University of Macau

Institute of Collaborative Innovation

collaborative with

Faculty of Health Sciences

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Building a Shiny Web Application to Promote Data Management in a Multi-Center Project with Rodent Shrapnel Model

by

Tianhao Chen, Student No: MB955389

Graduation Project Report submitted in partial fulfillment
of the requirements of the Degree of
Master of Science in Data Science (Precision Medicine)

Project Supervisor

Prof. Qi Zhao

07 May 2022

DECLARATION

I sincerely declare that:

1. I am the sole authors of this report,
2. All the information contained in this report is certain and correct to the best of my knowledge,
3. I declare that the thesis here submitted is original except for the source materials explicitly acknowledged and that this thesis or parts of this thesis have not been previously submitted for the same degree or for a different degree, and
4. I also acknowledge that I am aware of the Rules on Handling Student Academic Dishonesty and the Regulations of the Student Discipline of the University of Macau.

Signature : _____*Tianhao Chen*_____

Name : Tianhao Chen

Student No. : MB955389

Date : 07 May 2022

Acknowledgement

This project report is carried out under the enthusiastic direction of my supervisor Professor Xiaohua Douglas Zhang. I would like to gratefully acknowledge his advice, support, and encouragement during my studies at University of Macau. It is my great fortune to do research with him.

I truly appreciate Professor Qi Zhao and all examination committee members for their review and comments on this project report.

During my work on this project report, many people provided me with excellent support and advice, so I should own a special note of appreciation to them. I appreciate Ming Deng and Dongliang Leng for their help and the discussions in the early stage of this work. I would like to express thanks to Kuan Cheok (Johnny) Lei and Xizhi Luo for the assistance with transcriptomic data processing, especially in the quality control of the NGS raw data, and for answering my questions in programming. In building a web-based application, I thank Weimin Tan for identifying software requirements with me and introducing me to some tools and mindsets in software development, Yufei Li for the technical discussions and assistance in system design and implementation. I thank Jiaqi Xu for answering my questions in bioinformatics. I thank the members of the biotrainee for generously sharing their experiences and codes. Besides, I am grateful to collaborators that I have never met in the United States for offering me a chance to apply my data science skills to this high-quality biological research project. The rest of the members at Zhang's Lab also helped me a lot, I genuinely appreciate them.

Additionally, many thanks to the excellent staff members, Stephanie Ng, Lorna

Chau, Andrew Liu, and to Professor Terence Poon, and other faculty from the Faculty of Health Sciences and the Department of Computer and Informatics Sciences, University of Macau. I also want to express my thanks to Professor Baoqing Sun in First Affiliated Hospital of Guangzhou Medical University for her guidance during my stay here in 2020. I am grateful to all my friends, and anyone I met during my wonderful lifetime journey.

Last but not least, I would like to extend my gratitude to my family members for their endless support.

Abstract

Research data management refers to the strategies for handling data at all stages of a study and aims to promote data FAIR: Findable, Accessible, Interoperable, and Reusable. Therefore, it is considered a major part of any biomedical investigations, especially with multi-center collaborations, which may have multiple components from different institutions or regions due to research considerations. In this project report, I mainly contribute to data management with multi-center research data. The multi-center research project that I joined is carried out in response to the raising concerns on injuries with embedded metal fragments as the unfortunate consequences of modern armed conflicts, in which researchers have utilized a rodent model, integrating biokinetics, toxicology, and bioinformatics approaches, with the purpose of comprehensively investigate the long-term health effects of military-relevant pure metals implantation on rats. For the experiments part, rats were bilaterally implanted with pure metal pellets of any of Al, Co, Cu, DU, Fe, Ni, Pb, Ta, or W at their gastrocnemius muscles and assigned to cohorts to be followed up at 1-, 3-, 6-, 12-month. Euthanasia was performed on the rats at their designated experimental endpoints to harvest various biological samples, and research data was generated with specific research objectives. As a part of the greater collaboration on data management, I make a complemental effort to conduct data curation and hereby prepare a web-based platform for data storage and queries employing Shiny framework with Golem toolkit, where researchers could search and download data directly. The web application is currently available at: https://fhs1027.shinyapps.io/MetalEmbeddedStudies/.

Keywords: research data management, data FAIR, multi-center project, rodent model, embedded metal, web application, Shiny

Contents

Chapter 1 Introduction

1.1 Background and Significance

Injuries with embedded metal fragments are the unfortunate consequence of modern armed conflict. Having balanced the surgical risks and the long-term health risks of embedded metal fragments, standard surgical guidelines recommend leaving the fragments in place except for certain circumstances, so many U.S. defense personnel have to survive with the embedded fragments in their bodies. Traditionally, embedded metal fragments are considered inert. However, the lack of understanding of the long-term health effects of the embedded novel military materials, such as depleted uranium (DU), has raised many concerns. Since the first shrapnel injuries model for studying the health effects of rats with surgically implanted DU pellets in the gastrocnemius (leg) muscle was developed in 1996 at Armed Forces Radiobiology Research Institute in Bethesda, MD (AFRRI) (Castro, Benson, Bogo, Daxon, & Hogan, 1996), the rodent shrapnel model has used to investigate the biokinetics and toxicology of Du implantation (Pellmar, 1999). Metal-induced soft tissue sarcomas were observed in the rats implanted with DU pellets (Hahn, Guilmette, & Hoover, 2002). Later, the research object of embedded metal fragments extended to a range of military alloys, such as tungsten-based alloys. In the rodent model, tungsten/nickel/cobalt alloy embedded pellets can induce highly aggressive rhabdomyosarcomas (Kalinich et al., 2005). Still, the same result was not obtained in the case of tungsten/nickel/iron alloy, which may be related to the surface reactions between the materials and soft tissues (Emond & Kalinich, 2012). Additional to carcinogenic effects, neurotoxicity effects of

military-relevant metal implantation in rodent models (Fitsanakis et al., 2006) and reproductive toxicity effects in cell line models were also observed (Bardack, Dalgard, Kalinich, & Kasper, 2014). The above research shows that the biological effects and long-term health outcomes induced by different embedded metals are diverse. Collecting various samples, such as soft tissue samples near the implants, and urine and serum samples, is critical for conducting a comprehensive investigation in an animal study (Kalinich & Kasper, 2016).

1.2 Multi-Center Project with Rodent Shrapnel Model

1.2.1 Project Overview

In response to the increasing concerns about the health effects of embedded metal fragments, the U.S. Department of Defense (DoD) released the Health Affairs Policy Memo 07-029 (policy on analysis of metal fragments removed from Department of Defense personnel; 2007), which listed a panel of metals needing special attention, including Tungsten (W), Nickel (Ni), Copper (Cu), Lead (Pb), Cobalt (Co), Du, Aluminum (Al), iron (Fe), etc. Researchers noticed that previous research mainly focused on military alloys as implanted metals, while there is a lack of research on military-relevant pure metals. A multi-center research project with a rodent shrapnel model is carried out to bridge this research gap, integrating biokinetics, toxicology, and bioinformatics approaches, to comprehensively investigate the health effects of the above eight types of military-relevant pure metals implantation on rats. This project is led by the Armed Forces Radiobiology Research Institute in Bethesda (AFRRI) which involved the development of the rodent shrapnel model, while researchers from the University of Maryland School of Medicine, the Baltimore Department of Veterans' Affairs Medical Center (MDVA), the U.S. Food and Drug Administration (FDA), and

the University of Kentucky (UK), participate the project. Researchers from distinct centers proposed a variety of sub-studies based on their research interests and played heterogeneous roles with their interdisciplinary backgrounds and research approaches.

The multi-center research project can be divided into three stages. The first stage is performing animal experiments in AFRRI, in which rats were treated for collecting biological samples. Once all the animal experiments were finished, the project moved to the second stage of biological sample analysis. The samples were sent to the partner institutions following the research plan, and researchers analyzed the samples and generated research data for specific research objectives independently. The last stage is for data analysis, knowledge discovery, and publications. Researchers kept close communication in all the stages, the related issues included but were not limited to proposing sub-studies, biological samples distribution, and data management.

Here, I briefly illustrate the overall study design (figure 1.1). Animal experiments were performed at AFRRI after obtaining ethical approval (AFRRI IACUC protocol no. 2016-05-006). Three-month-old male Sprague-Dawley rats were randomly assigned to nine embedded metal exposure groups (eight military-relevant pure metals and tantalum (Ta) as the control group), with 8 biological replicates in each group. Each metal group had 4 timed cohorts of 1-, 3-, 6- and 12-month as designated experimental endpoints. A total of 288 rats (8 biological replications × 9 metal groups × 4 timed cohorts) were implanted with the corresponding metal pellets in their gastrocnemius. Urine, serum, and multiple tissue samples of rats in different embedded metal groups were collected at different time points (up to one-year follow-up). Urine samples are collected through the live animals, while biological tissues like muscle are collected after euthanasia was performed at the designated endpoints of the timed cohorts. Currently available data from two sub-studies is pointed out in the figure. Sub-

study of transcriptomics profiling was performed on muscle tissues, urine and serum samples through microarray experiments or sequencing experiments. Sub-study of metal concentrations was carried out via Inductively Coupled Plasma Mass Spectrometry (ICP-MS) with urine and serum samples. Additionally, researchers weighed each rat every week until euthanasia after metal implantation, and such records are provided.

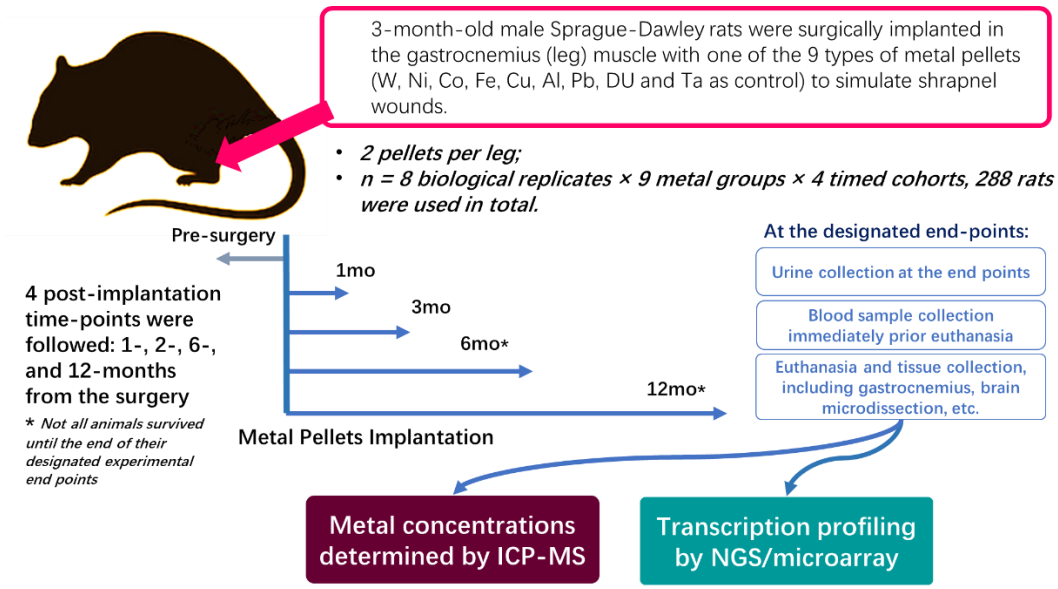


Figure 1.1: Study design of the multi-center project with currently available data.

By analyzing the generated data, researchers may identify the specific health risks caused by the implanted metals and find potential biomarkers indicative of those risks at an early enough stage. Analyses of parts of this dataset have been published in separate high-quality journals using muscle transcriptomics data and urine miRNA-Seq data to identify transcriptional changes in skeletal muscle tissue in response to embedded metal exposure (Wen et al., 2020) and to identify a group of miRNAs as potential urine biomarkers (Vechetti et al., 2021), respectively. Moreover, data analysis of the urine metal concentrations suggested that solubilized metals passed through the kidney and were excreted in the urine (Hoffman, Vergara, Fan, & Kalinich, 2021).

More data analysis and sample analysis are being conducted.

1.2.2 Data Management and Data FAIR

Research data management providing stagiaires for handling data at all stages of a study, thus it is acknowledged as a critical component in any scientific research (Yang & Yang, 2022). It involves a biomedical research data lifecycle, generally spanning from “Plan & Design”, “Collect & Create”, “Analyze & Collaborative”, “Store & Manage”, “Evaluate & Archive”, Share & Disseminate”, “Access & Reuse” (figure 1.2,) ("Biomedical Data Lifecycle,"). The second half of the data lifecycle, especially the part related to data storage and data sharing, supports the reuse of biomedical data in a wide-ranging academic community. To evaluate data reusability, a concise and measurable set of principles has been proposal. These are the famous FAIR guiding principles (shortened as data FAIRness): Findable, Accessible, Interoperable, and Reusable (figure 1.3), which good data management aims to achieve (Wilkinson et al., 2016).

Figure 1.2: Biomedical Data Lifecycle ("Biomedical Data Lifecycle,")

To be **Findable**:
- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be **Accessible**:
- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
  - A1.1 the protocol is open, free, and universally implementable
  - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be **Interoperable**:
- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be **Reusable**:
- R1. meta(data) are richly described with a plurality of accurate and relevant attributes
  - R1.1. (meta)data are released with a clear and accessible data usage license
  - R1.2. (meta)data are associated with detailed provenance
  - R1.3. (meta)data meet domain-relevant community standards

Figure 1.3: Illustration of data FAIR (Wilkinson et al., 2016)

Research data management plays an important role in multi-center biomedical research projects, especially in those with study subjects of human beings. In such a project, data is generated continuously from multiple institutes (i.e., enrollment or follow-up), where study execution, data analysis and quality control are considered highly strategic (Anderson et al., 2007; Johnson, Farach, Pelphrey, & Rozenblit, 2016). To achieve the strategic goal of good data management, web-based data platforms for data integration and validation are generally employed. Such a platform is described with a client-server model that contains an always-running host, called a server, handling requests from multiple "client" hosts anywhere. More specifically, the raw research data is stored at the server remotely, while researchers from different centers upload and query the data with their personal computer, say client, via the public Internet. It is actually a central database with a web interface designed to capture various research data and research-related activities and handle data queries and downloads. A web-based data platform is usually integrated with web applications supporting data communication, data harmonization, data preparation, and data documentation so that study execution is monitored in real-time and quality control is conducted automatically. As for data analysis and data sharing, with ethics and security being considered, the platform is limited open access to the public. Traditionally, such a web-based data platform is relatively expensive and is maintained by a data management team (Franklin, Guidry, & Brinkley, 2011) . Recently, some open-source tools like R have been applied to develop low-cost web applications in this domain, focusing on data sharing and secondary analysis, which helps promote data FAIR. For example, employing the R package Shiny, a web-based data capture has been developed for dementia evaluation and reporting of the University of Kansas Alzheimer's Disease Center (KU ADC) longitudinal Clinical Cohort database (Sharma

et al., 2021).

With respect to the multi-center research project that I joined, the study subjects were rats, and research data was generated at the distinct center independently via analyzing the shared biological samples. Each sub-study has its specific scientific questions and research hypotheses, the preliminary analysis of these research data is therefore performed by the researchers who designed the sub-study and generated the data. Still, promoting data management strategies that lead to FAIR data production is my role, which also defines my training spectrum and the main contribution of this project report. In this project, cohort management like continuously data integration and enrollment issues is no longer to be considered. Here, I focus on data communication and data documentation, which help promote effective communication among all members of the multi-center project. Furthermore, by integrating data from different sub-studies, a big picture of the whole project can be presented, which helps knowledge discovery and enhance reusability. The above practices can be achieved by conducting data curation to prepare reorganized datasets with rich metadata. Finally, I can build a web-based data platform for data storge and sharing, which provides a central platform for the multi-center project making data findable. In another cohort study with rodent behavior sequence data, a Shiny web application has been created as a center data platform providing detailed metadata and high-throughput data analysis, which encourages the publication of some metadata even when the data are kept private (Colomb & Winter, 2021). This earlier effort has inspired me to utilized Shiny for a web-based data platform implementation.

1.3 Research Objectives

The objectives of this project report are to make complemental effects to data

management in the above multi-center project with rodent shrapnel model. I will illustrate my contribution to conduct data curation on the currently available research data, and building a web-based data platform for data storage and data sharing, and deploy it to the public Internet so that researchers can search and download preprocessed research data and metadata through a given URL.

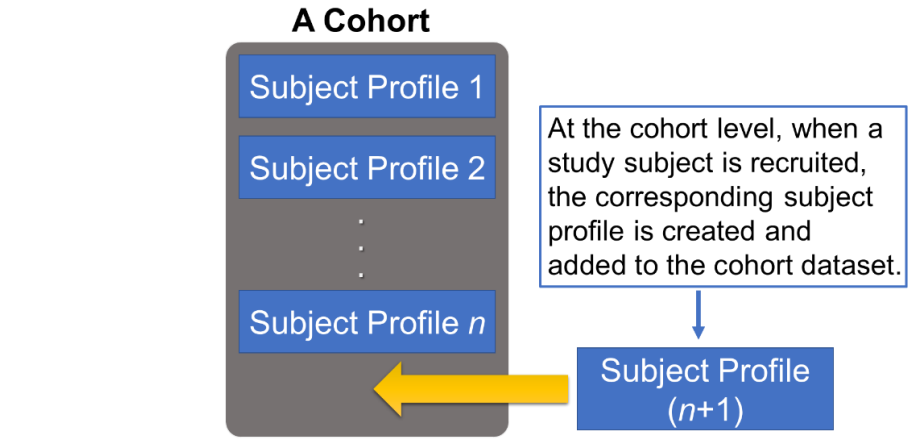
Chapter 2 Materials and Methods

2.1 Data Curation

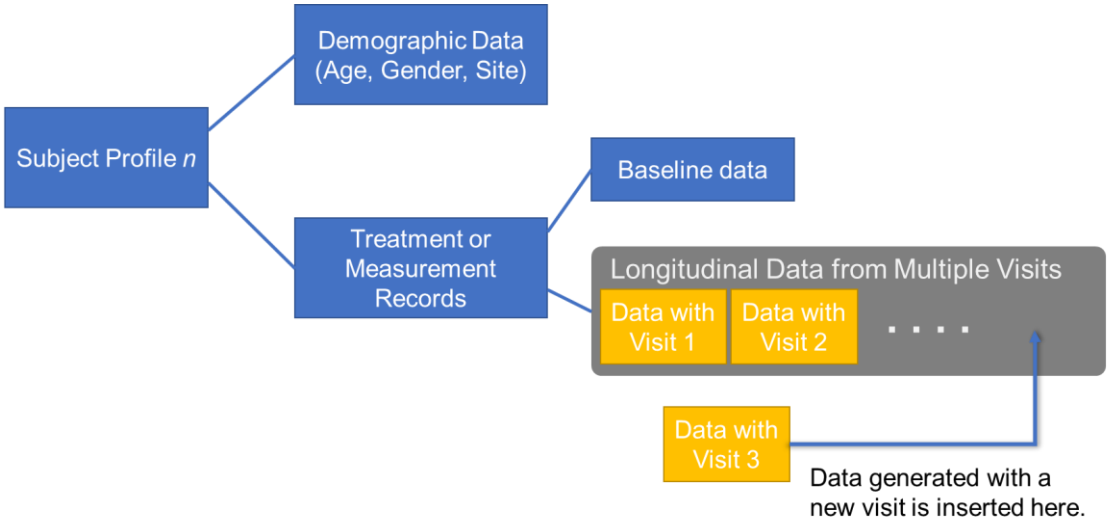
This Section presents data curation for building a web-based data platform. Although data curation is performed on currently available data, including expression profiling from skeletal muscle, serum, and urine, data related to metal concentrations from serum and urine, and body weight, my approaches could be expanded to the reorganization of coming data in our multi-center study.

2.1.1 Data Organization Manner

In a typical multi-center study with human cohorts, the data is traditionally organized in a "subject-centric manner", i.e., around a study subject on which data is collected (figure 2.1) (Das, Zijdenbos, Vins, Harlap, & Evans, 2012). As a component of the individual subjects' data, the longitudinal data is generated by tracking them over time. Additionally, the researchers would recruit subjects from multiple centers with different geographical locations to generate a study sample with similar demographic distribution to populations in a wide range of geographical regions, resulting in that the follow-up is always associated with a specific center.



(a)



(b)

Figure 2.1: Data organized in "subject-centric manner"; (a) Subject profiles organized at cohort level. (b) Data structure within a subject profile object.

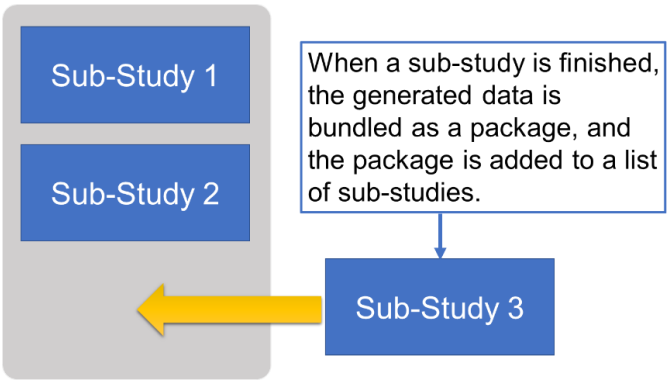
As we have seen in subsection 1.2.1, our multi-center research project is very different from the typical cohort study introduced above. On the one hand, our project is a follow-up design employing a rodent shrapnel model, where the individual subjects in the cohorts are rats instead of human beings. Researchers "followed up" a rat to evaluate the health effect of metal exposure by analyzing the biological samples from it, while a rat would never "return" for "multiple visits" once it was euthanized for harvesting biological tissues. As the rats we utilized in this project are typical model

animals, they are considered biological replicates of each other under the assumption that any individual rat shares the same demographic characteristics. Hence, for each metal group, the average values detected from rats' samples in different timed cohorts (i.e., euthanized at different designated timepoint post-implantation, or biological samples collected timepoint post-implantation) are considered as repeated observations at the individual level over designated periods of time. A cross-sectional panel can be defined given a metal group and a timed cohort. Study iterations were performed on the panels level instead of on an individual level by observing data from customized panels (i.e., a customized set of cross-sectional panels). As a result, in our case, individual rats are no longer as important as their counterparts in a typical study design with human cohorts. By contrast, the two characteristics of 1) in which metals were implanted (i.e., embedded metal exposure group) and 2) the euthanized timepoint post-implantation (i.e., timed cohort) are the most critical elements in identifying rats involved in study iterations.

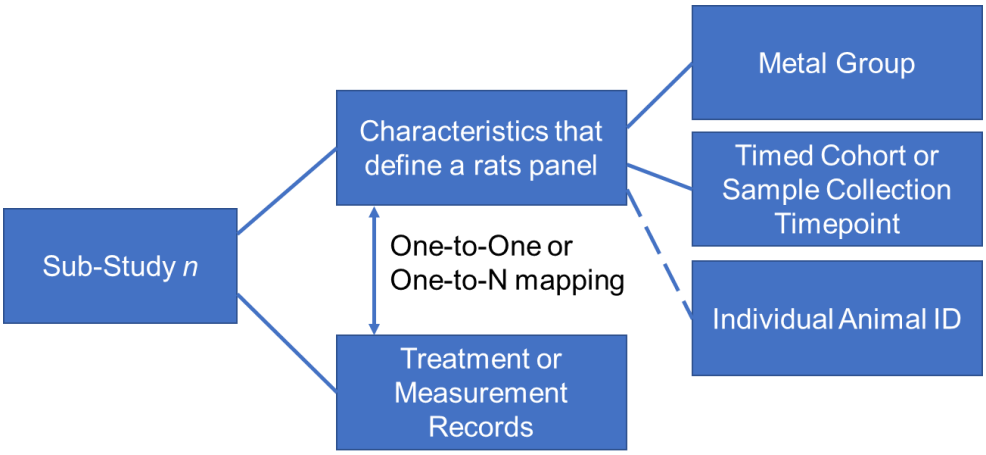
On the other hand, principal investigators from different centers contribute complementarily to this project due to their heterogeneous backgrounds. Therefore, a sub-study is usually associated with a specific center in our project, generating data with unique formats and structures. For example, the transcriptomic study was conducted at the University of Kentucky, while researchers at the AFRRI generated the metal concentrations and body weight data.

Based on the above analysis, I propose a "sub-study centered manner" to organize data around the sub-study with the rodent shrapnel model. We no longer need to create a table of individual rat profiles in this case. Instead, data from each sub-study should be saved as separate tables. The tables should offer any necessary identifying cross-sectional panels' information in the selected sub-study to enable researchers to filter

for subsets of the treatment/measurement records. Once customized panels are provided, we can filter data from different sub-studies separately and then integrate it across the sub-studies (figure 2.2).



(a)



(b)

Figure 2.2: Data organized in "sub-study centered manner"; (a) "Sub-study" data object organized at the sub-study level. (b) Data structure within a "sub-study" object.

2.1.2 Datasets Preparation

As mentioned in Chapter 1, currently available data is decomposed into three sub-study: 1) data related to expression profiling corresponds to a transcriptomic study, 2) data related to metal concentrations mapping a toxicology study, and 3) body weight records as monitoring of the rats' health status. All the available data can be

reorganized as electronic spreadsheets (i.e., data tables). Thus, it is homogeneous structured data.

Raw data, processed data, and metadata in the transcriptomic study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO) database. In this project report, we download the processed data and metadata from the GEO database, and 4 sub-datasets (i.e., GEO Series, shortened as "GSE") are included (Table 2.1). They provide processed expression profiling generated from muscle tissues, urine, and serum samples for bioinformatics downstream analysis. For each sub-datasets, a table in a wide format that reports the expression level of each gene in each sample (i.e., the gene expression matrix) is imported. Additionally, any identifying sub-study cohorts' information like "metal group" and "timed cohort" is extracted from the metadata (phenotype data) and organized in a long-format data table (i.e., a table in which columns are variables and rows are records from a sample). Thus, a total of 8 tables (2 tables per sub-datasets × 4 sub-datasets) are created for the transcriptomic study.

Table 2.1: GEO Series included in this project report

GEO Accession	Sub-Datasets Description	Number of Samples	Platform	Notes
GSE157921 (Wen et al., 2020)	mRNA array from skeletal muscle tissues	105	[Clariom_S_Rat] Affymetrix Clariom S Assay, Rat (Includes Pico Assay)	2~3 total RNA samples were pooled into one sample.
GSE157929 (Wen et al., 2020)	mRNA sequencing from skeletal muscle tissues	150	Illumina HiSeq 2000 (Rattus norvegicus)	Only 5 metal groups were included: Ni-, Pb-, Co-, DU-group, and Ta-group as control.
GSE167948 (Vechetti et al., 2021)	small RNA sequencing from urine samples	178	Illumina NextSeq 500 (Rattus norvegicus)	Urine samples collected before the metal implantations were included.
GSE168757	small RNA sequencing from serum samples	138	Illumina NextSeq 500 (Rattus norvegicus)	

Note: *The GSE168757 has not been published yet.*

The data related to metal concentrations were measured by inductively coupled plasma-mass spectroscopy (ICP–MS) from urine and serum samples, and manually recorded in the electronic laboratory notebook. Creatinine concentrations were also determined and recorded to normalize the metal concentrations data. Both the metal concentrations and creatinine data are currently organized as a long-format table, containing columns like the unique animal identification number of the rats, information on the cohorts (i.e., euthanized time-points post-implantation, in which metal was implanted), and the concentration type, and concentration values.

The data related to body weight was recorded weekly and is organized as a long-format table containing columns of the unique animal identification number of the rats, information on the cohorts, and the weighting timepoint post-implantation, which is longitudinal extensions repeatedly representing measurement, and the bodyweight

values.

2.2 Application Implementation

In this section, I act as a software developer to decompose the research objectives described in subsection 1.3 on the software design and implementation level via software requirements identification, to determine technical solutions for the application.

2.2.1 Software Requirements Identification

Software requirements identification plays a central role in software design and implementation (figure 2.3). From the perspective of software engineering, the requirements are categorized into functional and non-functional ones, which define the needs to be implemented in the software application. Developers conduct requirements analysis and determine the functional requirements, then design the technical solutions (design synthesis) according to the analysis results. System analysis, also called non-functional requirements analysis, involves all the software identification procedures.

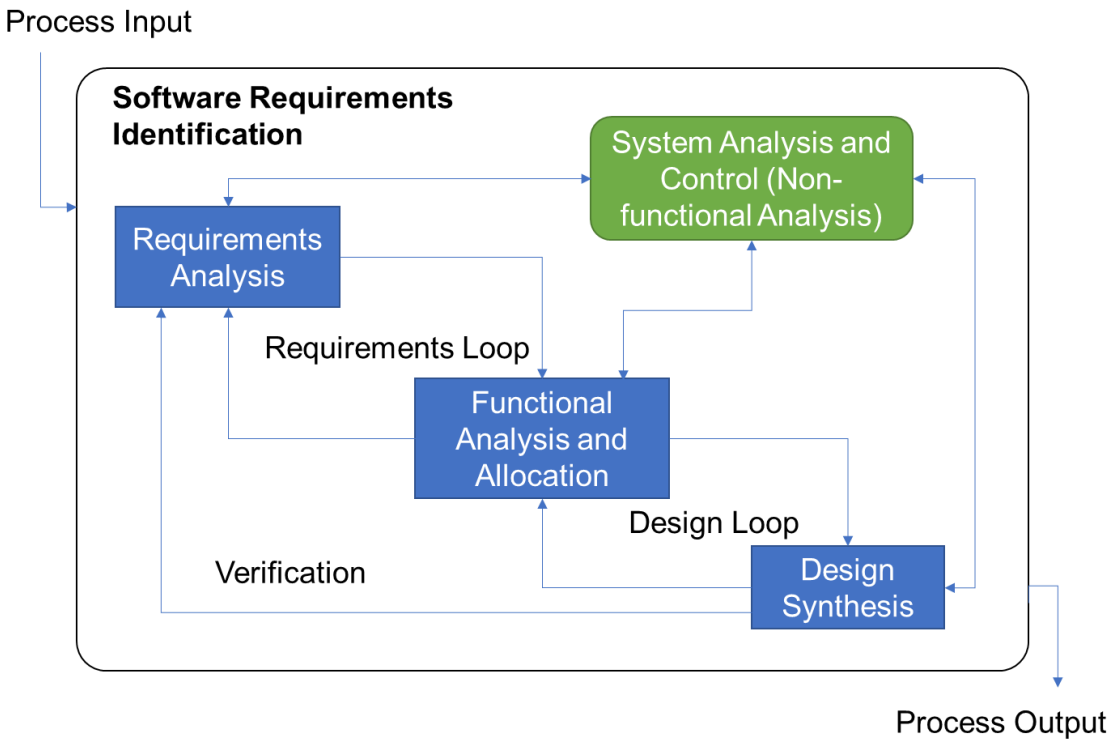


Figure 2.3: Software requirements identification from a software engineering perspective (Lightsey, 2001).

Functional Requirements

Functional requirements refer to "software must do <requirement>," specifying the results of a system. To identify functional requirements, I decompose them into user and business requirements and analyze them sequentially.

User Requirements Analysis

User requirements refer to the information and needs of the end-users. As suggested by a practical guide (Fay, Rochette, Guyader, & Girard, 2021), I try determining the needs by answering the at least 6 questions below:

1. *Who are the end-users of the application?*
   The end-users should be the researchers who want to reuse the research data for further knowledge discovery. They can be collaborators of the multi-center project, or anyone interested in the topic of embedded metal fragments.
   Needs: it is necessary to provide links to experimental design and data generation materials

and articles published earlier in the user interface.

2. *Are they tech-literate?*

   I assume that the end-users have a biomedical background but not necessarily a bioinformatics one. I assume they have no background in software development.

   Needs: The user interface should be designed self-explained, and an online introduction should be provided.

3. *In which context will they be using the app?*

   They would query and download the heterogeneous data across the sub-studies of this multi-center project via the web-based interface and integrate the data according to the customized panels offline.

   Needs: Provide online functions of query and download. If possible, the previews of the selected data tables would be helpful.

4. *On what machines (computer, tablet, smartphone, or any other device)?*

   On their personal computers, i.e., desktops or laptops.

   Needs: Front-end design should be based on the browsers on desktops.

5. *Are there any restrictions when it comes to the browser they are using? (For example, are they still using an old version of Internet Explorer?)*

   No.

6. *Will they be using the app in their office, on their phone while driving a tractor, in a plant, or while wearing a lab coat?*

   They use the web application in their lab office.

In summary, the end-users should be researchers equipped with biomedical knowledge on wet-lab or dry-lab, those interested in our multi-center research. They would like to inquire and download the research data via a desktop browser and conduct offline data analysis on metal fragments embedded, enjoying coffee in their lab office. I will detail how the end-users behave in the web application to determine the needs in business cases.

Business Requirements Analysis

Business requirements analysis aims to determine the high-level requirements taken from the business case for the software application, generating a concept map to present a big picture of the application's "who and what".

According to users' needs, the core of the business case is that users query research data through a desktop browser. More especially, a user would generate a customized

query by clicking the frontend buttons, and send it to the remote server. The server returns the selected data as a result, rendering it as data tableaus on the web interface, and download options are also provided. As discussed in Section 2.2, the dataset is organized in a “sub-study centric manner”. Thus, to generate a customized query, the user should select a sub-study at first, then generate the customized study panels of rats by determining the parameters of “metal group” and “timed cohorts”, and lastly determine the sub-study specific parameters. In addition to querying data, a user would ask for help, not because the web application is too complicated, but because he/her wants to get details of experiments and data generation for data reuse. Thus, as mentioned as one of the user needs, hyperlinks to the related materials should be prepared. As a result of the business requirements analysis, a concept map is designed (figure 2.4). We can conclude that our web application must provide functions of data query and download, hyperlinks for supplemental materials.

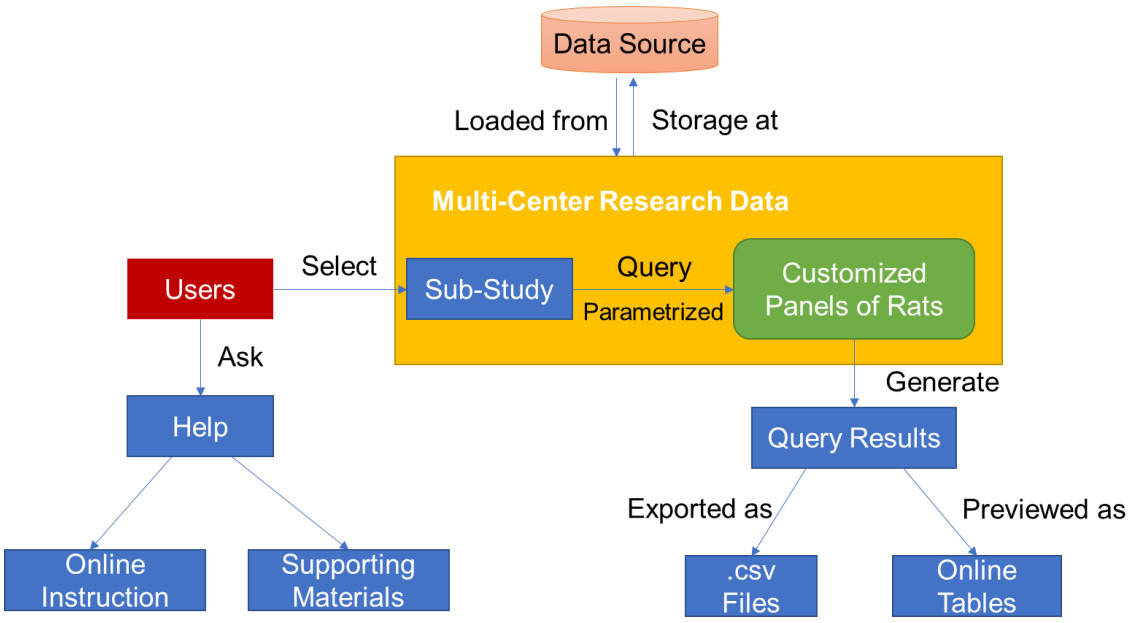


Figure 2.4: Concept map of business requirements.

Non-functional Requirements

Non-functional requirements are expressed in the form "software shall be

<requirement>" and fall under various categories, including usability, availability, reliability, scalability, performance, supportability, capacity, etc. Performing a non-functional analysis helps determine the system architecture. This subsection focuses on the requirements of supportability, which define if the app is easy to support and maintain throughout its life cycle and what kind of support it requires. In our case, both the developers and the (potential) maintainers come from a biostatistics lab. Thus, to make the application easier to be supported by other lab members (e.g., the principal investigator, the potential maintainers), R language, the most widely used programming language in biostatistics, is considered for use as the primary language in the server-side development. In addition, as the application is a part of the data management for biomedical research data, if an R-based pipeline of data analysis (e.g., bioinformatics analysis) needs to be integrated into our web application in the future, with an R-based server-side such a requirement could be implemented easily. As we have mentioned in subsection 1.2.2, the R package Shiny has been applied in building web-based data platform for data management, meanwhile, over 470 Shiny web applications have been developed for biomedical studies to now (Jia et al., 2021). Supported by the R community and the RStudio Ltd., Shiny applications can be easily deployed to the web in minutes. Due to the above considerations, our web application shall be based on the Shiny framework, and more details of the technical side are introduced in the following subsection.

2.2.2 Technical Solutions and Tools

Shiny Web Application with Golem

Shiny, an open-source R package first published in 2012, is a web framework to build applications (sometimes called Shiny web applications, or shortened as “shiny

app") that communicate with R, built in R, and working with R(Chang et al., 2021). Any Shiny web application requires a user interface (UI) and a server to form its basic architecture. The UI object controls the front-end layout and appearance of the application by compiling R code into web-friendly languages HTML, JavaScript, and CSS, while the server function contains both the business and the application logic needed to be implemented (figure 2.5).

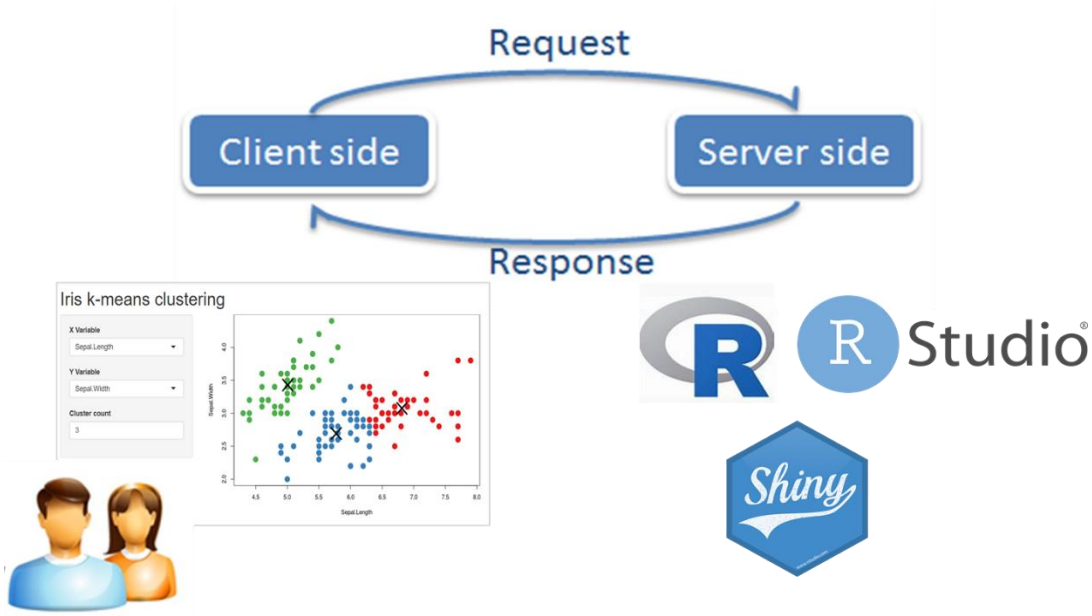


Figure 2.5: Schematic of a Shiny web application.

To build a successful Shiny web application, principles from software engineering should be followed, while R package Golem is designed to achieve that (Fay, Guyader, Rochette, & Girard, 2022). The Golem package is an optional toolkit to build Shiny applications which can be sent to production environments. As its name "Golem" implies, a developer uses the toolkit to create a "shapeless husk" of a Shiny application (i.e., a "golem" of the application, see figure 2.6) that integrates the rules and mindsets from software engineering, then "activates" the "golem" by implementing the functional needs using R programming, to build a production-ready application. A Shiny web application with Golem itself is structured as a typical R package instead of a series of files, bundling together the codebase, data source, and documentation,

and is easy to share with others.

```
golex
├── DESCRIPTION
├── NAMESPACE
├── R
│   ├── app_config.R
│   ├── app_server.R
│   ├── app_ui.R
│   └── run_app.R
├── dev
│   ├── 01_start.R
│   ├── 02_dev.R
│   ├── 03_deploy.R
│   └── run_dev.R
├── inst
│   ├── app
│   │   └── www
│   │       └── favicon.ico
│   └── golem-config.yml
└── man
    └── run_app.Rd
```

Figure 2.6: Files structure of a Shiny application's "golem".

In this project report, I implement a web application for data storage and queries employing Shiny and Golem.

System Architecture

The system architecture of our application is modeled as distinct layers (figure 2.7). Within the user-interface layer, each sub-study is prepared with a single front-end webpage, where an end-user can query and download data from the sub-study within the webpage and send requests to the remote server-side. The server layer hosts the R functions to respond to the requests from the end-users, mainly to conduct data queries from the data source and send the results to the front-end. The storage layer includes a

static storage component (i.e., the .rda files) in which the curation datasets described in Section 2.1 are bundled, hosting a "sub-study centered manner" structure of data tables within, to act as the application data source. The server layer and the data storage layer are hosted together via a server remotely, and the combination of them is called "server-side".

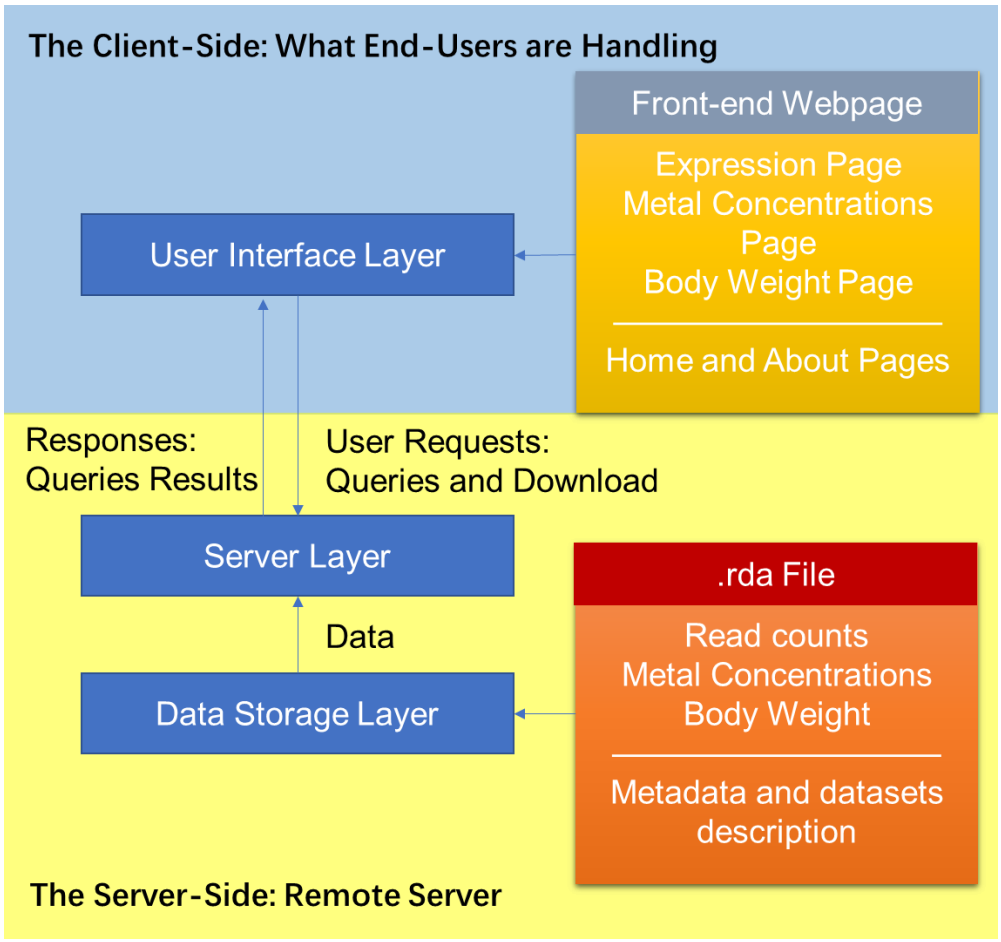


Figure 2.7: System architecture of our application.

Deployment and Remote Server

In general, a web application should be deployed to a remote server, setting up with a URL, so that end-users from anywhere could access the website via the internet. Thus, application deployment involves configuring a remote server. Luckily, the RStudio Limited provides an easy-to-use hosting server at a shinyapps.io platform

(https://www.shinyapps.io/#), which is specifically designed for deploying Shiny applications to the web. At this phase of the study, I deploy our application on the shinyapps.io platform, so I have no need to own a remote server and configure it.

Other Dependencies and Tools

In addition to Shiny and Golem, some other R packages are also imported into the web application, especially to improve the front-end performance. The reactable package (Lin, 2021) is used for interacting data tables for the front-end. The shinyjs package (Attali, 2021) performs common useful JavaScript operations in the Shiny application. The Bootstrap theme "paper" is applied as the front-end theme via the shinythemes package (Chang, 2021).

The version of the development language is R 4.0.3, and RStudio is utilized as the IDE. Git is employed to perform version control, integrating with the Github platform for source code management and software distribution.

Chapter 3 Results

3.1 Software Availability

The Shiny web application has been deployed on the public Internet, and one can access the website through:

https://fhs1027.shinyapps.io/MetalEmbeddedStudies/.

Additionally, our web application is open source under the MIT license. The source code, as well as the curation datasets, have been uploaded to a repository on GitHub platform with the URL of:

https://github.com/MuscleOne/MetalEmbeddedStudies/tree/dev.

By downloading the source code from GitHub, anyone equipped with basic knowledge of R and RStudio could run the application locally following the instruction provided online.

3.2 Front-End Views

From the perspective of an end-user, he/her will see a homepage as soon as he/her visits the website (figure 3.1). The homepage has a navigation bar, welcome words, and a brief introduction to our multi-center research with the rodent shrapnel model. Most importantly, tabs representing "sub-studies" are set to a prominent position so that the end-user can see the available sub-studies with available data at a glance. Each "sub-study" tab corresponds to a front-end webpage. By clicking a specific tab, one can jump to a webpage associated with the particular sub-study for data query, preview, and download.

As we mentioned in subsection 2.1.2, our data from different sub-studies is

organized as homogeneous structured data. Therefore, in webpage design on the front-end, the webpages associated with each sub-study share a similar webpage structure. We take the "Body Weight" webpage as an example to explain the webpage structure. As shown in figure 3.2, the webpage can be divided into four zones. A zone is a tab panel that acts as a filter. An end-user could generate customized queries by clicking the select input and send them to the back-end server. B zone provides the function of online data tables preview. Once the customized query is sent, the data tables will react accordingly within a second returning a reactive online preview as query results. C zone is a brief introduction to the sub-study data. A download button is located at the D zone, an end-user can download the query results, a customized spreadsheet returned by the server, by clicking the download button.

In addition to the Homepage and the webpages associated with the sub-studies, an About page is also designed to provide more details of this multi-center research, such as the introduction of the research teams from various centers and their contributions, and the earlier published high-quality research articles based on this project, etc.

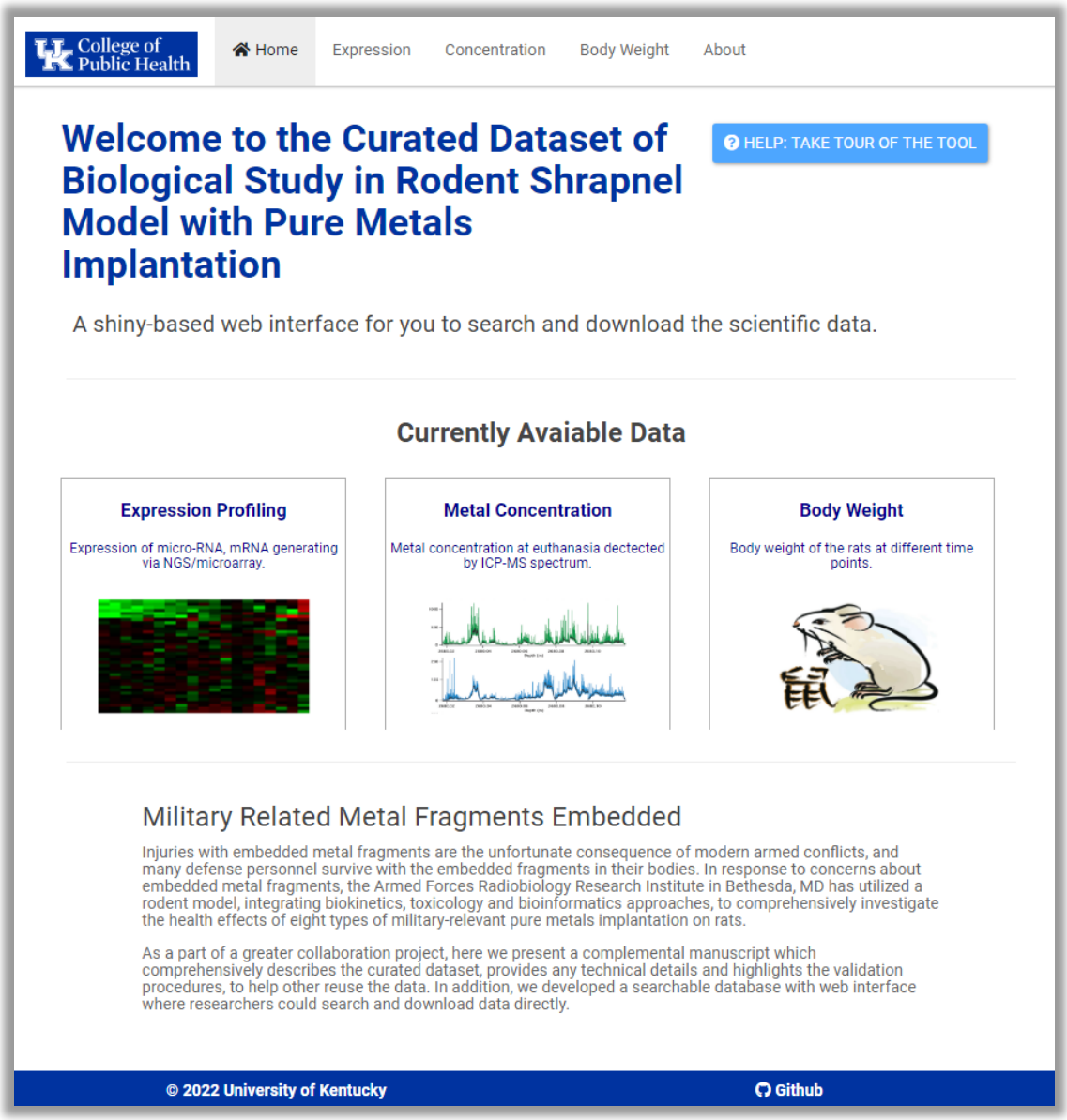


Figure 3.1: Homepage of the website.

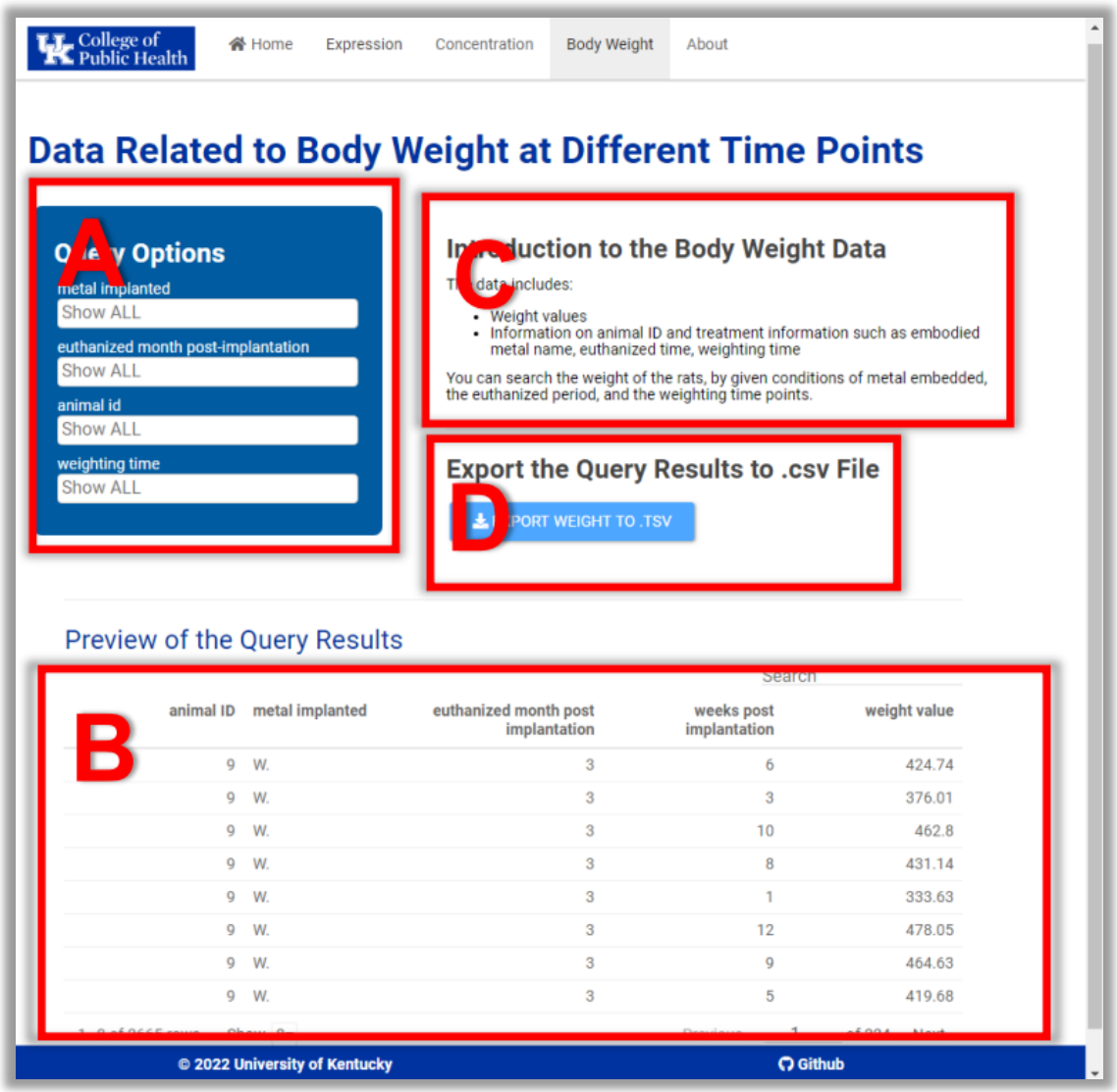


Figure 3.2: “Body Weight” page of the website with four functional zones (i.e., the A, B, C, D zone).

Chapter 4 Discussion and Conclusion

4.1 Data FAIRness Optimization

As introduced in Chapter 1, data FAIRness describes the level of findability, accessibility, interoperability, and reusability in research data, which the academic community has widely accepted as practical guiding principles for evaluating data management performance, especially in the stages of data storage and data sharing. This Section discusses how my complemental efforts, conducting data curation and building a web-based data platform, could optimize the data FIARness, bridging the gap of good data management for this multi-center biological research project.

For data findability, I provide a centralized and open access data platform for this multi-center research for the first time, hosting the curated datasets with rich metadata. Through this platform, researchers can find a big picture directly of this multi-center research project, instead of fitting together the pieces in separate publications, so as to promote the findability of the research and the research data. In the future, except that high-throughput data is uploaded to the GEO database as the general practice in the bioinformatics domain, all low-throughput biological data, as well as the metadata, will be deposited into the FigShare database, and all data will be assigned globally unique and persistent identifiers.

As for issues of accessibility, this work, and the research data we included in this work, have been or plan to be submitted to peer-reviewed journals for publication, so my efforts are expected to be retrievable.

For interoperable, our web-based data platform, as a centralized platform, provides researchers with data documents based on consistent terms. Moreover, the

data it provides is in a structured and analysis-ready form, which is suitable for manipulation with common data analysis software.

Finally, data reuse is also promoted by preparing a complete data descriptor as a part of the online data documents, emphasizing scientific value and potential for reuse. Through the descriptor, researchers can obtain any details of the whole experimental design and data generation, while the relationship between each sub-study is clearly presented. Therefore, data from different sub-studies but related to each other can be integrated for secondary analysis to promote knowledge discovery. Additionally, as the raw data, especially the low-throughput experimental data, is provided, the downstream usage is no longer limited. Currently, a biostatistics team from the University of Kentucky is conducting a time-course analysis, integrating sequencing profiling of kidney tissues and urine samples, and the low-throughput metal concentrations of urine, for investigating the effects of metal implantation on kidney and metabolic system in rats.

4.2 Advantages with Golem Toolkit

In Section 2.2, I mention that utilizing the R package Golem helps build a successful Shiny application. This Section details the benefits of using Golem in the web-based data platform, and further analyzes how Golem enhances the reliability and supportability of our software.

Building an application with Golem generates an application's "golem" at the beginning phase as a template for the application's engineering structure. Firstly, the application built based on the Golem-generated template is built into a typical R package, so it has all the advantages of an R package, including but not limited to: splitting the application into small functions to decompose the codebase into small

pieces; building a NAMESPACE file for handling all dependencies of the application; generating documents and rich metadata of the application to promote FAIRness. Secondly, the template encourages us to build modules that are sub-applications made up of a series of R functions. Thus, functional needs are decomposed, and the business-logic functions can also be separated from the interactive-logic functions. In a Shiny application with Golem, modules are divided into two categories. One of them is reused through the application. In our case, the module implementing the download function and the module previewing query results belong to this category, as they are called on the separate webpages of different sub-studies. Another is modules for more specific requirements, such as processing the input conditions of a "sub-study" queries and generating customized queries sent to the data source. In this situation, as the data structures of the distinct sub-studies are different, the targeted design of modules is unavoidable, and a larger module often includes some smaller modules as building elements. Thirdly, this template provides a framework for software testing and encourages us to test the small pieces of functions and modules. In software engineering, any application should be tested before being deployed in a production environment. However, testing is lacking in many prior published shiny applications in the biomedical domain.

Reliability, as mentioned in prior non-functional requirements analysis, is a term in software engineering which defines the probability of failure-free software operation for a period of time and involves the complexity and codebase quality in practice. A Shiny application with Golem can reduce the system complexity by decomposing the application into modules. Besides, formal testing avoids solving coding problems with temporary approaches, resulting in good quality codebase reducing possible problems in the subsequent integration. These two features improve

the reliability level of our application.

By employing modules and testing, it would be easier to implement new business needs (e.g., integrating datasets from a new sub-study) in our web-based data platform in the future, promoting supportability with respect to continuous integration. The Shiny application itself can be considered the largest module, consisting of at least three smaller modules (called sub-modules) corresponding to the three "sub-studies" available now. By building a new module, I can integrate the data of the new sub-study into the application without modifying the existing tested sub-modules. Ideally, a collaborator as a developer does not need to read a large volume of the existing code when he/she integrates new functions in the application.

4.3 Summary and Future Work

Data management, involving all stages of a study with a final objective to promote data FAIR, is acknowledged as a critical component in any biomedical investigations, especially with a multi-center research collaboration which may have multiple sub-studies from different institutes due to research considerations. In this project report, I participated in collaborative data management efforts for a multi-center research project with a rodent shrapnel model. I mainly contribute to conducting data curation, including but not limited to reorganizing the preprocessed data and preparing data documents to highlight the scientific values. Moreover, a production-grade web-based data platform powered by R Shiny and Golem is built for data storage and sharing. Such a Shiny application has been deployed on the public Internet so that anyone can access it to search and download the research data. All my work has finally promoted data management, especially data FAIRness.

In the future, I will continue to integrate the latest generated research data in this

multi-center research into the curated datasets and our web-based data platform. After all the sub-studies have been completed, a comprehensive manuscript of data descriptor, along with the description of this data platform, will be written and submitted to a peer-reviewed journal for publication. All the high-throughput data would be published on the GEO database, while the rest would be deposited into the FigShare database as open-access data, further facilitating data findability.

Reference

Anderson, N. R., Lee, E. S., Brockenbrough, J. S., Minie, M. E., Fuller, S., Brinkley, J., & Tarczy-Hornoch, P. (2007). Issues in Biomedical Research Data Management and Analysis: Needs and Barriers. *Journal of the American Medical Informatics Association, 14*(4), 478-488. doi:10.1197/jamia.m2114

Attali, D. (2021). shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds.

Bardack, S., Dalgard, C., Kalinich, J., & Kasper, C. (2014). Genotoxic Changes to Rodent Cells Exposed in Vitro to Tungsten, Nickel, Cobalt and Iron. *International Journal of Environmental Research and Public Health, 11*(3), 2922-2940. doi:10.3390/ijerph110302922

Biomedical Data Lifecycle. Retrieved from https://datamanagement.hms.harvard.edu/about/what-research-data-management/biomedical-data-lifecycle

Castro, C., Benson, K., Bogo, V., Daxon, E., & Hogan, J. (1996). *Establishment of an Animal Model to Evaluate the Biological Effects of Intramuscularly Embedded Depleted Uranium Fragments*. Retrieved from

Chang, W. (2021). shinythemes: Themes for Shiny.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., . . . Borges, B. (2021). shiny: Web Application Framework for R.

Colomb, J., & Winter, Y. (2021). Creating Detailed Metadata for an R Shiny Analysis of Rodent Behavior Sequence Data Detected Along One Light-Dark Cycle. *Frontiers in neuroscience, 15*, 742652-742652. doi:10.3389/fnins.2021.742652

Das, S., Zijdenbos, A., Vins, D., Harlap, J., & Evans, A. (2012). LORIS: a web-based data management system for multi-center studies. *5*. doi:10.3389/fninf.2011.00037

Emond, C. A., & Kalinich, J. F. (2012). Biokinetics of embedded surrogate radiological dispersal device material. *Health Phys, 102*(2), 124-136. doi:10.1097/HP.0b013e31823095e5

Fay, C., Guyader, V., Rochette, S., & Girard, C. (2022). golem: A Framework for Robust Shiny Applications.

Fay, C., Rochette, S., Guyader, V., & Girard, C. (2021). *Engineering Production-Grade Shiny Apps*: Chapman and Hall/CRC.

Fitsanakis, V. A., Erikson, K. M., Garcia, S. J., Evje, L., Syversen, T., & Aschner, M. (2006). Brain Accumulation of Depleted Uranium in Rats Following 3- or 6-Month Treatment With Implanted Depleted Uranium Pellets. *Biological Trace Element Research, 111*(1-3), 185-198. doi:10.1385/bter:111:1:185

Franklin, J. D., Guidry, A., & Brinkley, J. F. (2011). A partnership approach for Electronic Data Capture in small-scale clinical trials. *Journal of Biomedical Informatics, 44*, S103-S108. doi:10.1016/j.jbi.2011.05.008

Hahn, F. F., Guilmette, R. A., & Hoover, M. D. (2002). Implanted depleted uranium fragments cause soft tissue sarcomas in the muscles of rats. *Environmental health perspectives, 110*(1), 51-59. doi:10.1289/ehp.0211051

Hoffman, J. F., Vergara, V. B., Fan, A. X., & Kalinich, J. F. (2021). Effect of embedded metal fragments on urinary metal levels and kidney biomarkers in the Sprague-Dawley rat. *Toxicology Reports, 8*, 463-480. doi:10.1016/j.toxrep.2021.02.023

Jia, L., Yao, W., Jiang, Y., Li, Y., Wang, Z., Li, H., . . . Zhang, H. (2021). Development of interactive biological web applications with R/Shiny. *Briefings in Bioinformatics, 23*(1). doi:10.1093/bib/bbab415

Johnson, S. B., Farach, F. J., Pelphrey, K., & Rozenblit, L. (2016). Data management in clinical research: Synthesizing stakeholder perspectives. *Journal of Biomedical Informatics, 60*, 286-293. doi:10.1016/j.jbi.2016.02.014

Kalinich, J. F., Emond, C. A., Dalton, T. K., Mog, S. R., Coleman, G. D., Kordell, J. E., . . . McClain, D. E. (2005). Embedded weapons-grade tungsten alloy shrapnel rapidly induces metastatic high-grade rhabdomyosarcomas in F344 rats. *Environmental*

health perspectives, 113(6), 729-734. doi:10.1289/ehp.7791
Kalinich, J. F., & Kasper, C. E. (2016). Are Internalized Metals a Long-term Health Hazard for Military Veterans? *Public Health Reports, 131*(6), 831-833. doi:10.1177/0033354916669324
Lightsey, B. (2001). *Systems engineering fundamentals*. Retrieved from
Lin, G. (2021). reactable: Interactive Data Tables Based on 'React Table'.
Pellmar, T. (1999). Distribution of uranium in rats implanted with depleted uranium pellets. *Toxicological Sciences, 49*(1), 29-39. doi:10.1093/toxsci/49.1.29
Sharma, P., Montgomery, R. N., Graves, R. S., Meyer, K., Hunt, S. L., Vidoni, E. D., . . . Mudaranthakam, D. P. (2021). CONSENSUS: a Shiny application of dementia evaluation and reporting for the KU ADC longitudinal Clinical Cohort database. *JAMIA Open, 4*(3). doi:10.1093/jamiaopen/ooab060
Vechetti, I. J., Wen, Y., Hoffman, J. F., Alimov, A. P., Vergara, V. B., Kalinich, J. F., . . . Peterson, C. A. (2021). Urine miRNAs as potential biomarkers for systemic reactions induced by exposure to embedded metal. *Biomarkers in Medicine, 15*(15), 1397-1410. doi:10.2217/bmm-2021-0120
Wen, Y., Vechetti, I. J., Alimov, A. P., Hoffman, J. F., Vergara, V. B., Kalinich, J. F., . . . Peterson, C. A. (2020). Time-course analysis of the effect of embedded metal on skeletal muscle gene expression. *Physiological Genomics, 52*(12), 575-587. doi:10.1152/physiolgenomics.00096.2020
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. doi:10.1038/sdata.2016.18
Yang, Q., & Yang, F. (2022). Biomedical Data. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 116-118). Cham: Springer International Publishing.