose.

While the number of identified metabolites was similar to what was found in LMMs, the observed increase in the amount of significant results could stem from the simpler estimation approach that the GEE provides. In addition to requiring less data assumptions, a great advantage of the GEE is its amenability to misspecifications of the “working” correlation structure while maintaining the consistency of the parameter estimates. Evidently, the efficiency of those estimates increases when the correlation structure is correctly specified, particularly when dealing with smaller sample sizes (as in the case of this study) [30].

Diagnostic plots of the significant metabolites showed a linear pattern of constant variance for the residuals of each model (examples shown in Figures 3.11 to 3.14), therefore confirming the assumption of data homoscedasticity. Additional investigation of these 30 models discovered that 17 of them were statistically different from the null model (Figure 3.15), highlighting them as the most significant and interesting results to explore.



Figures 3.11 to 3.14. Diagnostic plots of the GEE residuals of four metabolites against the fitted values of their respective GEE models. In addition to the horizontal marker on the residuals with zero value, each plot has a second black line that is used to represent the general pattern in which the residuals are distributed.

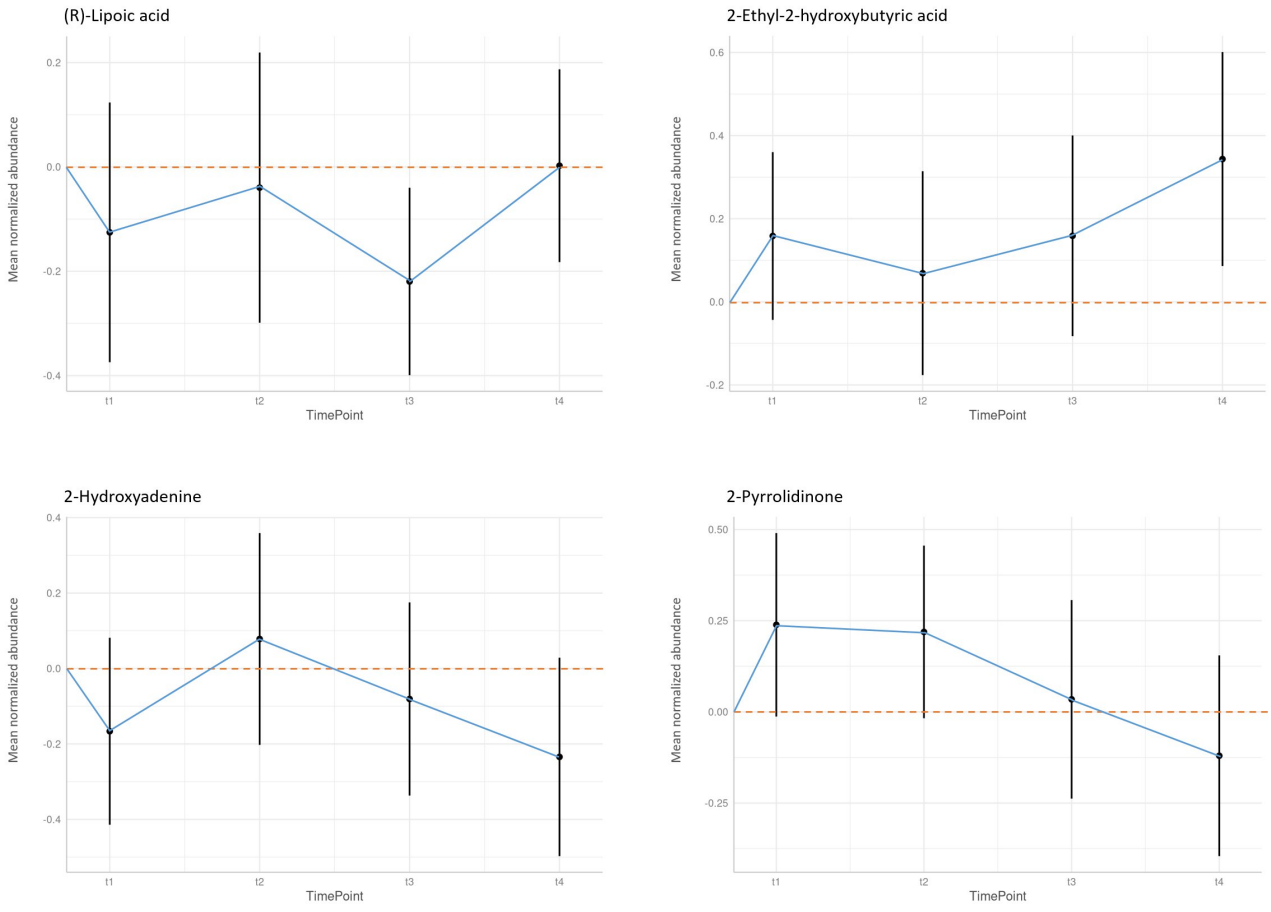
```
##    metabolite    Chi    df     Pr(>Chi)  p.value adj confidence level
## 17         D8 34.72321   3 1.393826e-07 4.181479e-06             0.01
## 23         F5 30.75550   3 9.570162e-07 1.435524e-05             0.01
## 1          A1 16.48554   3 9.015375e-04 6.844357e-03             0.01
## 5          B3 15.98744   3 1.140726e-03 6.844357e-03             0.01
## 8          C3 16.41669   3 9.313638e-04 6.844357e-03             0.01
## 10         C7 15.27222   3 1.598196e-03 7.990979e-03             0.01
```

Figure 3.15. ANOVA results of the top ranked metabolites with the lowest FDR-adjusted p-values obtained for the comparison between the GEE models and their null counterparts. The first four metabolites on the list correspond to hypoxanthine (D8), L-proline (F5), (R)-lipoic acid (A1), and 5'-methylthioadenosine (B3).

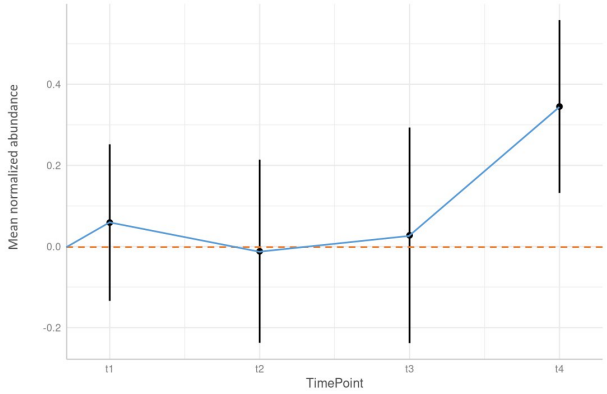
Overall, the GEE strategy seemed to be a better fit for modeling the longitudinal metabolomic data in comparison to LMM. This outcome may be, at least partially, due to the small sample, which hinders the use of models that require normally distributed data as they are more easily achieved with larger sample sizes. Moreover, the utilization of population means for coefficient estimation in GEE allowed to bypass the individual data scattering observed in the LMMs and obtain a more coherent collection of data points, with less prediction error, that focused on the main objective of finding the consensus of metabolite responses over time.

3.4. Behavioral and biological connections

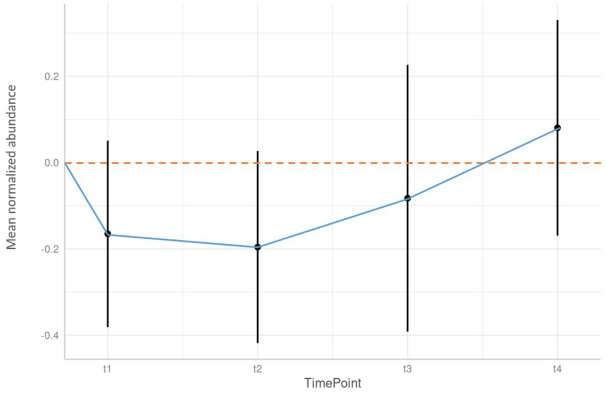
After their validation, the different GEE models (Figures 3.16 to 3.45) were analyzed in more detail to delve into the behavioral resemblances and disparities between the selected metabolites.



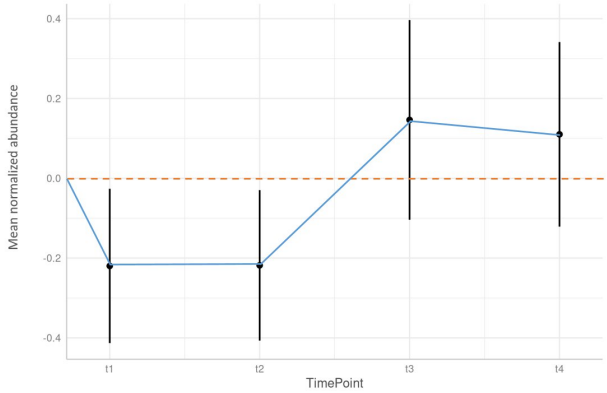
5'-Methylthioadenosine (MTA)



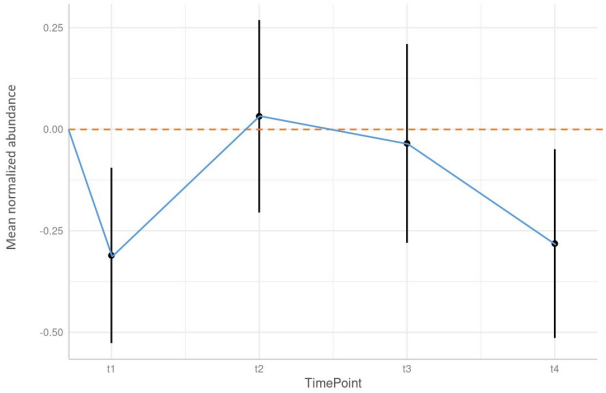
5-Methylcytidine



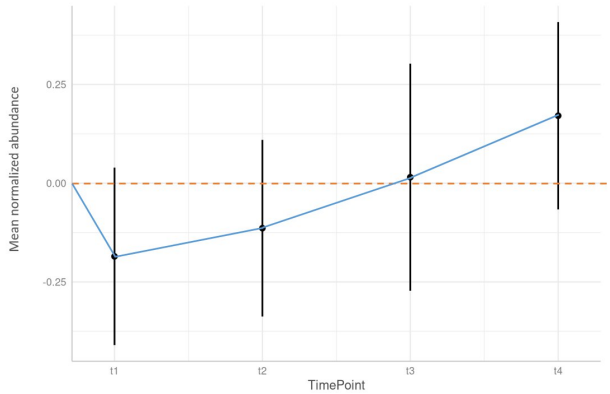
Adenosine 3',5'-diphosphate (PAP)



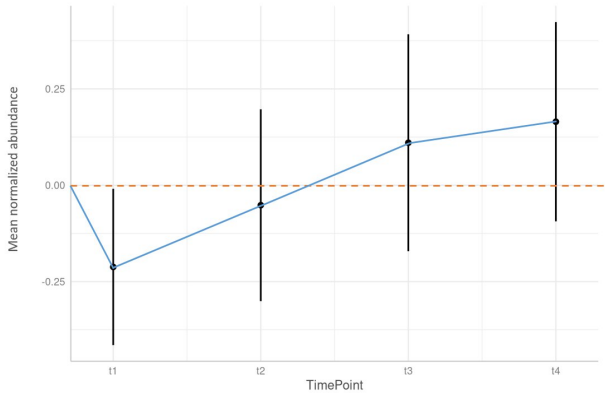
Adenosine monophosphate (AMP)



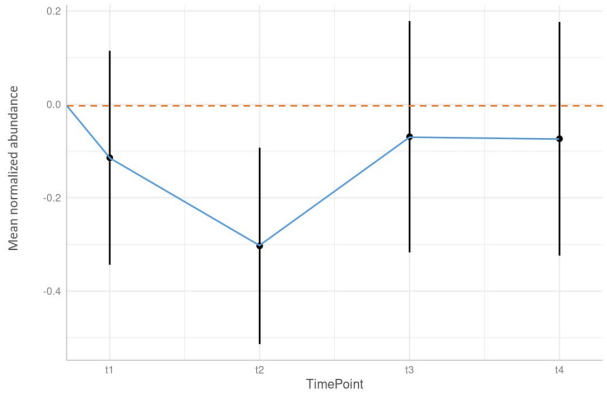
Creatine



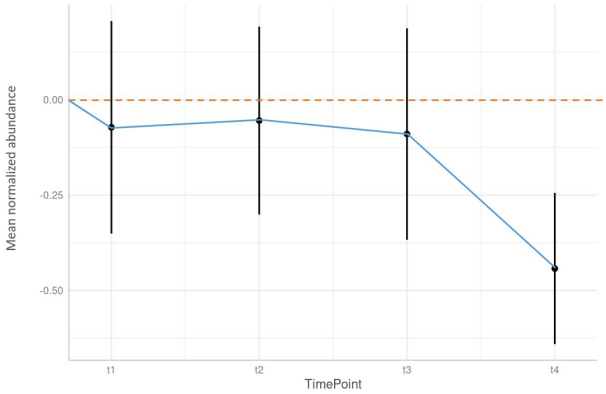
Creatinine



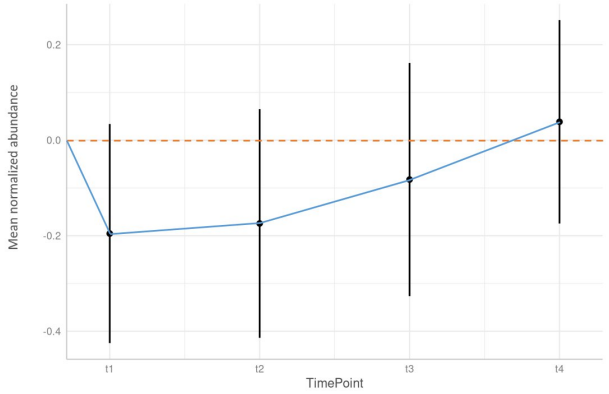
D-Arginine



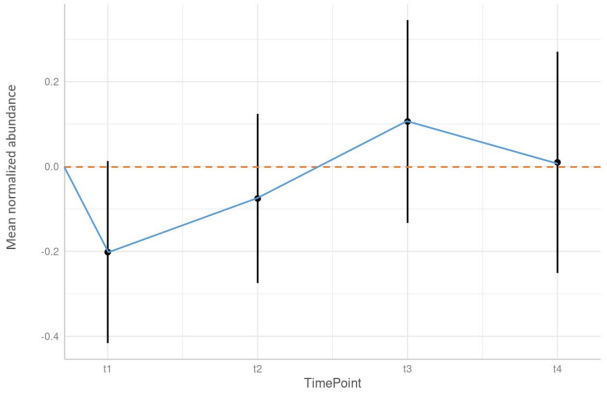
Glucaric acid



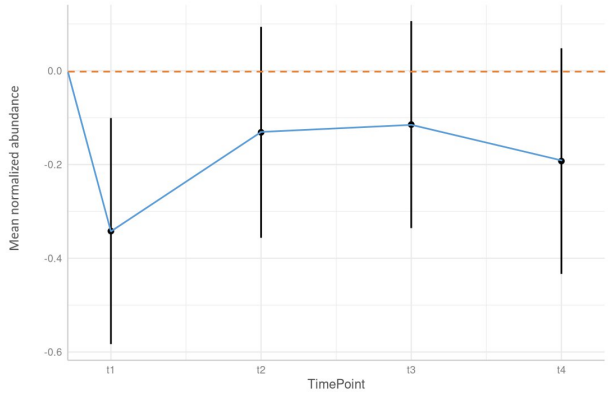
Glycine



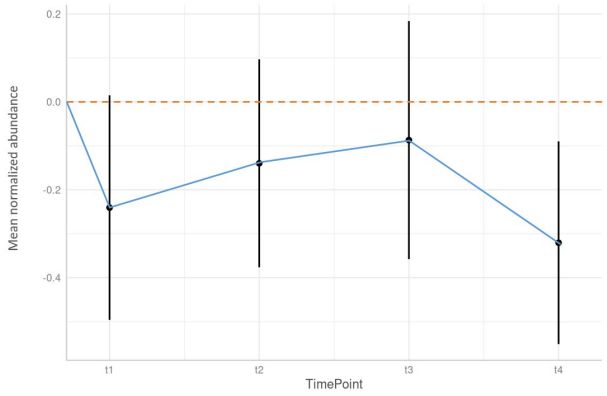
Guanosine 5'-diphosphate (GDP)



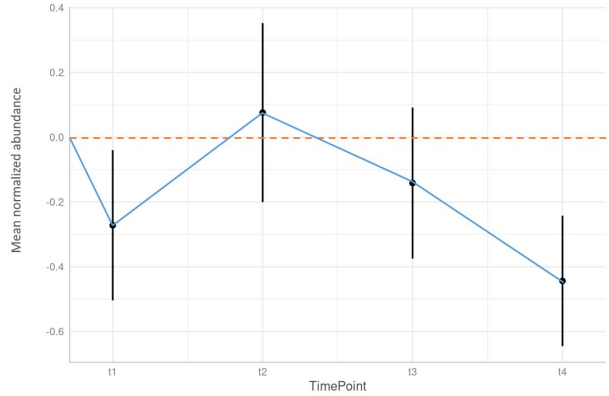
Guanosine monophosphate (GMP)



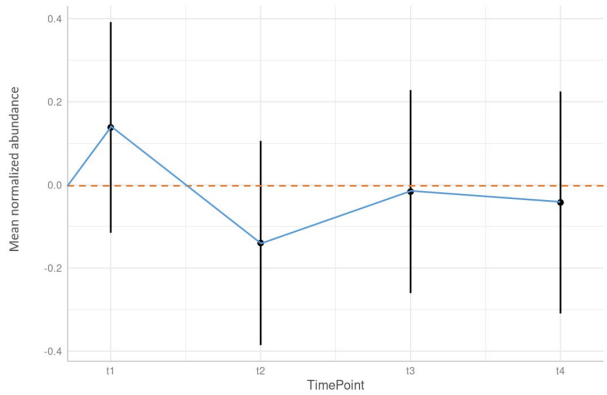
Hyodeoxycholic acid



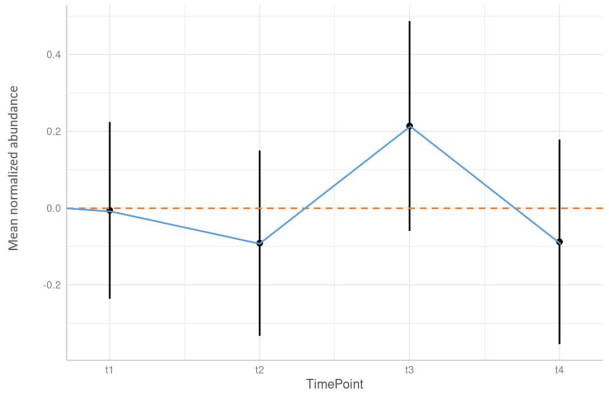
Hypoxanthine



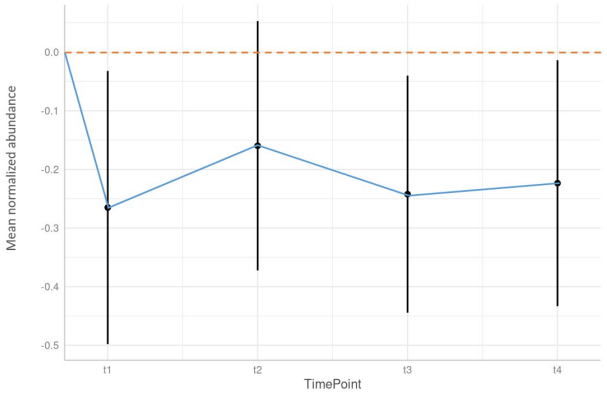
Isomaltose



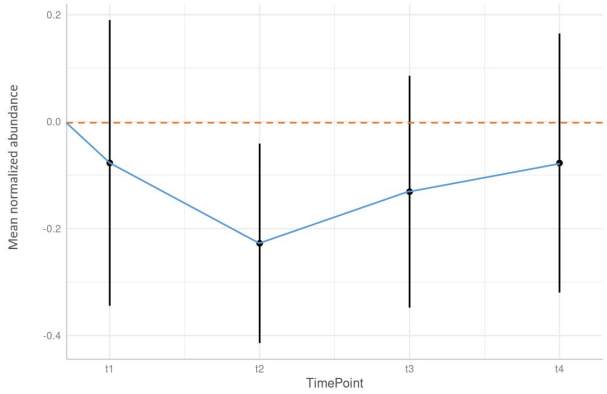
Isopropyl alcohol



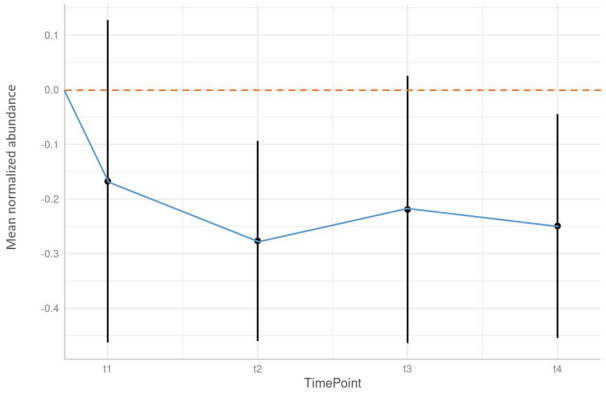
L-Carnitine



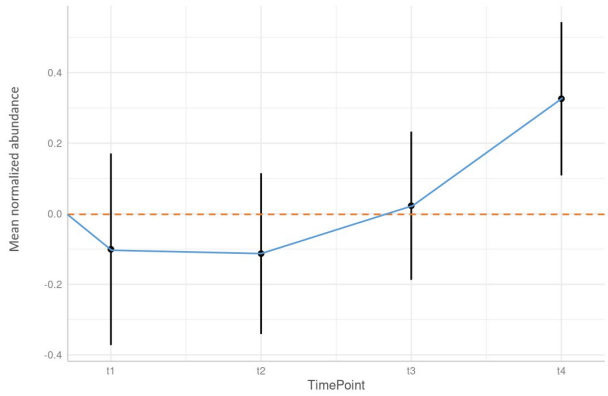
L-Methionine



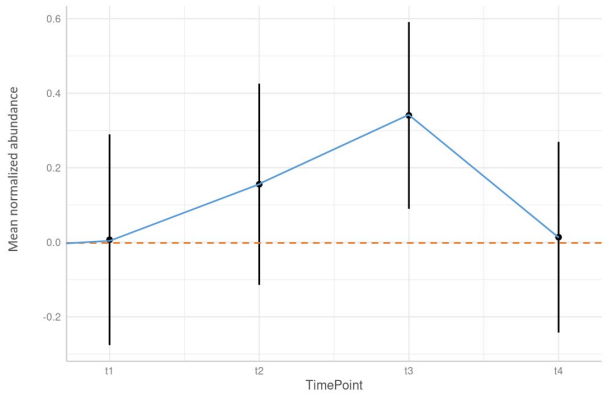
L-Phenylalanine



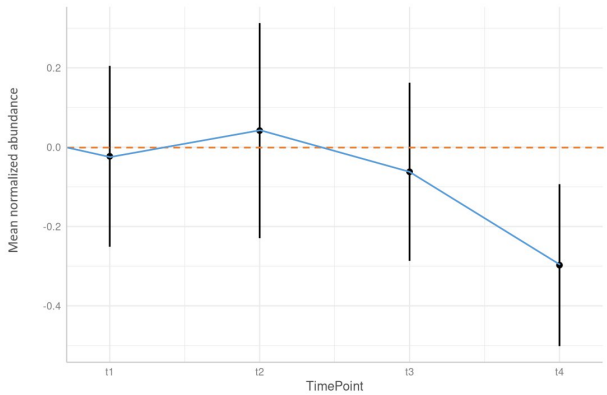
L-Proline



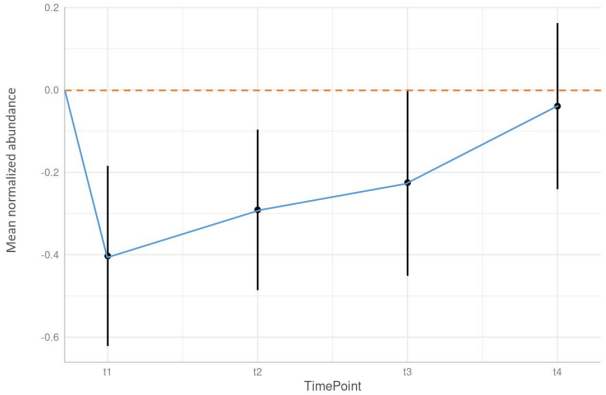
Phenylacetaldehyde



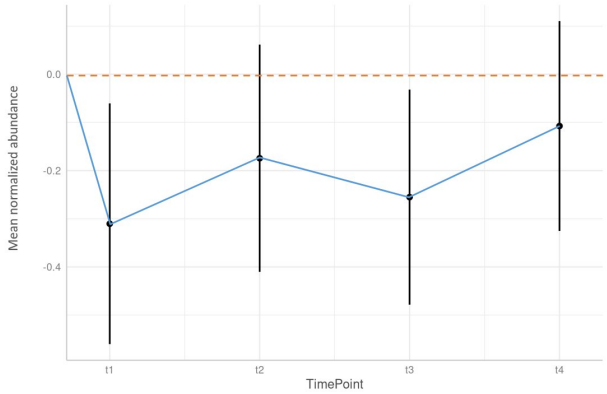
Phosphoric acid



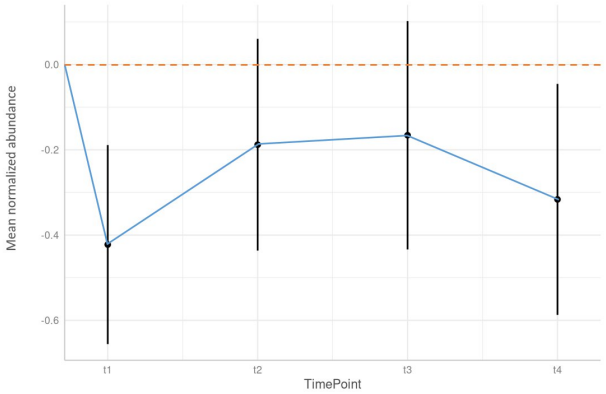
Pyridine

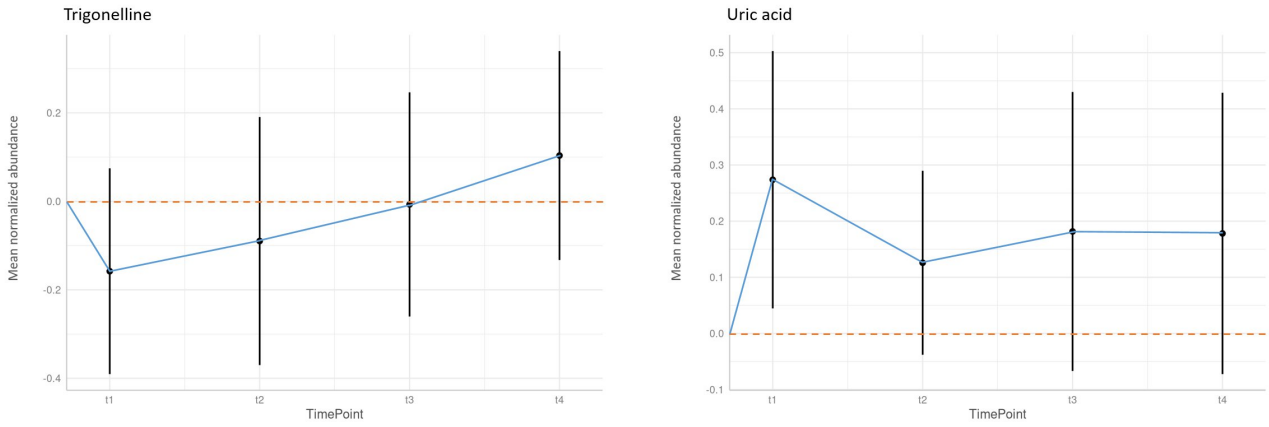


Sarcosine



Stearic acid





Figures 3.16 to 3.45. Plots of the GEE models constructed from the normalized abundance changes observed in 30 different metabolites over time during COVID-19 vaccination. Each of these metabolites was found to have at least one statistically significant GEE coefficient (with FDR-adjusted p-values < 0.1). Every plot is comprised by four black dots and corresponding vertical lines, which represent the intercept coefficients for the different time points and the confidence intervals obtained for their estimates. Also present in each plot, the blue line is used to depict the transition between the concentrations in each time point, while the dotted orange line highlights the baseline value of the initial metabolite abundances.

Most of the significant alterations recorded for the differentially abundant metabolites were observed at the t1 and t4 time points, corresponding to the effects of each vaccine dose after 24-72 hours and 30 days, respectively. This pattern of response indicates that there are possible correlations between the metabolic changes registered and the COVID-19 vaccination. More evidence, however, is required in order to establish any sort of causation relationship concerning these events.

The contribution of each significant metabolite to the development of immunity can have multiple sources, such as being a direct consequence of changes in RBC activity, or a byproduct of cell modifications whose metabolic products are released into the bloodstream and captured by red blood cells as a transportation mechanism to reach other parts of the body. Some of these metabolites might also have intermediate or secondary roles in the pathways of immune response, thus not being directly responsible for the body's reaction to the vaccine.

To explore which metabolic pathways were more associated with the RBC metabolites found to have significant abundance variations during COVID-19 vaccination, pathway and enrichment analyzes were performed through MetaboAnalyst. Within this software, the pathway analysis is a module that integrates enrichment analysis and pathway topology analysis, whereas the dedicated enrichment analysis is used to achieve an over representation analysis (ORA) of the metabolites [57,58].

Pathway topology analysis evaluates the topological roles of metabolites and provide insights into how changes in specific metabolites can affect the dynamics of metabolic pathways, highlighting critical nodes that may be driving the observed effects. Through the construction of pathway network models, this method examines the structural roles and interactions of metabolites within these pathways to understand their biological significance. Hence, the pathway topology analysis can simulate metabolic fluxes based on the presence or absence of metabolites in a given data set, relying on the topological properties and structure of the pathway network to allow for the identification of pathways with altered fluxes and metabolic bottlenecks. It is, for that reason, a method that focuses on the pathway connectivity within a network to pinpoint the key metabolites and metabolic pathways that are most significantly affected by a particular condition or treatment [59-63].

On the contrary, enrichment analysis aims to identify over-represented or enriched metabolic pathways in a given data set compared to a reference database of known metabolite-pathway relationships, instead of focusing on the interactions between metabolites within a network of metabolic pathways. This analysis is often applied in metabolomics to identify metabolic pathways or biological processes (associated with a specific set of metabolites) that are significantly altered in a particular condition or disease state. By identifying significantly enriched terms, researchers can pinpoint the potential regulatory mechanisms that are behind specific phenotypes or diseases, which in turn can help guide them to prioritize research directions that focus on promising areas for further investigation, such as the development of new therapeutic strategies. It is important to note, however, that the lack of metabolite abundance data can limit the sensitivity and specificity of the enrichment analysis, therefore generating less informative and reliable results with lower accuracy and an increased risk of false positive findings when the input of the analysis is a simple list of compounds [59,61,62,64-66].

While both pathway topology and enrichment analyzes can provide valuable insights into the biological mechanisms and metabolic profiles that characterize a certain experimental condition, playing a crucial role in the biological interpretation of metabolomics data, the two methods lead to the production of different types of information due to their distinct features. This is apparent by the distinct MetaboAnalyst results obtained from each analysis, which are depicted in Tables 3.2 and 3.3.

Table 3.2. Pathway analysis conducted through MetaboAnalyst 6.0 using as input a list of compound names from the 30 metabolites found to have significant GEE model coefficients.

Pathway name	Number of hits	Matched compounds	P-value	FDR
Purine metabolism	5/70	GDP; AMP; GMP; Hypoxanthine; Uric acid	0.0013836	0.11069
Phenylalanine metabolism	2/8	Phenylacetaldehyde; L-Phenylalanine	0.0040995	0.16398
Glycine, serine and threonine metabolism	3/33	Glycine; Sarcosine; Creatine	0.0074972	0.19992
Lipoic acid metabolism	2/28	(R)-Lipoate; Glycine	0.047542	0.7981
Phenylalanine, tyrosine and tryptophan biosynthesis	1/4	L-Phenylalanine	0.049881	0.7981
Cysteine and methionine metabolism	2/33	MTA; L-Methionine	0.063949	0.85256
Arginine and proline metabolism	2/36	Creatine; L-Proline	0.074599	0.85256
Ascorbate and aldarate metabolism	1/9	D-Glucarate	0.10891	1.0
D-Amino acid metabolism	1/15	D-Arginine	0.17515	1.0
Starch and sucrose metabolism	1/18	Isomaltose	0.20649	1.0
Pantothenate and CoA biosynthesis	1/20	PAP	0.22675	1.0
Glutathione metabolism	1/28	Glycine	0.30299	1.0
Lysine degradation	1/30	Carnitine	0.3209	1.0
Porphyrin metabolism	1/31	Glycine	0.32969	1.0
Glyoxylate and dicarboxylate metabolism	1/32	Glycine	0.33838	1.0
Biosynthesis of unsaturated fatty acids	1/36	Octadecanoic acid	0.37205	1.0
Primary bile acid biosynthesis	1/46	Glycine	0.44926	1.0

Abbreviations: AMP, Adenosine monophosphate; GDP, Guanosine diphosphate; GMP, Guanosine monophosphate; MTA, 5-methylthioadenosine; PAP, Adenosine 3',5'-diphosphate.

Presented in Table 3.2, the outcome of the pathway analysis shows that 20 out of the 30 metabolites included in the search were matched to one or more metabolic pathways. Most of the featured metabolic pathways are linked to some of the major metabolic pathways of cells, namely the biosynthesis and/or degradation of several amino acids, as well as the metabolism of purines, lipids, and carbohydrates. However, what is also observed from this analysis is that only 7 of these pathways have an associated p-value lower than 0.1, and none of them were found to have significant FDR values. These results imply that the associations established between the metabolites and pathways are likely due to chance, and that there is no sufficient evidence to affirm that these metabolites have a significant impact within the structural context of the identified metabolic pathways.

Table 3.3. Enrichment analysis conducted through MetaboAnalyst 6.0 using as input a list of compound names from the 30 metabolites found to have significant GEE model coefficients.

Metabolite set	Number of hits	Matched compounds	P-value	FDR
SLC-mediated transmembrane transport	12/155	L-Carnitine; Hypoxanthine; Phenylalanine; Proline; Uric acid; Creatinine; Methionine; (R)-Lipoic acid; Stearic acid; Phosphate; Glycine; GMP	3.26E-13	5.54E-10
Transport of small molecules	13/208	L-Carnitine; Hypoxanthine; Phenylalanine; Proline; Uric acid; Creatinine; Methionine; (R)-Lipoic acid; Stearic acid; Phosphate; Glycine; GMP; GDP	3.34E-13	5.54E-10
Metabolism of amino acids and derivatives	13/283	L-Carnitine; Creatine; Phenylalanine; Proline; Sarcosine; Creatinine; Methionine; MTA; Phosphate; AMP; Glycine; GDP; PAP	1.8E-11	1.99E-8
SLC transporter disorders	8/80	L-Carnitine; Phenylalanine; Proline; Uric acid; Methionine; Stearic acid; Phosphate; Glycine	1.1E-9	9.09E-7
Purine metabolism and related disorders	7/58	Hypoxanthine; Uric acid; Methionine; Phosphate; AMP; GMP; GDP	3.91E-9	2.59E-6
Metabolism	19/1370	L-Carnitine; Creatine; Hypoxanthine; Phenylalanine; Proline; Sarcosine; Uric acid; Creatinine; Methionine; Pyridine; MTA; GDP; Phenylacetaldehyde; Stearic acid; PAP; Phosphate; AMP; Glycine; GMP	5.02E-9	2.69E-6
Disorders of transmembrane transporters	8/98	L-Carnitine; Phenylalanine; Proline; Uric acid; Methionine; Stearic acid; Phosphate; Glycine	5.68E-9	2.69E-6
Transport of bile salts and organic acids, metal ions and amine compounds	7/76	L-Carnitine; Phenylalanine; Proline; Uric acid; Creatinine; Methionine; Glycine	2.71E-8	1.12E-5
Methionine metabolism leading to sulfur amino acids and related disorders	5/23	AMP; Glycine; Sarcosine; Methionine; Phosphate	4.14E-8	1.17E-5
Translation	7/83	Phenylalanine; Proline; Methionine; Phosphate; AMP; Glycine; GDP	5.05E-8	1.17E-5
3-Phosphoglycerate dehydrogenase deficiency	6/49	AMP; Creatine; Glycine; Sarcosine; Methionine; (R)-Lipoic acid	5.64E-8	1.17E-5
Dihydropyrimidine dehydrogenase deficiency (DHPD)	6/49	AMP; Creatine; Glycine; Sarcosine; Methionine; (R)-Lipoic acid	5.64E-8	1.17E-5
Dimethylglycine dehydrogenase deficiency	6/49	AMP; Creatine; Glycine; Sarcosine; Methionine; (R)-Lipoic acid	5.64E-8	1.17E-5
Non-ketotic hyperglycinemia	6/49	AMP; Creatine; Glycine; Sarcosine; Methionine; (R)-Lipoic acid	5.64E-8	1.17E-5
Sarcosinemia	6/49	AMP; Creatine; Glycine; Sarcosine; Methionine; (R)-Lipoic acid	5.64E-8	1.17E-5
Transcription/translation	5/25	AMP; Phenylalanine; Proline; Methionine; GMP	6.5E-8	1.27E-5
Biochemical pathways: part I	12/445	AMP; PAP Creatine; Glycine; Hypoxanthine; Phenylalanine; Proline; Sarcosine; Uric acid; Methionine; GDP; GMP	7.99E-8	1.47E-5
Glycine, serine and threonine metabolism	6/56	AMP; Creatine; Glycine; Sarcosine; Methionine; (R)-Lipoic acid	1.28E-7	2.24E-5

Disease	12/470	L-Carnitine; Hypoxanthine; Phenylalanine; Proline; Uric acid; Methionine; GDP; Phosphate; Stearic acid; AMP; Glycine; GMP	1.48E-7	2.45E-5
Metabolism of proteins	10/295	Phenylalanine; Proline; Methionine; MTA; Stearic acid; Phosphate; AMP; Glycine; GDP; PAP	2.07E-7	3.13E-5
tRNA aminoacylation	6/66	Phenylalanine; Proline; Methionine; Phosphate; AMP; Glycine	3.49E-7	3.13E-5
Adenine phosphoribosyltransferase deficiency (APRT)	6/66	AMP; Glycine; Hypoxanthine; Uric acid; GDP; GMP	3.49E-7	3.13E-5

Abbreviations: AMP, Adenosine monophosphate; GDP, Guanosine diphosphate; GMP, Guanosine monophosphate; MTA, 5-methylthioadenosine; PAP, Adenosine 3',5'-diphosphate; SLC, Solute carrier.

The results of the enrichment analysis shown in Table 3.3 seem to be highly significant all-around, even after the p-value correction for multiple testing. These low FDR values indicate that the observed enrichments are statistically significant and, thus, are unlikely to be due to chance. The enriched metabolite sets appear to be a mixture of specific disorder and deficiency-related pathways, varying categories of biomolecule metabolism, and broad biological processes. Collectively, these pathways give a general idea of the potential functions of these metabolites, which are mainly tied to the metabolism of amino acids, the purine salvage pathway, and the transport of small molecules through the cellular membrane, as well as some associated disorders.

Recent research on the erythrocyte metabolism supports these discoveries, with some of the RBC metabolites being explained as intrinsic to the good functioning of such cells. One example of that is during metabolic degradation, where purine bases (like hypoxanthine) and nucleosides are transported or “salvaged” from the circulation into the erythrocyte and re-utilized in a series of reactions to generate purine nucleotides (such as AMP and GMP). These molecules can be subsequently converted to ADP and GDP, respectively, which is then used in glycolysis to produce ATP. Because of the important role that these enzymatic reactions play in the glycolytic flux regulation of erythrocytes, it has been discovered that changes in cellular energy charge can be detected through the level of AMP abundance in RBCs. The concentration of AMP in erythrocytes is, therefore, a measure of glycolytic alterations in these cells, emphasizing it as a potential biomarker for infections or other conditions where these metabolic dysregulations often occur [9].

Amino acids such as methionine, arginine, and carnitine have also been well described in RBCs. The methionine salvage pathway, for instance, plays an essential role in the synthesis of polyamines and the repair of protein damage, whereas arginine is utilized for nitric oxide production, a critical gaseous signaling molecule that is relevant in many redox reactions. Redox homeostasis is particularly important for the preservation of the erythrocyte membrane, as well as its cytoskeletal composition and properties. Since carnitine metabolism is deeply involved in the repair of oxidized membrane lipids, its dysregulation results in an excessive peroxidation of the RBC membrane that can lead to decreased deformability and eryptosis (cell death). When occurring at the systemic level, this extreme oxidative damage can trigger inflammatory responses that cause membrane fluidity reduction and impaired microcirculation, both of which have been observed in various diseases, including COVID-19 [9].

In addition to the metabolism of amino and nucleic acids, the transmembrane transport of solutes is also a vital mechanism of cells that allows them to trade molecules between the intracellular and extracellular spaces, which is believed to be one of the main functions of RBCs besides oxygen loading, transport, and delivery. It is thanks to the pumps, transporters, and channels contained within the RBC membrane that the erythrocyte can control and maintain its shape, volume, osmotic balance, and ultimately, its structural integrity and deformability [9,67].

Finally, the connection of the RBC metabolites to several disorders and deficiencies is a possible indication that they are simultaneously encountered more often in cases of cell stress, tissue damage or even diseases and infections. Such situations are usually related to an increase in the activity of the immune system, something that is also expected to occur during the COVID-19 vaccination process, from which the experimental results were obtained. Everything considered, this enrichment analysis points to a probable link between RBC metabolites and the development of immune responses.

As for the differences between the outcome of the pathway and enrichment analyzes, it presents an interesting matter to discuss. Considering that both methods include an enrichment step, which produced significant ORA results, it is plausible that the non-significant outcomes of the first analysis are related to the use of a topology measure to determine the structure and connectivity of the different pathway networks. Given the small number of metabolites identified within each metabolic pathway, it makes sense that their contribution to modifications in those pathways would not be sufficient to generate meaningful changes in the metabolic activity of the network.

Following the study of metabolic pathways, a behavioral analysis was conducted from the model coefficients of the 30 significant metabolites that resulted in the establishment of seven different groups of metabolites with analogous response trajectories (Table 3.4), emphasizing them as potential biological associations within the RBC metabolism.

Table 3.4. Behavioral analysis of metabolites with similar response trajectories to COVID-19 vaccination. The symbols used to express variation indicate that, from one time point to another, a decrease (↘) or increase (↗) in abundance occurred. Each row comprises a group of metabolites that have a unique trajectory of response. Groups of metabolites with opposite responses are highlighted with the same colors. Non-highlighted groups do not share any similarities.

Similar responses	t1	t2	t3	t4
(R)-Lipoic acid; L-Carnitine; Sarcosine	↘	↗	↘	↗
Isomaltose; Uric acid	↗	↘	↗	↘
2-Hydroxyadenine; AMP; Glucaric acid; Hypoxanthine; Phosphoric acid	↘	↗	↘	↘
2-Ethyl-2-hydroxybutyric acid; MTA	↗	↘	↗	↗
5-Methylcytidine; L-Methionine; L-Proline	↘	↘	↗	↗
Creatine; Creatinine; Glycine; Pyridine; Trigonelline	↘	↗	↗	↗
2-Pyrrolidinone	↗	↘	↘	↘
D-Arginine; Isopropyl alcohol; L-Phenylalanine	↘	↘	↗	↘
PAP; GDP; GMP; Hyodeoxycholic acid; Stearic acid	↘	↗	↗	↘
Phenylacetaldehyde	↗	↗	↗	↘

Abbreviations: AMP, Adenosine monophosphate; GDP, Guanosine diphosphate; GMP, Guanosine monophosphate; MTA, 5-methylthioadenosine; PAP, Adenosine 3',5'-diphosphate; t1, 24-72h after the first vaccine dose; t2, before the second vaccine dose; t3, 24-72h after the second vaccine dose; t4, 30 days after the last vaccine dose.

Depicted in Table 3.4 are several correlations that were detected between the responses of different metabolites, with up to seven metabolites in a single group that share an identical or opposite trajectory of abundance across time points. In contrast, phenylacetaldehyde did not receive any pairing, hence being the only statistically significant metabolite with a unique response to the vaccination process. When looking at their magnitudes of response, the closest metabolite trajectories were found in the variations of creatine, trigonelline, and creatinine, followed in second place by L-carnitine and sarcosine with proximate results as well.

Mainly used to regenerate ATP within the cell, creatine is a naturally occurring compound that is believed to have anti-inflammatory effects by reducing the production of pro-inflammatory cytokines. Its abundance increase during immune responses, such as the one observed in this work, is thought to be associated with a rise in ATP synthesis in order to boost leukocyte proliferation and activity, reason why creatine is labeled in some studies as a predictor of greater disease severity in COVID-19 [68-71].

Creatinine, a breakdown product of creatine, is considered a waste product of cellular metabolism that is normally eliminated by the kidneys through urinary excretion. Because of that, high levels of creatinine are suggestive of kidney injury or failure, which may arise during sepsis, serious kidney infections or certain viral infections. This greater abundance of creatinine was accompanied by elevated levels of neutrophils and leukocytes in severe and fatal COVID-19 cases, hence serving as a potential biomarker for the disease [72-75].

Trigonelline is a product of niacin (vitamin B3) metabolism that is excreted in the urine. Its higher abundance levels in COVID-19 have been associated with increased activation of the NAD+ salvage pathway as an effort to improve NAD+ recycling, which is fundamental for redox reactions. Through this action, trigonelline is responsible for reducing oxidative stress, while also mitigating inflammatory responses by attenuating the expression of pro-inflammatory cytokines [76-79].

As mentioned earlier, carnitine is responsible for repairing oxidatively damaged lipids to enable restoration and maintenance of cellular membrane integrity, thus protecting cells from ROS-induced damage. This amino acid has been demonstrated to reduce inflammation and oxidative stress in COVID-19 patients by preventing the accumulation of lipid peroxidation end-products, as well as being able to improve oxygen delivery via increase of oxygen saturation levels. Consequently, higher levels of carnitine have been linked to decreased disease severity and increased oxygen saturation leading to a faster recovery [9,80-84].

Sarcosine is a natural amino acid that is mostly involved in the physiological recovery from viral infections and immune activities. It is responsible for boosting antigen-presenting cell activity and autophagy, which is essential for the antiviral response mediated through the direct removal of virus, the display of viral antigens, and the suppression of excessive inflammation [85-87].

These are some examples of metabolites that have relevant functions in the development of an effective immune response against infections, such as COVID-19. Their purposes revolve around enhancing certain metabolic pathways to boost lymphocyte production and reducing oxidative stress, while at the same time decreasing inflammation to counteract the cytokine storm (excessive release of pro-inflammatory signaling molecules) that is triggered by the innate immune system in response to an infection.

A look at the behavior of these metabolites over time shows that the actions which promote the body's immunization are coherent with the abundance changes recorded for the aforementioned metabolites, where an initial plummet is observed after the first (and sometimes second) dose is administered, likely due to the inflammatory reactions that are triggered by the vaccine (similar to what occurs in the acute phase of a disease). This is followed by a rise in concentration during the recovery periods of each dose, where the immune system attempts to reinstate homeostasis through an anti-inflammatory and anti-oxidant approach, alongside the development of more permanent defenses against future invading pathogens. However, it is important to clarify that the abundance declines of these metabolites were, for the most part, more severe than its subsequent increments, meaning that their concentration values remained negative for the majority of the immunization process (in comparison to the baseline measured at the start of the experiment, before vaccination). Still, the observed trends of metabolite abundance

variation are in agreement with the mechanisms of immune response, solidifying the conviction on the potential immunological roles of these metabolites.

Altogether, the connections found between several RBC metabolites and the defense mechanisms commonly used by cells to fight viral infections offer a strong clue for the immune functions that those metabolites may have in the human body, suggesting that the RBC metabolism could play a significant part in the generation of an immune response to COVID-19 vaccination.

4. CONCLUSIONS AND FUTURE PERSPECTIVES

Red blood cells have emerged in the last years as important modulators of the immune system. Despite existing evidence of alterations in RBC functionality being associated with the COVID-19 disease, there was still no concrete information regarding the impact of RBC activity on the immune response to COVID-19 vaccination.

The results of the statistical analysis conducted in this work have demonstrated that several RBC metabolites suffer significant changes in abundance during the immunization process triggered by the COVID-19 vaccine. These metabolites appear to be mainly linked to the metabolism of amino acids and purines, as well as the transport of small molecules through the cellular membrane. Some of the highlighted metabolites were found to have important functions related to lymphocyte production, and the reduction of oxidative stress and inflammation, all of which contribute to the development of an effective immune response. Overall, the connections found between RBC metabolites and the defense mechanisms of cells against foreign invaders present a compelling hypothesis for the immune functions that those metabolites may have in the human body, thereby implying that the RBC metabolism may have a predominant role in the immunization mechanisms of COVID-19 vaccination.

Work is in progress to integrate and correlate proteomic data retrieved from the same longitudinal experiment for a comprehensive depiction of the RBC function in the COVID-19 vaccine-induced immunization process. Potential future investigations could be made to evaluate the relevance of other explanatory variables (namely age, gender and BMI) in the metabolic variations that were found to occur during COVID-19 vaccination. In addition, alternative imputation strategies could be tested on the metabolomics data set and compared to the results in this work in order to provide a more complete understanding of the effects of different imputation algorithms on this data.

5. REFERENCES

[1] Sánchez, A., García-Pardo, G., Gómez-Bertomeu, F., López-Dupla, M., Foguet-Romero, E., Buzón, M. J., Almirante, B., Olona, M., Fernández-Veledo, S., Vidal, F., Chafino, S., Rull, A., & Peraire, J. (2023). Mitochondrial dysfunction, lipids metabolism, and amino acid biosynthesis are key pathways for COVID-19 recovery. *iScience*, *26*(10), 107948. https://doi.org/10.1016/j.isci.2023.107948

[2] Mendonça, M. M., Da Cruz, K. R., Da Silva Pinheiro, D., Moraes, G. C. A., Ferreira, P. M., Ferreira-Neto, M. L., Da Silva, E. S., Gonçalves, R. V., Pedrino, G. R., Fajemiroye, J. O., & Xavier, C. H. (2022). Dysregulation in erythrocyte dynamics caused by SARS-CoV-2 infection: possible role in shuffling the homeostatic puzzle during COVID-19. *Hematology Transfusion and Cell Therapy*, *44*(2), 235–245. https://doi.org/10.1016/j.htct.2022.01.005

[3] De Souza Nogueira, J., Santos-Rebouças, C. B., Piergiorge, R. M., Valente, A. P., Gama-Almeida, M. C., El-Bacha, T., Moreira, M. L. L., Marques, B. S. B., De Siqueira, J. R., De Carvalho, E. M., Da Costa Ferreira, O., Porto, L. C., Da Silva Fidalgo, T. K., & Santos, G. C. D. (2023). Metabolic Adaptations Correlated with Antibody Response after Immunization with Inactivated SARS-CoV-2 in Brazilian Subjects. *Journal of Proteome Research*, *22*(6), 1908–1922. https://doi.org/10.1021/acs.jproteome.3c00014

[4] Diray-Arce, J., Conti, M. G., Petrova, B., Kanarek, N., Angelidou, A., & Levy, O. (2020). Integrative metabolomics to identify molecular signatures of responses to vaccines and infections. *Metabolites*, *10*(12), 492. https://doi.org/10.3390/metabo10120492

[5] Haj, A. K., Hasan, H., & Raife, T. J. (2022). Heritability of Protein and Metabolite Biomarkers Associated with COVID-19 Severity: A Metabolomics and Proteomics Analysis. *Biomolecules*, *13*(1), 46. https://doi.org/10.3390/biom13010046

[6] Wang, Y., Wang, X., Luu, L. D. W., Chen, S., Jin, F., Wang, S., Huang, X., Wang, L., Zhou, X., Chen, X., Cui, X., Li, J., Tai, J., & Zhu, X. (2022). Proteomic and metabolomic signatures associated with the immune response in healthy individuals immunized with an inactivated SARS-COV-2 vaccine. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.848961

[7] Dagla, I., Iliou, A., Benaki, D., Gikas, E., Mikros, E., Bagratuni, T., Kastritis, E., Dimopoulos, M. A., Terpos, E., & Tsarbopoulos, A. (2022). Plasma metabolomic alterations induced by COVID-19 vaccination reveal putative biomarkers reflecting the immune response. *Cells*, *11*(7), 1241. https://doi.org/10.3390/cells11071241

[8] He, M., Huang, Y., Wang, Y., Liu, J., Han, M., Xiao, Y., Zhang, N., Gui, H., Qiu, H., Cao, L., Jia, W., & Huang, S. (2022). Metabolomics-based investigation of SARS-CoV-2 vaccination (Sinovac) reveals an immune-dependent metabolite biomarker. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.954801

[9] Chatzinikolaou, P. N., Margaritelis, N. V., Paschalis, V., Theodorou, A. A., Vrabas, I. S., Kyparos, A., D'Alessandro, A., & Nikolaidis, M. G. (2024). Erythrocyte metabolism. *Acta Physiologica*, *240*(3). https://doi.org/10.1111/apha.14081

[10] Nemkov, T., Reisz, J. A., Xia, Y., Zimring, J. C., & D'Alessandro, A. (2018). Red blood cells as an organ? How deep omics characterization of the most abundant cell in the human body highlights other systemic metabolic functions beyond oxygen transport. *Expert Review of Proteomics*, *15*(11), 855–864. https://doi.org/10.1080/14789450.2018.1531710

[11] Russo, A., Tellone, E., Barreca, D., Ficarra, S., & Laganà, G. (2022). Implication of COVID-19 on erythrocytes functionality: red blood cell biochemical implications and Morpho-Functional Aspects. *International Journal of Molecular Sciences*, *23*(4), 2171. https://doi.org/10.3390/ijms23042171

[12] Thomas, T., Stefanoni, D., Dzieciatkowska, M., Issaian, A., Nemkov, T., Hill, R. C., Francis, R. O., Hudson, K. E., Buehler, P. W., Zimring, J. C., Hod, E. A., Hansen, K. C., Spitalnik, S. L., & D'Alessandro, A. (2020). Evidence of Structural Protein Damage and Membrane Lipid Remodeling in Red Blood Cells from COVID-19 Patients. *Journal of Proteome Research*, *19*(11), 4455–4469. https://doi.org/10.1021/acs.jproteome.0c00606

[13] Sun, J., & Xia, Y. (2023). Pretreating and normalizing metabolomics data for statistical analysis. *Genes & Diseases*, *11*(3), 100979. https://doi.org/10.1016/j.gendis.2023.04.018

[14] Sysi-Aho, M., Katajamaa, M., Yetukuri, L., & Orešič, M. (2007). Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, *8*(1). https://doi.org/10.1186/1471-2105-8-93

[15] Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. (2018). Missing value imputation approach for mass spectrometry-based metabolomics data. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-017-19120-0

[16] Van Den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & Van Der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, *7*(1). https://doi.org/10.1186/1471-2164-7-142

[17] GeeksforGeeks. (2024, May 23). *Normalization and scaling*. GeeksforGeeks. https://www.geeksforgeeks.org/normalization-and-scaling/

[18] GeeksforGeeks. (2024, February 14). *What is Zero Mean and Unit Variance Normalization*. GeeksforGeeks. https://www.geeksforgeeks.org/what-is-zero-mean-and-unit-variance-normalization/

[19] Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd ed.). John Wiley & Sons, Inc.

[20] Mwangi, M., Verbeke, G., Njagi, E. N., Mwalili, S., Ivanova, A., Bukania, Z. N., & Molenberghs, G. (2021). Improved longitudinal data analysis for cross-over design settings, with a piecewise

linear mixed-effects model. *Communications in Statistics Case Studies Data Analysis and Applications*, *7*(3), 413–431. https://doi.org/10.1080/23737484.2021.1959468

[21] Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in Medicine*, *16*(20), 2349-2380. https://doi.org/10.1002/(SICI)1097-0258(19971030)16:20<2349::AID-SIM667>3.0.CO;2-E

[22] Laird, N. M., & Ware, J. H. (1982). Random-Effects Models for longitudinal data. *Biometrics*, *38*(4), 963. https://doi.org/10.2307/2529876

[23] Lopes, V. S. S. (2022). *Self-reported Diabetes in Portugal and its association with health-related quality of life and medical resources consumption using a nationwide prospective cohort* [Master's thesis in Biostatistics]. Faculdade de Ciências da Universidade de Lisboa.

[24] Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Series in Statistics.

[25] *What is the difference between generalized estimating equations and GLMM?* (2011). Cross Validated. Retrieved November 26, 2024, from https://stats.stackexchange.com/questions/17331/what-is-the-difference-between-generalized-estimating-equations-and-glmm

[26] *When to use generalized estimating equations vs. mixed effects models?* (2011). Cross Validated. Retrieved November 26, 2024, from https://stats.stackexchange.com/questions/16390/when-to-use-generalized-estimating-equations-vs-mixed-effects-models

[27] Wikipedia contributors. (2024, October 29). *Generalized linear model*. Wikipedia. https://en.wikipedia.org/wiki/Generalized_linear_model

[28] Grace-Martin, K. (2024, January 2). *Approaches to repeated measures data: Repeated measures ANOVA, marginal, and mixed models*. The Analysis Factor. https://www.theanalysisfactor.com/repeated-measures-approaches/

[29] Ruíz, J. S., López, O. a. M., Ramírez, G. H., & Hiriart, J. C. (2023). Generalized linear mixed models for non-normal responses. In *Springer eBooks* (pp. 113–127). https://doi.org/10.1007/978-3-031-32800-8_4

[30] Wang, M. (2014). Generalized Estimating Equations in Longitudinal Data Analysis: A review and Recent developments. *Advances in Statistics*, *2014*, 1–11. https://doi.org/10.1155/2014/303728

[31] Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

[32] R Core Team (2023). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/

[33] Pang, Z., Lu, Y., Zhou, G., Hui, F., Xu, L., Viau, C., Spigelman, A. F., MacDonald, P. E., Wishart, D. S., Li, S., & Xia, J. (2024). MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Research*, *52*(W1), W398–W406. https://doi.org/10.1093/nar/gkae253

[34] Meyer, D., & Buchta, C. (2022). proxy: Distance and Similarity Measures. R package version 0.4-27. https://CRAN.R-project.org/package=proxy

[35] Lazar, C., & Burger, T. (2022). imputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation. R package version 2.1. https://CRAN.R-project.org/package=imputeLCMD

[36] Moeller, J. (2015). A word on standardization in longitudinal studies: don't. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01389

[37] Debastiani, V. J., & Pillar, V. D. (2012). SYNCSA — R tool for analysis of metacommunities based on functional traits and phylogeny of the community components. *Bioinformatics*, *28*(15), 2067–2068. https://doi.org/10.1093/bioinformatics/bts325

[38] Hamner, B., & Frasco, M. (2018). Metrics: Evaluation Metrics for Machine Learning. R package version 0.1.4. https://CRAN.R-project.org/package=Metrics

[39] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

[40] Lenth, R. (2024). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.10.0. https://CRAN.R-project.org/package=emmeans

[41] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). Springer-Verlag, NY.

[42] Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion

[43] Ford, C. (2021). Getting Started with Generalized Estimating Equations. *University of Virginia Library*. https://library.virginia.edu/data/articles/getting-started-with-generalized-estimating-equations

[44] Højsgaard, S., Halekoh, U., & Yan, J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, *15*(2), 1-11. https://www.jstatsoft.org/article/view/v015i02

[45] Yan, J., & Fine, J. P. (2004). Estimating Equations for Association Structures. *Statistics in Medicine*, *23*(6), 859-880. https://doi.org/10.1002/sim.1650

[46] Yan, J. (2002). geepack: Yet Another Package for Generalized Estimating Equations. *R News*, *2/3*, 12-14.

[47] Lüdecke, D. (2018). ggeffects: Tidy Data Frames of Marginal Effects from Regression Models. *The Journal of Open Source Software*, *3*(26), 772. https://doi.org/10.21105/joss.00772

[48] *Model validation with GEE-GLM*. (2020). Cross Validated. Retrieved November 26, 2024, from https://stats.stackexchange.com/questions/447717/model-validation-with-gee-glm

[49] Vanegas, L., Rondón, L., & Paula, G. (2024). glmtoolbox: Set of Tools to Data Analysis using Generalized Linear Models. R package version 0.1.10. https://CRAN.R-project.org/package=glmtoolbox

[50] Zhang, B., Hu, S., Baskin, E., Patt, A., Siddiqui, J., & Mathé, E. (2018). RAMP: A comprehensive relational database of metabolomics pathways for pathway enrichment analysis of genes and metabolites. *Metabolites*, *8*(1), 16. https://doi.org/10.3390/metabo8010016

[51] Bobbitt, Z. (2021, October 4). *MAE vs. RMSE: Which Metric Should You Use?* Statology. https://www.statology.org/mae-vs-rmse/

[52] Chugh, A. (2024, January 18). MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? *Medium*. https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

[53] Hiregoudar, S. (2022, March 4). Ways to Evaluate regression Models - towards data science. *Medium*. https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70

[54] *Relative Absolute Error*. (n.d.). Statistics How To. Retrieved November 26, 2024, from https://www.statisticshowto.com/relative-absolute-error/

[55] Hyndman, R. J., & Athanasopoulos, G. (2018). 3.1 Some simple forecasting methods. In *Forecasting: Principles and Practice* (2nd ed.). OTexts: Melbourne, Australia. https://otexts.com/fpp2/simple-methods.html. Accessed on November 26, 2024.

[56] Czado, C. (n.d.). *Lecture 10: Linear Mixed Models (Linear Models with Random Effects)* [Slide show]. TU München.

[57] Chong, J., & Xia, J. (2023, July 24). *Pathway analysis*. Retrieved November 26, 2024, from https://www.metaboanalyst.ca/resources/vignettes/Pathway_Analysis.html

[58] Pang, Z., Chong, J., & Xia, J. (2023, July 24). *Enrichment analysis*. Retrieved November 26, 2024, from https://www.metaboanalyst.ca/resources/vignettes/Enrichment_Analysis.html

[59] Chong, J., Wishart, D. S., & Xia, J. (2019). Using MetaboAnalyst 4.0 for comprehensive and integrative metabolomics data analysis. *Current Protocols in Bioinformatics*, *68*(1). https://doi.org/10.1002/cpbi.86

[60] Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichiţa, C., & Drăghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*, *4*. https://doi.org/10.3389/fphys.2013.00278

[61] Ma, J., Shojaie, A., & Michailidis, G. (2019). A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics*, *20*(1). https://doi.org/10.1186/s12859-019-3146-1

[62] Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., & Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics*, *16*(1). https://doi.org/10.1186/s12859-015-0751-5

[63] Ihnatova, I., Popovici, V., & Budinska, E. (2018). A critical comparison of topology-based pathway analysis methods. *PLoS ONE*, *13*(1), e0191154. https://doi.org/10.1371/journal.pone.0191154

[64] Lu, Y., Pang, Z., & Xia, J. (2022). Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data. *Briefings in Bioinformatics*, *24*(1). https://doi.org/10.1093/bib/bbac553

[65] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

[66] Marco-Ramell, A., Palau-Rodriguez, M., Alay, A., Tulipani, S., Urpi-Sarda, M., Sanchez-Pla, A., & Andres-Lacueva, C. (2018). Evaluation and comparison of bioinformatic tools for the enrichment analysis of metabolomics data. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-017-2006-0

[67] Nemkov, T., Reisz, J. A., Xia, Y., Zimring, J. C., & D'Alessandro, A. (2018b). Red blood cells as an organ? How deep omics characterization of the most abundant cell in the human body highlights other systemic metabolic functions beyond oxygen transport. *Expert Review of Proteomics*, *15*(11), 855–864. https://doi.org/10.1080/14789450.2018.1531710

[68] *Human Metabolome Database: Showing metabocard for Creatine (HMDB0000064)*. (n.d.). Retrieved November 26, 2024, from https://hmdb.ca/metabolites/HMDB0000064

[69] Kreider, R. B., & Stout, J. R. (2021). Creatine in health and disease. *Nutrients*, *13*(2), 447. https://doi.org/10.3390/nu13020447

[70] Recktenwald, S. M., Simionato, G., Lopes, M. G., Gamboni, F., Dzieciatkowska, M., Meybohm, P., Zacharowski, K., Von Knethen, A., Wagner, C., Kaestner, L., D'Alessandro, A., & Quint, S. (2022). Cross-talk between red blood cells and plasma influences blood flow and omics phenotypes in severe COVID-19. *eLife*, *11*. https://doi.org/10.7554/elife.81316

[71] Albóniga, O. E., Jiménez, D., Sánchez-Conde, M., Vizcarra, P., Ron, R., Herrera, S., Martínez-Sanz, J., Moreno, E., Moreno, S., Barbas, C., & Serrano-Villar, S. (2022). Metabolic snapshot of plasma samples reveals new pathways implicated in SARS-COV-2 pathogenesis. *Journal of Proteome Research*, *21*(3), 623–634. https://doi.org/10.1021/acs.jproteome.1c00786

[72] *Human Metabolome Database: Showing metabocard for Creatinine (HMDB0000562)*. (n.d.). Retrieved November 26, 2024, from https://hmdb.ca/metabolites/HMDB0000562

[73] Deng, X., Liu, B., Li, J., Zhang, J., Zhao, Y., & Xu, K. (2020). Blood biochemical characteristics of patients with coronavirus disease 2019 (COVID-19): a systemic review and meta-analysis. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *58*(8), 1172–1181. https://doi.org/10.1515/cclm-2020-0338

[74] Ghazanfari, T., Salehi, M. R., Namaki, S., Arabkheradmand, J., Rostamian, A., Chenary, M. R., Ghaffarpour, S., Ardestani, S. K., Edalatifard, M., Naghizadeh, M. M., Mohammadi, S., Mahloujirad, M., Izadi, A., Ghanaati, H., Beigmohammadi, M. T., Vodjgani, M., Shirazi, B. M., Mirsharif, E. S., Abdollahi, A., ... Faghihzadeh, S. (2021). Interpretation of Hematological, Biochemical, and Immunological Findings of COVID-19 Disease: Biomarkers Associated with Severity and Mortality. *Iranian Journal of Allergy Asthma and Immunology*. https://doi.org/10.18502/ijaai.v20i1.5412

[75] Melo, A. K. G., Milby, K. M., Caparroz, A. L. M. A., Pinto, A. C. P. N., Santos, R. R. P., Rocha, A. P., Ferreira, G. A., Souza, V. A., Valadares, L. D. A., Vieira, R. M. R. A., Pileggi, G. S., & Trevisani, V. F. M. (2021). Biomarkers of cytokine storm as red flags for severe and fatal COVID-19 cases: A living systematic review and meta-analysis. *PLoS ONE*, *16*(6), e0253894. https://doi.org/10.1371/journal.pone.0253894

[76] *Human Metabolome Database: Showing metabocard for Trigonelline (HMDB0000875)*. (n.d.). Retrieved November 26, 2024, from https://hmdb.ca/metabolites/HMDB0000875

[77] Omidi-Ardali, H., Lorigooini, Z., Soltani, A., Balali-Dehkordi, S., & Amini-Khoei, H. (2019). Inflammatory responses bridge comorbid cardiac disorder in experimental model of IBD induced by DSS: protective effect of the trigonelline. *Inflammopharmacology*, *27*(6), 1265–1273. https://doi.org/10.1007/s10787-019-00581-w

[78] Khalili, M., Alavi, M., Esmaeil-Jamaat, E., Baluchnejadmojarad, T., & Roghani, M. (2018). Trigonelline mitigates lipopolysaccharide-induced learning and memory impairment in the rat due to its anti-oxidative and anti-inflammatory effect. *International Immunopharmacology*, *61*, 355–362. https://doi.org/10.1016/j.intimp.2018.06.019

[79] Lonati, C., Berezhnoy, G., Lawler, N., Masuda, R., Kulkarni, A., Sala, S., Nitschke, P., Zizmare, L., Bucci, D., Cannet, C., Schäfer, H., Singh, Y., Gray, N., Lodge, S., Nicholson, J., Merle, U., Wist, J., & Trautwein, C. (2023). Urinary phenotyping of SARS-CoV-2 infection connects clinical diagnostics with metabolomics and uncovers impaired NAD+ pathway and SIRT1 activation. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *62*(4), 770–788. https://doi.org/10.1515/cclm-2023-1017

[80] *Human Metabolome Database: Showing metabocard for L-Carnitine (HMDB0000062)*. (n.d.). Retrieved November 26, 2024, from https://hmdb.ca/metabolites/HMDB0000062

[81] Fortin, G. (2011). L-Carnitine and intestinal inflammation. *Vitamins and Hormones*, 353–366. https://doi.org/10.1016/b978-0-12-386960-9.00015-0

[82] Nikolaos, S., George, A., Telemachos, T., Maria, S., Yannis, M., & Konstantinos, M. (2000). Effect of L-Carnitine Supplementation on Red Blood Cells Deformability in Hemodialysis Patients. *Renal Failure*, *22*(1), 73–80. https://doi.org/10.1081/jdi-100100853

[83] Talebi, S. S., Ghasemi, M., Etminani-Esfahani, M., Mohammadi, Y., & Haddadi, R. (2022). Effects of L-carnitine supplementation in patients with mild-to-moderate COVID-19 disease: a pilot study. *Pharmacological Reports*, *74*(6), 1296–1305. https://doi.org/10.1007/s43440-022-00402-y

[84] Al-Dhuayan, I. S. (2022). Biomedical role of L-carnitine in several organ systems, cellular tissues, and COVID-19. *Brazilian Journal of Biology*, *82*. https://doi.org/10.1590/1519-6984.267633

[85] *Human Metabolome Database: Showing metabocard for Sarcosine (HMDB0000271)*. (n.d.). Retrieved November 26, 2024, from https://hmdb.ca/metabolites/HMDB0000271

[86] Anwardeen, N. R., Cyprian, F. S., Yassine, H. M., Al-Thani, A. A., Abdallah, A. M., Emara, M. M., & Elrayess, M. A. (2023). The retrospective study of the metabolic patterns of BCG-vaccination in type-2 diabetic individuals in COVID-19 infection. *Frontiers in Immunology*, *14*. https://doi.org/10.3389/fimmu.2023.1146443

[87] Chen, X., Gu, M., Li, T., & Sun, Y. (2021). Metabolite Reanalysis Revealed Potential Biomarkers for COVID-19: A Potential Link with Immune Response. *Future Microbiology*, *16*(8), 577–588. https://doi.org/10.2217/fmb-2021-0047

6. ANNEXES

Annex 6.1. Function “knnImputation” created in this work for the imputation of zero values present in a metabolomic data set through the kNN algorithm. This function requires the input of two lists of data frames, one for concentration matrices (n patients vs p metabolites) and another for distance matrices (n patients vs n patients).

```
knnImputation <- function(valueMatrixList,distanceMatrixList) {

  impt_list <- list()

  for (m in 1:length(distanceMatrixList)) {
    dist_matrix <- as.data.frame(distanceMatrixList[m])
    val_matrix <- as.data.frame(valueMatrixList[m])
    impt <- data.frame(matrix(nrow=nrow(val_matrix),ncol=ncol(val_matrix)))

    for (n in 1:nrow(dist_matrix)) {
      ids <- c()

      while (length(ids)<5) {
        min_dist=min(dist_matrix[n,],na.rm=TRUE)
        id <- which(dist_matrix[n,]==min_dist)
        if (min_dist==0) {
          dist_matrix[n,id] <- NA }
        else {
          ids <- append(ids,id)
          dist_matrix[n,id] <- NA }}

      for (j in 1:ncol(val_matrix)) {
        if (val_matrix[n,j]==0) {
          sum=0
          for (id in ids)
            sum=sum+val_matrix[id,j]
          impt[n,j] <- sum/length(ids) }
        else
          impt[n,j] <- val_matrix[n,j] }

    rownames(impt) <- rownames(val_matrix)
    colnames(impt) <- colnames(val_matrix) }

    impt_list <- append(impt_list,list(impt)) }

  return(impt_list)
}
```

Step by step explanation:

The “knnImputation” function begins by creating an empty list for imputed data frames. Then, it cycles through the different matrices in “distanceMatrixList” and “valueMatrixList”, saving each one as a data frame. For every data frame of concentration values, the function creates a new data frame with the same number of rows and columns to hold the new imputed data. Next, it goes through the rows of each data frame and searches for the minimum value within them, saving the corresponding “PatientID” (column name) in the “id” variable. Since the first minimum value discovered in every row is always zero (distance between a patient and himself found in the matrix diagonal), this value is quickly replaced by NA and the cycle progresses to find the second minimum in that row. The next five “id” numbers are saved in the “ids” vector created earlier while the minimum values found are substituted with NA every time to allow for a new minimum to appear.

In the following step, the function cycles through the columns of the concentration matrix and replaces the zeros of a particular metabolite with a mean concentration calculated from the k nearest neighbors of a specific patient. This is done by summing the concentrations of that metabolite in every patient identified in “ids” and dividing this sum by the length of the “ids” vector, which corresponds to the parameter k=5 defined for the kNN algorithm (meaning the number of neighbors that is being considered). When the metabolite’s concentration is different than zero, the same value is returned to the imputed matrix. Finally, the row and column names of the imputed data frame are copied from the concentration matrix (n patients vs p metabolites) and the new data set is added to the “impt_list” created at the beginning of the function. The list of imputed data frames is returned to the user.

Annex 6.2. Function "dataNormalization" created in this work for the normalization of the metabolite abundances based on their initial concentrations at the beginning of the experiment. This function requires the input of a data frame with n patients vs p metabolites. Additional columns at the beginning of the data frame provide the patient and time point information. In the case of this work, the parameters used were n=20 and p=82, with 5 time points.

```
dataNormalization <- function(dataframe) {

  met_names <- colnames(dataframe[,3:84])
  columns <- c("PatientID","TimePoint",met_names)
  norm_data <- data.frame(matrix(nrow=0,ncol=length(columns)))

  ind=1

  while(ind<nrow(dataframe)+1) {

    for(t in seq(0,4)) {

      x=0
      norm_values <- list()

      for(met in dataframe[ind+t,3:84]) {

        if(is.na(met)) norm = NA
        else if(max(dataframe[ind:(ind+4),3+x],na.rm=TRUE)==0) norm = 0
        else norm = (met-dataframe[ind,3+x])/(max(dataframe[ind:(ind+4),3+x],na.rm=TRUE)-min(dataframe[ind:(ind
+4),3+x],na.rm=TRUE))

        norm_values <- append(norm_values,sub("00+$","",norm))
        x=x+1
      }

      dataframe$TimePoint <- as.character(dataframe$TimePoint)
      norm_data[,2] <- as.character(norm_data[,2])

      new_row <- c(dataframe[ind,1],dataframe[ind+t,2],norm_values)
      norm_data <- rbind(norm_data,new_row)

      dataframe$TimePoint <- as.factor(dataframe$TimePoint)
      norm_data[,2] <- as.factor(norm_data[,2])
    }

    ind=ind+5
  }

  colnames(norm_data) <- columns
  return(norm_data)
}
```

Step by step explanation:

The "dataNormalization" function starts by defining the column names for the empty normalized matrix as a vector of "PatientID", "TimePoint" and metabolite names. Next, it goes through the different patients in the data set and cycles between each time point within them, creating an empty list for the normalized values. Then, the function calculates the normalized metabolite concentrations by subtracting the concentration of that metabolite measured in t0 for a particular patient (baseline) and then diving the result by the biological range of the same metabolite in that patient, which is obtained from the difference between the max and min concentration values recorded for this metabolite over time. In the cases where the concentration of a metabolite is equal to zero for every time point (hence preventing the application of the formula above), a zero value is returned since no deviations from the baseline occurred. Lastly, NA is returned when a missing value is encountered in the data set. After removing unnecessary trailing zeros, these results and saved in the "norm_values" list for every metabolite of a particular patient measured in a specific time point.

Continuing, the function creates the new row to be incorporated in the final data frame by joining the "PatientID" and "TimePoint" in a vector with the normalized values. To make sure that the time points are presented as t0-t4, it is important to convert to characters the values of the respective columns in the original and normalized data frames before adding the new row, returning them to the factor class after the row has been added. After that, the function moves on to the next patient by advancing five rows at a time, which corresponds to the five time points recorded for each patient. To finish, the previously defined vector of column names is applied to the normalized data frame and this data set is returned to the user.

Annex 6.3. Function "knnImputationForNA" created in this work for the imputation of N/A values present in a metabolomic data set through the kNN algorithm. This function requires the input of two lists of data frames, one for concentration matrices (n patients vs p metabolites) and another for distance matrices (n patients vs n patients).

```
knnImputationForNA <- function(valueMatrixList,distanceMatrixList) {

  impt_list <- list()

  for (m in 1:length(distanceMatrixList)) {
    dist_matrix <- as.data.frame(distanceMatrixList[m])
    val_matrix <- as.data.frame(valueMatrixList[m])
    impt <- data.frame(matrix(nrow=nrow(val_matrix),ncol=ncol(val_matrix)))

    for (n in 1:nrow(dist_matrix)) {
      ids <- c()

      while (length(ids)<5) {
        min_dist=min(dist_matrix[n,],na.rm=TRUE)
        id <- which(dist_matrix[n,]==min_dist)
        if (min_dist==0) {
          dist_matrix[n,id] <- NA }
        else {
          ids <- append(ids,id)
          dist_matrix[n,id] <- NA }}

      for (j in 1:ncol(val_matrix)) {
        if (is.na(val_matrix[n,j])) {
          sum=0
          for (id in ids) {
            if (is.na(val_matrix[id,j])==FALSE)
              sum=sum+val_matrix[id,j] }
          impt[n,j] <- sum/length(ids) }
        else
          impt[n,j] <- val_matrix[n,j] }

    rownames(impt) <- rownames(val_matrix)
    colnames(impt) <- colnames(val_matrix) }

    impt_list <- append(impt_list,list(impt)) }

  return(impt_list)
}
```

Step by step explanation:

The "knnImputationForNA" function begins by creating an empty list for imputed data frames. Then, it cycles through the different matrices in "distanceMatrixList" and "valueMatrixList", saving each one as a data frame. For every data frame of concentration values, the function creates a new data frame with the same number of rows and columns to hold the new imputed data. Next, it goes through the rows of each data frame and searches for the minimum value within them, saving the corresponding "PatientID" (column name) in the "id" variable. Since the first minimum value discovered in every row is always zero (distance between a patient and himself found in the matrix diagonal), this value is quickly replaced by NA and the cycle progresses to find the second minimum in that row. The next five "id" numbers are saved in the "ids" vector created earlier while the minimum values found are substituted with NA every time to allow for a new minimum to appear.

In the following step, the function cycles through the columns of the concentration matrix and replaces the missing values (NA) of a particular metabolite with a mean concentration calculated from the k nearest neighbors of a specific patient. This is done by summing the concentrations of that metabolite in every patient identified in "ids" and dividing this sum by the length of the "ids" vector, which corresponds to the parameter k=5 defined for the kNN algorithm (meaning the number of neighbors that is being considered). As a result of using the same length (ids) for all calculations, neighboring metabolite concentrations with an NA value are treated as zero since they are not added to the sum. When the metabolite's concentration is different than NA, the same value is returned to the imputed matrix. Finally, the row and column names of the imputed data frame are copied from the concentration matrix (n patients vs p metabolites) and the new data set is added to the "impt_list" created at the beginning of the function. The list of imputed data frames is returned to the user.