1. **Title Page**

- Title of the report (e.g., *Data Cleaning and Management for Mouse Phenotype Data Analysis*).
- Group number and name of database
- URL of the GitHub repository with scripts and database backups.

2. **Abstract**

- Summarize the project's objective, key steps, results, and conclusion.
- Highlight the importance of phenotypic data analysis and dashboard development.

3. **Introduction**

- Provide background on the **International Mouse Phenotyping Consortium (IMPC)**:
    - Purpose and methodology of the IMPC (e.g., knockout mice, phenotype testing, p-value thresholds, etc.).
    - Significance of mouse models for understanding human diseases.
- Define the goals of this project:
    - Data cleaning and management of IMPC data.
    - Creating a **MySQL database** and an **RShiny dashboard** for data exploration.
- Overview of the collaborator's requests:
    - Visualizing phenotype scores by genotype and phenotype
    - Querying the database for four genotypes of interest.

4. **Methods**

4.1 Data Cleaning

- Detail the data cleaning steps performed on the .csv files:
    - Handling missing or inconsistent data.
    - Standardization and formatting of variables (e.g., phenotypic parameters).
    - Explain why each step was necessary for accurate downstream analysis.

4.2 Data Collation

- Explain how cleaned data was structured in R for:
    - Tidy data format.
    - Efficient transfer to MySQL and compatibility with RShiny.

4.3 MySQL Database Design

- Describe the database schema:
    - Tables created (e.g., phenotypes, genotype-phenotype mappings, disease links).
    - Key relationships and normalization strategies.
- Explain how the schema supports:
    - Querying genotypes of interest.
    - Adding new data (scalability).
- Include an annotated diagram of the schema

4.4 RShiny Dashboard

- Outline the dashboard features:
    - Interface design and user interaction.
    - Plots for genotype-specific and phenotype-specific statistical scores.
    - Gene cluster visualization.
- Briefly describe the RShiny implementation process.

5. Results

5.1 Database Queries

- Provide MySQL commands used to retrieve information on the four genotypes of interest.
- Explain the results, including:
    - Parameters, p-values, and associated diseases.
    - Phenotypic procedures and groupings.

5.2 Dashboard Outputs

- Include and describe the four plots generated for the collaborator's requests:
    1. Phenotype scores of all phenotypes tested for a selected knockout.
    2. Scores of all knockouts for a selected phenotype.
    3. Clusters of genes with similar phenotype scores.
- Discuss insights from the visualizations.

6. Discussion

- Evaluate the workflow:
    - Strengths and limitations of data cleaning, collation, and management steps.
    - Challenges in database design (e.g., parameter grouping, scalability).
    - Usability and performance of the RShiny dashboard.
- Reflect on how well the project met the collaborator's requirements.

7. Conclusion

- Summarize the achievements and their importance for the IMPC and biomedical research.
- Provide suggestions for future improvements (e.g., expanding parameter groupings, enhancing the dashboard).

8. References

- Cite any sources used, including the IMPC homepage, papers, and any external tools or libraries.

Appendices (if needed)

- Additional materials:
- Full MySQL schema.
- Sample R code snippets for critical steps.
- Backup files or supplementary figures.

Since you're unable to proceed with the coding for now, you can focus on preparing foundational sections of your report and planning the workflow. Here's what you can work on:

1. **Flesh Out the Introduction**

- **Background Research**:
    - Expand on the IMPC's impact on biomedical research by summarizing key findings or examples from their publications (accessible from the IMPC website).
    - Describe how data from the IMPC has been used in human disease research or therapeutic developments.
- **Significance of the Project**:
    - Highlight how data cleaning, database creation, and dashboard visualization improve the utility of large datasets like those from the IMPC.

2. **Methods (Framework and Planning)**

Although you can't run the coding tasks, you can:

a. Data Cleaning Plan

- **Outline Steps**:
    - Document potential cleaning steps like handling missing data, identifying inconsistencies, removing duplicates, or normalizing column formats.
    - Describe how you might group parameters into meaningful categories (e.g., weight, imaging, neurological traits).
- **Plan for Missing Data**
    - Decide on strategies for handling missing values, such as imputation, exclusion, or flagging.

b. Database Design

- **Draft a Schema**:
    - Plan the structure of your MySQL database.
    - Identify tables (e.g., phenotypes, genotypes, parameter groupings, human disease associations) and define key relationships.
    - Sketch a normalized schema with relationships (e.g., ER diagrams).
- **Write Database Queries**:
    - Draft potential MySQL queries to retrieve collaborator-required outputs, such as phenotypes for specific genotypes or linked diseases.

c. Dashboard Planning

- **Feature Mapping**:
    - Define key features for the dashboard:
    - Selection menus for genotypes and phenotypes.
    - Interactive plots for phenotype significance or gene clustering.
    - Clear, user-friendly layout for intuitive data exploration.
- **Visualization Ideas**:
    - Plan visualizations (e.g., bar charts, heatmaps, scatter plots).

3. **Reflection Section Draft**

Begin drafting insights for the reflection section:

- **Ease of Replicating Technical Work**:
    - Note any challenges with the coursework instructions or tools (e.g., TRE access).
- **Data Challenges**:
    - Highlight anticipated issues, like dealing with complex parameter groupings or inconsistencies in the data.

4. Figures and Diagrams

- **Database Schema**:
    - Create a draft ER diagram for your MySQL database.
- **Workflow Diagram**:
    - Sketch out the workflow for the project, from data cleaning to visualization.
- **Dashboard Mockups**:
    - Design simple mockups of the RShiny dashboard's layout and functionality.

5. Literature Review and References

- Research articles or resources that discuss:
    - The significance of mouse models in human disease research.
    - Best practices for data cleaning, database design, or dashboard creation.
- Organize your references to save time when finalizing the report.

6. Writing Additional Sections

- **Abstract**:
    - Refine the draft abstract to align with your planned methodology and goals.
- **Conclusion**:
    - Outline how your work will provide meaningful insights and enhance data usability for IMPC collaborators.

Potential Titles

Data Cleaning and Management for Phenotypic Scoring Data: A Case Study with IMPC Mouse Models”

“Designing Scalable Databases and Interactive Dashboards for Mouse Phenotype Analysis”

“Streamlining Mouse Phenotype Data Analysis: A Data Management and Visualization Approach”

“Building Reproducible Pipelines for Cleaning and Managing IMPC Phenotypic Data

“From Raw Data to Insights: Developing a Data Infrastructure for Mouse Phenotype Research

“Integrating MySQL and RShiny for Interactive Exploration of Mouse Phenotype Data

“Data-Driven Insights: Cleaning, Managing, and Visualizing IMPC Phenotypic Data

“An Analytical Framework for Exploring Gene-Phenotype Relationships in Mouse Models

“Enhancing Biomedical Research with Clean Data and Visual Dashboards: A Study on Mouse Phenotypes”

“A Functional Data Pipeline for Phenotype-Genotype Mapping in IMPC Research”

CONTENTS

ABSTRACT

The International Mouse Phenotyping Consortium (IMPC) generates extensive phenotypic data to study the functional roles of genes in mice, offering critical insights into human diseases. This project aimed to process, manage, and visualize pre-analyzed phenotypic scoring data to meet the requirements of IMPC collaborators. Key objectives included cleaning and standardizing phenotype data, designing a scalable MySQL database, and developing an interactive RShiny dashboard for data exploration. Phenotypic data analysis is essential for uncovering gene-disease relationships, supporting biomedical research, and guiding therapeutic discovery. To maximize the value of IMPC data, this project aimed to process, manage, and visualize pre-analyzed phenotypic scoring data through robust data cleaning, database management, and dashboard development.

Data cleaning ensured the consistency, completeness, and accuracy of datasets, addressing missing values, standardizing variables and grouping parameters for downstream analyses and biologically meaningful categories. Cleaned data were structured into tidy formats and used to design a scalable. Cleaned data were structured into tidy formats and used to design a normalized MySQL database schema, enabling efficient querying of genotype-phenotype associations and parameter groupings. MySQL database that efficiently stores genotype-phenotype associations, enabling seamless data retrieval and query optimization. An interactive RShiny dashboard was developed to facilitate intuitive visualizations of phenotypic data, enabling the exploration of genotype-specific and phenotype-specific results, as well as clustering genes with similar phenotypic profiles.

The workflow's success was demonstrated by the accurate retrieval of phenotype information for selected genotypes, alongside insightful visualizations that highlighted significant gene-phenotype associations. This integrated workflow highlights the transformative potential of phenotypic data analysis and dashboard tools in making complex datasets accessible, actionable, and interpretable for researchers. Challenges in data cleaning and parameter grouping underscored the need for further refinements to enhance scalability and usability. This study provides a framework for managing and exploring complex phenotypic data, but also laid a foundation for scalable and reproducible data management with applications in biomedical research and more.

INTRODUCTION

The International Mouse Phenotyping Consortium (IMPC) is a global initiative dedicated to systematically identifying the function of every protein-coding gene in the mouse genome. By creating and analyzing knockout mouse strains—where specific genes are inactivated—the IMPC aims to build a comprehensive catalog linking genes to phenotypes, thereby enhancing our understanding of mammalian gene function and its relevance to human health.

Purpose and Methodology of the IMPC:

- **Gene Knockout Production:** The IMPC generates homozygous knockout mouse strains for each protein-coding gene, typically producing cohorts of seven males and seven females per gene.
- **Phenotype Testing:** These mice undergo a standardized phenotyping pipeline to assess a wide array of biological and physiological parameters. The pipeline includes evaluations at various life stages:
- **Viability Assessment:** Conducted at the pre-weaning stage to determine if homozygous knockouts are viable.
- **Early Adult Pipeline:** Viable homozygous mice are subjected to comprehensive phenotyping to detect deviations from baseline wildtype characteristics.
- **Embryo Pipeline:** For knockouts that are non-viable, additional analyses are performed during embryonic development to ascertain the stage and cause of lethality.
- **Data Standardization and Quality Control:** Phenotyping centers across five continents adhere to uniform protocols to ensure data consistency. This includes maintaining standardized sample sizes and continually testing wildtype mice to provide baseline values for comparison.

The International Mouse Phenotyping Consortium (IMPC) is a global initiative dedicated to uncovering the function of every protein-coding gene in the mouse genome. This is achieved by systematically producing knockout mouse strains, where specific genes are inactivated, to assess their impact on biological and physiological traits. Using standardized protocols, the

IMPC generates cohorts of homozygous knockout mice, typically consisting of seven males and seven females for each gene.

Phenotype testing is conducted across multiple life stages to capture a comprehensive profile of gene function. For viable homozygous knockouts, an early adult phenotyping pipeline evaluates a broad range of biological parameters. If homozygous knockouts are non-viable, additional analyses are performed during embryonic development to determine the stage and cause of lethality. Pre-weaning viability assessments are also carried out to confirm survival beyond early developmental stages.

To ensure data reliability, phenotyping centers on five continents follow uniform procedures, including standardized sample sizes and baseline wildtype testing. These efforts facilitate high-quality, reproducible data collection, which is critical for establishing robust genotype-phenotype associations.

Significance of Mouse Models for Understanding Human Diseases:

Mice serve as invaluable models in biomedical research due to their genetic and physiological similarities to humans. Approximately 99% of mouse genes have human counterparts, and mice possess comparable organ systems and biological functions. These parallels enable researchers to study human disease mechanisms, gene functions, and potential therapeutic interventions in a controlled and ethical manner.

The IMPC's systematic approach to gene function analysis facilitates the identification of novel gene-phenotype relationships, many of which have direct implications for human health. By uncovering the roles of genes and their associations with specific phenotypes, the IMPC contributes to the development of new disease models, enhances our understanding of genetic disorders, and supports the discovery of targeted treatments.

In summary, the IMPC's efforts in generating and analyzing knockout mouse models are pivotal for advancing our comprehension of gene functions and their relevance to human diseases, thereby driving progress in medical research and therapeutic development.

- Define the goals of this project:

The International Mouse Phenotyping Consortium (IMPC) is a global initiative committed to uncovering the functional roles of all protein-coding genes in the mouse genome. By systematically creating knockout mouse strains, where individual genes are inactivated, the IMPC aims to identify gene-phenotype associations that provide crucial insights into mammalian biology and human diseases. Knockout mice undergo rigorous testing through a standardized phenotyping pipeline to measure a wide array of biological and physiological parameters. This approach is implemented across multiple life stages, including pre-weaning viability checks, early adult assessments for viable knockouts, and embryonic analyses for non-viable knockouts. To maintain data consistency and quality, all participating phenotyping centers adhere to uniform protocols, enabling the generation of reproducible, high-quality datasets.

Mice are invaluable models for biomedical research due to their genetic and physiological similarities to humans, with approximately 99% of mouse genes having human counterparts. The IMPC's systematic approach to gene function analysis allows for the discovery of novel gene-phenotype relationships, many of which have direct relevance to human health. By contributing to the development of new disease models and targeted therapies, the IMPC plays a pivotal role in advancing medical research and therapeutic innovation.

The primary goals of this project are to address key data management and exploration challenges faced by IMPC collaborators. This includes cleaning and standardizing phenotypic scoring data to ensure accuracy and consistency, as well as organizing the cleaned data into a scalable MySQL database for efficient storage and retrieval. Furthermore, an interactive RShiny dashboard will be developed to facilitate data exploration. The dashboard will enable users to visualize statistical scores for specific knockout genotypes, examine phenotype-specific patterns, and identify clusters of genes with similar phenotypic profiles. These tools aim to enhance the accessibility and usability of IMPC data, empowering researchers to derive meaningful insights from complex datasets.

IMPC's Role in Biomedical Advancement

Flesh Out the Introduction

Background Research:

The International Mouse Phenotyping Consortium (IMPC) plays a pivotal role in advancing biomedical research through its systematic exploration of mouse genetics and phenotypes. Established to create a comprehensive catalog of mouse gene functions, the IMPC employs a standardized methodology that includes the generation of knockout mice, extensive phenotype testing, and rigorous statistical analysis. By utilizing knockout mice—where specific genes are intentionally inactivated—the IMPC enables researchers to observe the resultant phenotypic changes and infer the gene's role in biological processes. This methodology is critical for understanding gene function and its implications in human diseases.

One of the significant contributions of the IMPC is its focus on sex differences in biomedical research. A recent study conducted by the IMPC, published in Nature Communications, highlights the importance of considering sexual dimorphism in research. The study analyzes 234 physical characteristics across more than 50,000 mice, revealing that sex significantly influences 56.6% of quantitative traits, such as body composition and metabolic profiles, and 9.9% of qualitative traits, including head shape and coat characteristics. In mutant mice, where specific genes are switched off, sex modifies the effects of mutations in 13.3% of qualitative traits and up to 17.7% of quantitative traits. This finding emphasizes that male and female mice exhibit distinct biological responses, which should be taken into account in experimental designs.

The implications of these findings extend beyond mouse models to human health. The understanding that sex differences impact various traits informs the design of clinical trials and therapeutic developments. Historically, biomedical research has exhibited a bias toward male subjects, with approximately 85% of animal studies and 41% of clinical trial participants being male. Such biases can obscure critical information about how diseases manifest and respond to treatments in females. Dr. Natasha Karp, lead author of the IMPC study, articulates that the oversight of sex differences in research represents a significant blind spot, potentially

resulting in ineffective treatments for women and a misunderstanding of disease prevalence and severity.

Moreover, the data generated by the IMPC is instrumental in elucidating the genetic underpinnings of human diseases. By establishing a comprehensive database of gene functions and associated phenotypes, the IMPC facilitates the identification of gene variants linked to diseases such as cardiovascular disorders, autoimmune diseases, and metabolic syndromes. Researchers can draw parallels between mouse models and human conditions, enabling them to develop targeted therapies that consider genetic and sex-related factors. This approach aligns with the movement toward precision medicine, which seeks to tailor medical treatments to individual patient characteristics.

The IMPC's systematic approach to phenotype characterization, combined with its emphasis on sex differences, signifies a transformative shift in biomedical research. As highlighted by Professor Steve Brown, a key figure in the IMPC, the neglect of sex differences in research has likely led to the oversight of critical scientific information. By advocating for the inclusion of both sexes in research designs, the IMPC not only enhances the robustness of scientific findings but also paves the way for more effective and inclusive healthcare solutions.

In conclusion, the IMPC's contributions to biomedical research underscore the importance of comprehensive genetic studies and the consideration of sex differences. By leveraging mouse models to explore gene function and disease mechanisms, the IMPC stands as a vital resource in the quest for improved understanding and treatment of human diseases.

Reference for **IMPC's Role in Biomedical Advancement:**
https://www.ucl.ac.uk/news/2017/jun/sex-differences-important-medical-research

INTRO SECTION 2

The International Mouse Phenotyping Consortium (IMPC) is a global initiative dedicated to elucidating the functions of all protein-coding genes in the mouse genome. By systematically creating knockout mouse strains—where specific genes are inactivated—the IMPC aims to establish a comprehensive catalog linking genes to phenotypes. This endeavor significantly enhances our understanding of mammalian biology and its relevance to human health.

The IMPC's impact on biomedical research is profound. Through large-scale phenotyping, the consortium has generated extensive data that serve as a valuable resource for the scientific community. For instance, analysis of genome-wide knockout mouse data has identified numerous genes associated with specific phenotypes, providing insights into gene function and disease mechanisms.

Moreover, the IMPC's efforts have led to the development of new disease models. Based on data from 3,328 genes, 360 new disease models have been identified, allowing researchers to investigate the molecular mechanisms underpinning human genetic diseases and explore new therapeutic interventions.

The data generated by the IMPC have been instrumental in human disease research and therapeutic development. By providing comprehensive phenotypic information, the consortium enables the identification of gene-phenotype associations that mirror human conditions. This facilitates the study of disease pathways and the development of targeted treatments. For example, the IMPC's comprehensive knockout phenotyping has underpinned the study of human disease by evaluating the recapitulation of human phenotypes for known gene–disease associations in mouse knockouts and uncovering potential novel gene–phenotype associations.

In summary, the IMPC's systematic approach to gene function analysis has significantly advanced biomedical research. By generating and analyzing knockout mouse models, the consortium provides invaluable resources that enhance our understanding of gene functions and their implications for human diseases, thereby driving progress in medical research and therapeutic development.

Reference: INTRO SECTION 2

Main:

https://www.sciencedirect.com/science/article/pii/S1871678415001624

https://www.sanger.ac.uk/group/mouse-pipelines/

Additional Citations:

https://www.nature.com/articles/s41598-022-19710-7?utm_source=chatgpt.com

https://pmc.ncbi.nlm.nih.gov/articles/PMC6061128/?utm_source=chatgpt.com

https://academic.oup.com/nar/article/51/D1/D1038/6777802?utm_source=chatgpt.com

https://pmc.ncbi.nlm.nih.gov/articles/PMC8490014/

https://pmc.ncbi.nlm.nih.gov/articles/PMC6061128/

https://academic.oup.com/nar/article/51/D1/D1038/6777802

https://academic.oup.com/bioinformatics/article/40/8/btae469/7721930

https://www.ebi.ac.uk/training/events/international-mouse-phenotyping-consortium-webinar/

https://www.genome.gov/Funded-Programs-Projects/Knockout-Mouse-Phenotyping-Project-KOMP2

https://www.nature.com/articles/s41598-022-19710-7

https://news.mit.edu/2023/new-way-evaluate-impact-medical-research-0814

Methods
Data Cleaning

Data Collation

MySQL Database Design

Database Creation and Schema Design

The MySQL database was designed to effectively manage and analyse genotype-phenotype associations in mice, incorporating key phenotypic data, genotype information. This schema consists of five core tables: Mouse, Phenotype, Parameter_Group, Procedures, and Human_Disease. Careful normalisation and relational design were employed to ensure efficient querying, supporting future scalability.

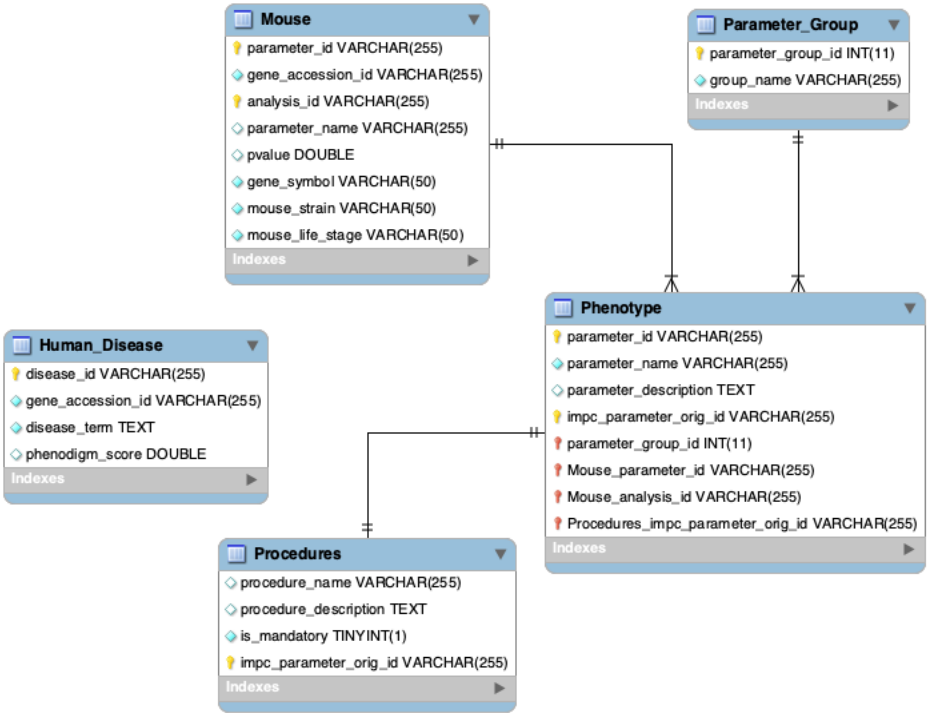

Figure : ER diagram of the MySQL database schema. The schema consists of five tables Mouse, Phenotype, Parameter_Group, Procedures, and Human_Disease, with relationships defined through foreign keys to support efficient querying and data management. The Phenotype table acts as a central table, linking to Mouse, Procedures, and Parameter_Group through foreign keys.

The Mouse table stores genotype information along with parameters, p-values, and metadata such as the mouse strain and life stage. A composite primary key (parameter_id, analysis_id) ensures the uniqueness of each parameter and analysis combination. This table is directly linked to the Phenotype table via the parameter_id column, allowing queries to retrieve related phenotypic data for specific genotypes.

The Phenotype table contains detailed information about the observed phenotypic parameters, including names, descriptions, and their association with parameter groups. To

ensure data integrity, a composite primary key (parameter_id, impc_parameter_orig_id) was implemented. Foreign key constraints were established to link the Phenotype table with both the Parameter_Group and Procedures tables. This design allows the database to categorise parameters under specific groups, such as “Weight Parameters” or “Cardiovascular Parameters,” and to associate each phenotype with a defined procedure.

The Parameter_Group table was created to group phenotypic parameters into predefined categories, facilitating efficient querying and organisation of related data. The parameter_group_id column serves as the primary key and is referenced by the Phenotype table to assign each parameter to its respective group. Parameter groups were manually assigned based on the parameter file provided . The SQL UPDATE statements ensured that all parameters in the Phenotype table were linked to their respective parameter groups in the Parameter_Group table. The groupings are as follows Weight, Image, Brain, Cardiovascular, Heart and Electrolytes.

Normalisation was applied to the database by moving redundant information, into separate tables, which are then referenced using foreign keys. This approach ensures that updates to group names or procedures only need to be made in a single location, thereby improving the maintainability of the database and enhancing its scalability for future expansions.

The Procedures table documents the experimental procedures used to collect phenotypic data. Each procedure is uniquely identified by an impc_parameter_orig_id, which also serves as the primary key. This table ensures that every phenotype has a corresponding procedure, maintaining consistency in the recorded experimental methods.

Finally, the Human_Disease table stores human disease information, including associated genes and phenodigm scores, which ideally should quantify phenotype similarity between mice and humans. This table can be queried in conjunction with the Mouse and Phenotype tables to explore potential genotype-disease relationships.

The database schema aims to eliminate redundancy. One to many relationships were established where appropriate, such as between Parameter_Group and Phenotype, and

between Procedures and Phenotype. These relationships enable efficient data retrieval without duplicating information across tables. For example, a query to retrieve all phenotypic data associated with a specific genotype, such as Phyhd1, involves joining the Mouse, Phenotype, Parameter_Group, and Procedures tables

The design also ensures scalability, allowing for seamless integration of new data. Adding new phenotypes, procedures, or parameter groups is straightforward, as the schema accommodates extensions without requiring structural modifications. For instance, when a new parameter group is introduced, it can be added to the Parameter_Group table, and any phenotypes related to that group can reference its ID in the Phenotype table.

New phenotypes and their associated procedures can be added using simple SQL insert statements. The auto increment feature for parameter_group_id in the Parameter_Group table ensures that newly added groups receive a unique identifier automatically, minimising manual effort and potential errors.

By enforcing foreign key constraints and ensuring proper indexing on key columns such as parameter_id and gene_symbol, the schema guarantees data consistency and enhances query performance. This relational design enables efficient retrieval of data for analysis, whether querying specific genotypes, identifying significant phenotypes, or exploring genotype-disease correlations.

Query Design for Genotype Analysis

The schema was designed to facilitate efficient querying of genotype related data, particularly for identifying phenotypes and their associated parameters, parameter groups, and procedures. By ensuring normalisation and defining foreign key constraints, complex queries can be executed without data inconsistency issues. Once the database was populated, queries were developed to retrieve detailed information on the collaborator's four genotypes of interest (Phyhd1, Cwc22, Slc7a13, Emcn). These queries were structured to identify associated phenotypes for each genotype, retrieve parameter group details, extract relevant procedural information, display statistical significance through p-values.

In addition to supporting complex queries, the schema was designed with future data expansion in mind. As new genotypes, phenotypes, and experimental procedures are discovered, they can be readily integrated into the existing database structure without disrupting current functionality.

This carefully structured MySQL database not only meets current research requirements but also provides a frame for future investigations. The ability to efficiently query genotypes of interest, retrieve relevant phenotypes, and explore parameter groupings and procedures makes it a valuable tool for collaborative research and analysis.

RShiny Dashboard

Results
Database Queries

Dashboard Outputs

Discussion

Reflection

Literature Review

Conclusion

Prompts

Are you able to draft an ER diagram. "The mice from each analysis were tested for a particular phenotype parameter that has been assigned a unique identifier and name (parameter_id and parameter_name). To reduce the parameter space the collaborator would like to start grouping parameters together based on phenotype test similarity and/or naming similarity. Design the database to reflect new groupings for weight, images and brain parameters, plus at least 3 others of your choice. They are interested in using the MySQL database and the RShiny dashboard to query four genotypes that they are working on."