Conceptual Design and Prototyping for a Primate Health History Model

Martin Q. Zhao [1,*], Elizabeth Maldonado [2], Terry B. Kensler [2], Luci A.P. Kohn [3], Debbie Guatelli-Steinberg [4], Qian Wang [5, *]

1. Department of Computer Science, Mercer University, Macon, GA
2. Caribbean Primate Research Center, University of Puerto Rico Medical Sciences Campus, San Juan, PR
3. Department of Biological Sciences, Southern Illinois University Edwardsville, Edwardsville, IL
4. Department of Anthropology, The Ohio State University, Columbus, OH
5. Department of Biomedical Sciences, Texas A&M University College of Dentistry, Dallas, TX

Page: 10

Table: 0

Figure: 7 (All in color)

SHORT TITLE: Designing a Primate Health History Knowledge Model

KEY WORDS: Cayo Santiago, rhesus macaques, demography, graphical user interfaces, relational database design

[*] Correspondence to: Qian Wang, Ph.D., Department of Biomedical Sciences, Texas A&M University College of Dentistry, 3302 Gaston Ave, Dallas, TX 75246. Tel: 214-370-7002; Fax: 214-874-4835; Email: qian.wang@tamu.edu; and to Martin Q. Zhao, Ph.D., Department of Computer Science, Mercer University, 1501 Mercer University Drive, Macon, GA 31207. Tel: 478-301-2425; Fax: 478-301-2276; Email: zhao_mq@mercer.edu.

SUMMARY

Primate models are important for understanding human conditions, especially in studies of ageing, pathology, adaptation, and evolution. However, how to integrate data from multiple disciplines and render them compatible with each other for datamining and in-depth study is always challenging. In a long-term project, we have started a collaborative research endeavor to examine the health history of a free-ranging rhesus macaque colony at Cayo Santiago, and build a knowledge model for anthropological and biomedical/translational studies of the effects of environment and genetics on bone development, aging, and pathologies. This paper discusses the conceptual design as well as the prototyping of this model and related graphical user interfaces, and how these will help future scientific queries and studies.

INTRODUCTION

In 1938, a group of rhesus macaques (*Macaca mulatta*) from India were introduced to Cayo Santiago (CS), Puerto Rico to insure a steady supply for research and vaccine development in the continental U.S. during WWII (Sade et al., 1985; Rawlins and Kessler, 1986; Kessler,1989; Turnquist and Hong, 1989; Wang et al., 2006a,b, 2007, 2016a,b; Dunbar, 2012; Wang, 2012; Kessler and Rawlins, 2016; Kessler et al., 2016; Li et al., 2018).

Systematic daily tracking of all rhesus monkeys on the island began in 1956 under NIH's Laboratory of Perinatal Physiology (LPP) in San Juan. The LPP closed in 1970, and the Caribbean Primate Research Center (CPRC) was established under the University of Puerto Rico (UPR) School of Medicine with base support coming from NIH. The daily census, begun in 1956, has continued uninterrupted to the present day. The colony has naturally divided to more than 26 matrilineal families. The data collected during the past 64 years for over 10 generations of monkeys makes the rhesus colony at Cayo Santiago one of the most useful primate database in biomedical and anthropological research. In addition, in 1971, the CPRC rhesus monkey skeletal collection was established and at present, up to eight generations are in the collection. This is a unique translational resource for genetic and age-related studies: ancestors of non-human primates available in a skeletal collection plus their descendants living in similar conditions (Dunbar, 2012; Wang, 2012; Kessler and Rawlins, 2016). However, there is no integrated database on the Cayo rhesus colony and the derived skeletal collection, limiting the use of this rhesus resource for the reconstruction of the health history of the colony for the purpose of biomedical and anthropological studies. The need to integrate data warrants the construction of a complete demographic profile of the colony at Cayo Santiago.

We have started a collaborative research project with a group of scientists in four universities in the US involved to carry out the studies. In our long-term project, there are three aims toward building a searchable database of Cayo Santiago monkey health history for anthropological and biomedical/translational studies of the effects of environment and genetics on bone development, aging, and pathologies.

1. *Document morphological and pathological conditions of the Cayo Santiago skeletal collection.*
2. *Build a Cayo Santiago rhesus health database.*
3. *Test hypotheses about secular trends and familial disparities in health and other features using the Cayo Santiago rhesus health database.*

This paper discusses the conceptual design and prototyping of this database (DB) and related graphical user interfaces (GUI) to include all necessary information and facilitate searchable outputs, while allowing the continuing input of information in the future, similar to a knowledge model (Zhao, 2010, 2012). Meanwhile, it must be pointed out that, *like any database of human subjects, and following regulations and requirements set by Caribbean Primate Research Center, this project treats every monkey individual as a patient and thus protects its privacy as we practice with human patients*.

DATA NEEDS AND CONCEPTUAL DATA MODELS

Building the proposed database and related GUIs need to have thorough plans that addresses all phases of the application development lifecycle (Zhao, 2018), as illustrated in Fig 1. In this early stage of the development process, the key activity is to collect and analyze requirements for the proposed system and come up with conceptual design models. In this section, we will discuss data needs and conceptual data models. Functional requirements in terms of use cases and GUI design concepts will be discussed in the next section.

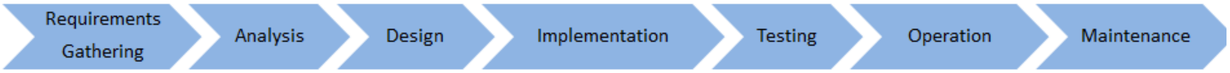


Figure 1 Illustration of the Application Development Lifecycle (ADLC)

In our planned project, individual skeletal remains will be screened and measured for documenting morphological and pathological conditions for a wide spectrum of bone and tooth health and pathology. Five sets of data will be collected for each skeleton: [1] The demographic and genealogical information of all specimens, and body mass and sitting height when available; [2] The bone conditions of all available skeletal parts, including age-related and pathological features (such as abnormalities, diseases, and trauma) and non-metric harmless bone features (such as supernumerary teeth, suture type at the pterion, hyperostosis, etc.). Color images of morphological and pathological features of interest will be generated using a high-quality camera Cannon EOS-50D; [3] Bone density will be measured by a portable OmniSense 8000S Mobile Sonometer Bone Densitometry System. Specimens of special interest will be further examined using X-ray or quantitative CT scanning facilities; [4] Measurements of skeletal size of both cranial and postcranial skeletons (Wang et al., 2007; Kohn and Bledsoe, 2012) ; and [5] linear enamel hypoplasia (developmental defects of enamel) from dental replicas under a Leica DMS 1000 digital monocular microscope and then with a measuring microscope (Spectra Services) and VisionGauge software to measure perikymata (growth increment) spacing. For data collection, an interactive bone survey program will be generated using Qualtrics, a web-based data collection tool that is secure and Texas A&M University has a license for all faculty and staff to use.

The proposed integrative database will incorporate health data obtained in this project (scans, measurements, and observation data) with subject genealogy information of the rhesus families maintained by the CPRC. A standard relational data model will be used to provide a normalize the database (Fig. 2).

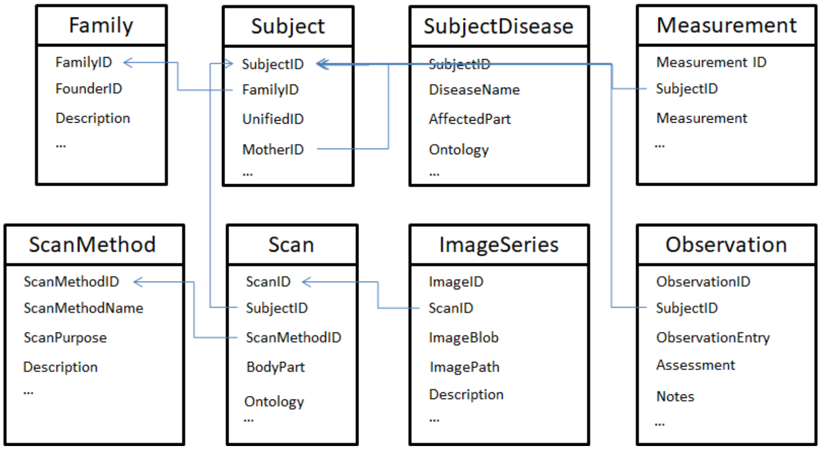


Figure 2 Conceptual Data Model for the Cayo Santiago Rhesus Health Database. Per regulations and requirements set by Caribbean Primate Research Center, this project treats every monkey individual as a patient and thus protects its privacy as we practice with human patients. All assigned IDs will be a coded ID, not original tattoos.

Originally, the genealogical data are recorded using Excel files, each for a different family. In these files, each row keeps track of a leaf node (i.e., a subject with no descendant) in the respective family tree, including information of its own (gender, birth and death dates) and information about its female ancestors all the way back to the family founder.

To remove the redundancy in the original dataset is to split it into two tables, one for family and the other for subjects. The Subject table with one row for each subject, which is related to the corresponding family through a foreign key (FK) FamilyID to the primary key (PK) in the Family table. Each subject entry is related to its mother through another foreign key MotherID, which is the mother's SubjectID, to keep mother-child mapping.

Additional tables will be used to store subject morphology and pathology information (such as bone density, disease, and image) when they become available. Similar relational design has been used (Seo et al., 2013) and shows great extensibility. More tables may be added as needed to demonstrate a high-level abstraction of the database schema that captures the major data sets (i.e., database tables) and the relationships among them (Figure 3). As will be discussed in the following section, other kinds of information need to be stored to keep track valid users, access control, user activities, etc. Additional tables will be added to in later stages (such as detailed design phase) to store those kinds of information.

Unified coding scheme

The unified code includes all the information regarding the subject with a FI-GE-MSEQ-SS-G-BY-DY-SSEQ pattern, which is detailed in in the unified coding scheme. When used in various length (including certain parts in the multi-part pattern), it can present data needed in various scenarios. For instance, FI (family ID) and SSEQ (subject sequence number within the family) can uniquely identify a subject; FI, MSEQ (mother sequence number) and SS (sibling sequence number) can also identify a subject and focuses on a subject's social status.

The FI-GE-MSEQ-SS-G-BY-DY-SSEQ pattern consists of the following parts:

- FI: two-digit family ID.
- GE: two-digit generation number within a family, with 01 for the family founder.
- MSEQ: four-digit subject sequence number within family for the mother of this subject; 0000 is used for family founder, whose mother is not in this database.
- SS: sibling sequence number for direct children from the same mother subject.
- G: one-character code for individual's sex, with valid values 'f', 'm', and 'u'.
- BY: Exact birth date with four-digit birth year, two-digit birth month, and two-digit birth day.
- DY: Exact death date with four-digit death year, two-digit death month, and two-digit death day.
- 'RR' and '..' used for subjects that are removed and still alive, respectively, with exact date of removal.
- CPRC-LPM-CM: Current CPRC LPM's skeletal catalog number assigned to each skeleton.
- SSEQ: four-digit subject sequence number within family.

A segment of the family tree representing subjects in family 11 (an arbitrary number for now) using the unified code is given in Fig. 3.

A UnifiedCode column is added to the Subject table to protect subject's privacy. On the other hand, ontology links are added in certain tables to be compliant with medical informatics standards. To be specific, Uberon (Mungall et al, 2012) numbers will be added for diseases (stored in the SubjectDisease table) and body part (in the Scan table).

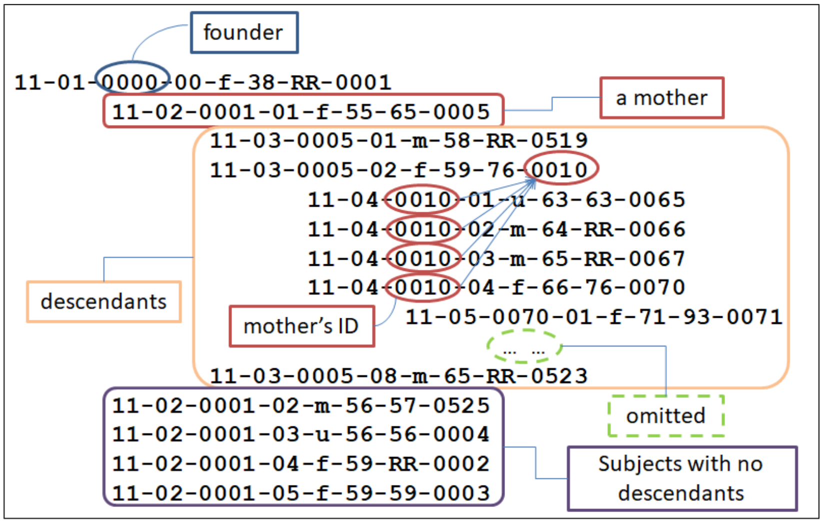

founder
11-01-0000-00-f-38-RR-0001
11-02-0001-01-f-55-65-0005
a mother
11-03-0005-01-m-58-RR-0519
11-03-0005-02-f-59-76-0010
11-04-0010-01-u-63-63-0065
11-04-0010-02-m-64-RR-0066
11-04-0010-03-m-65-RR-0067
11-04-0010-04-f-66-76-0070
descendants
mother's ID
11-05-0070-01-f-71-93-0071
… …
11-03-0005-08-m-65-RR-0523
11-02-0001-02-m-56-57-0525
11-02-0001-03-u-56-56-0004
11-02-0001-04-f-59-RR-0002
11-02-0001-05-f-59-59-0003
omitted
Subjects with no descendants

Figure 3 A Partial Family Tree with Subjects Represented in the FI-GE-MSEQ-SS-G-BY-DY-SSEQ Pattern

USE CASES & APPLICATION DESIGN CONCEPTS

Convenient user interfaces need to be developed to support various kinds of users to use the database we are building. Users can be categorized into the following types: (1) staff in charge of collecting and entering data, (2) researchers involved in this project using the data to "*test hypotheses on secular trends and familial disparities*", and (3) general public in research communities searching for related information to support their research.

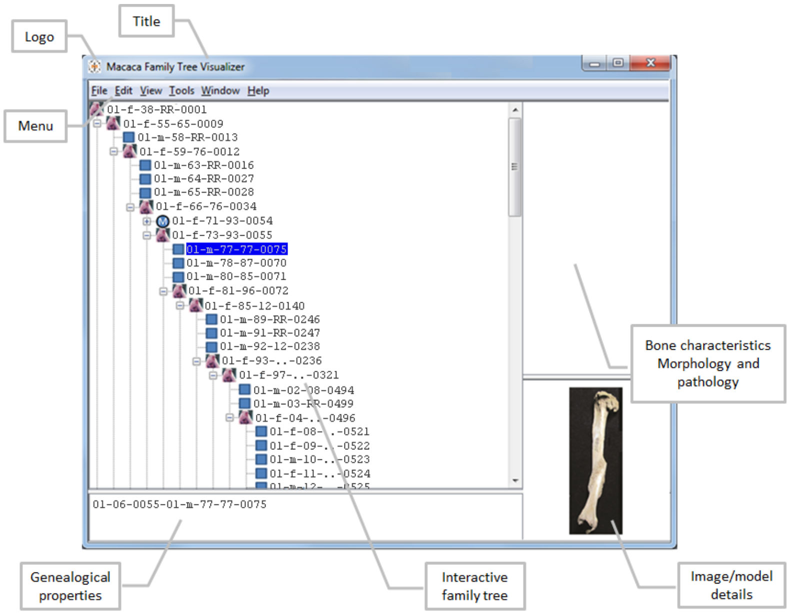
While the three groups may overlap with one another, the first two groups of users are closely related to the collaborative research project. It is appropriate to develop two sets of interfaces: a window-based GUI app for staff operators/researchers to use, such as for loading and editing data, and for data visualization and manipulation; and a web-based application for general research communities to access data and use in their studies.

The web-based application is designed to be a searchable and computer-interoperable knowledge model to discover previously unknown associations from the Rhesus family data. It will be developed in the last stage of the multi-year project when the database is constructed and loaded with newly collected data. In this paper, our focus will be on the window-based interfaces for staff and project researchers.

Like many information systems, this proposed system will need to support the following use cases:

- User authentication and role-based accessibility control: Only authorized users will be able to log into the system with valid credentials. Based on their job functions (or roles), they will be able to access (search only or manipulate) the right types of data (e.g., family and/or pathology, etc.).
- Data entry and editing: The system will provide support in several different approaches.
    - Converting existing datasets (Excel spreadsheets) used to track subject and family information, and automatically load them into the new relational database.
    - Providing convenient forms and/or dialog boxes to allow for manual data entry and editing, as well as data quality checking.

    - Integrating with other specialized data collection tools such as Qualtrics to load data into targeted tables.
- Provide a comprehensive window (Fig. 4) with multiple panels or tabs to support various kinds of data visualization and analytics tasks. This window will include
    - A main panel family tree that displays an interactive family tree with all subjects from the selected family displayed, and indicate as a mother, a male, female, or unknown gender subject with no descendant.
    - A side panel displaying morphological and pathological data in tabular form for an individual selected in the family tree.
    - Separate panels will display scan images for the selected subject when available.
    - Other panels displaying additional notes/descriptions
- The comprehensive window will anchor menus or tabs to allow insider researchers to conduct searching, filtering, and analytical tasks, such as
    - Displaying distribution and trends with common charts (e.g. histogram, scatter plots).
    - Partial family tree with filtering criteria in place and/or with father information available from DNA to be conducted in this project.


Logo
Title
Macaca Family Tree Visualizer
Menu
File Edit View Tools Window Help
Bone characteristics Morphology and pathology
Genealogical properties
Interactive family tree
Image/model details

```
01-f-38-RR-0001
⊟ 01-f-55-65-0009
    01-m-58-RR-0013
  ⊟ 01-f-59-76-0012
      01-m-63-RR-0016
      01-m-64-RR-0027
      01-m-65-RR-0028
    ⊟ 01-f-66-76-0034
      ⊞ 01-f-71-93-0054
      ⊟ 01-f-73-93-0055
          01-m-77-77-0075
          01-m-78-87-0070
          01-m-80-85-0071
        ⊟ 01-f-81-96-0072
          ⊟ 01-f-85-12-0140
              01-m-89-RR-0246
              01-m-91-RR-0247
              01-m-92-12-0238
            ⊟ 01-f-93-..-0236
              ⊟ 01-f-97-..-0321
                  01-m-02-08-0494
                  01-m-03-RR-0499
                ⊟ 01-f-04-..-0496
                    01-f-08-..-0521
                    01-f-09-..-0522
                    01-m-10-..-0523
                    01-f-11-..-0524
                    01-m-12-..-0525
```

```
01-06-0055-01-m-77-77-0075
```

Figure 4 Illustration of the Layout of the Window-based User Interface

DATABASE & APPLICATION PROTOTYPING

Efforts to parse subject information have been tested in three families stored in Excel spreadsheets, which was by prepared by E.M., and released by CPRC to Q.W. at the early stage of preparing for this project. Original data for each subject includes a unique code (or ID), sex, years born and died, and the mother's code. Additional data derived from the dataset include generation within the family, family number, sibling sequence number, and each subject's life span. As mentioned above, a unified coding scheme is proposed to provide a unique identifier for each subject. Some pilot studies have been conducted to test the effectiveness of the conceptual data models and provide a prototype that implements the design concepts. Before a detailed plan for new data collection contents and procedures are established, a ***simplified database schema*** is developed in these proof-of-concept efforts. It can manage the existing subject data and can associate various kinds of scan imagery and conventional physical measuring data related to the subjects. With detailed data management needs provided, this simplified schema can be extended to store all collected data as proposed.

A simplified implementation was prototyped in fall of 2019 using SQL Server on a virtual machine at Mercer University's Computer Science Department (Fig. 5). The Family and Subject tables are populated with data originally from CPRC, with derived values like generation, lifespan, and a unified subject ID. Dummy data collected from online sources are used to populate other tables (such as Scan and ImageSeries) to provide test data necessary for the graphical interface development efforts. To be compliant with standard bioinformatics ontologies (such as Uberon) for interoperability with other data sources and search tools, Uberon IDs for body parts (bones or teeth) and diseases will be included in related tables.

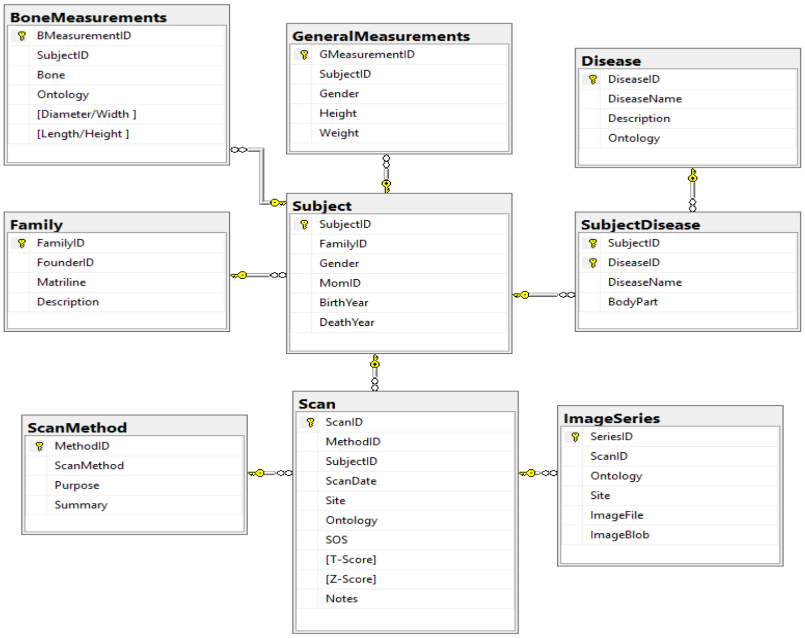


Figure 5. Relational Schema of a Prototype Database for Testing Purpose. Note: Though this looks similar in the flowchart, Fig. 2 presents a data "conceptual" data model, developed in "analysis" phase of the ADLC, with certain details omitted intentionally to deal with layers of complexity. Fig. 5 is a design model or DB schema that is specialized for this simplified prototype implementation of the more abstract analysis model. Again, per regulations and requirements set by Caribbean Primate Research Center, this project treats every monkey individual as a patient and thus protects its privacy as we practice with human patients. All assigned IDs will be a coded ID, not original tattoos.

Framework of the GUI design for the window-based application was developed in spring of 2020. It includes a comprehensive window that can anchor components for displaying a family tree, series of scan images, tabular presentation of basic body measures. dialog boxes that support data entry and update, as well as some data analytics tasks.

Interactive family tree (shown in screenshot Fig. 6) can be used to display all subjects in the same family. The Search menu can facilitate selecting a family by family ID and a matriline family tree starting from the founder can be generated using data from the DB and displayed in the panel. Various icons are used to indicate mother, female or male descendants, as well as subjects whose gender is labeled as U (for unknown). Subtrees starting from each mother node can be expanded or collapsed to show or hide details. When a subject node is selected, related imagery and measurement data will be displayed in corresponding panels as illustrated in Fig. 4.

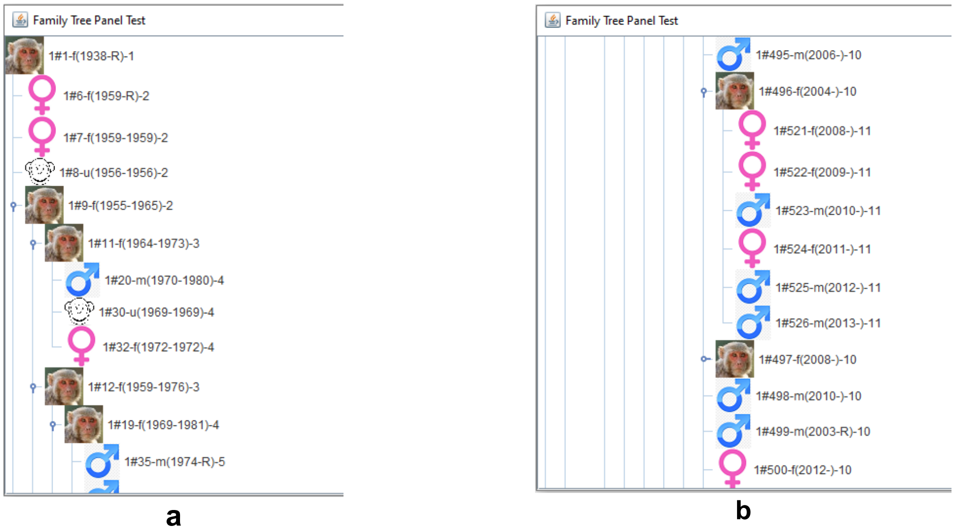


Figure 6 Illustrations of the Interactive Family Tree Panel. a. Family Tree Starting from Founder. b. Tree Expanded to 11th Generation (with M, F, U, and Mom icons).

Certain GUIs that can be used to support data entry/editing and data analytics tasks have also been developed in the prototype system. Dialog boxes for manipulating data in a table are provided to support operational staff to view and edit data in the DB. These dialog boxes can pop up from the menus. A screenshot of a dialog box used for manipulating entries in the ImageSeries table is shown in Fig. 7a. Operation staff can use this interface to edit data collected from the scan, add a new image, or delete an image, as necessary. Values in primary key and/or foreign key fields are not changeable in this interface.

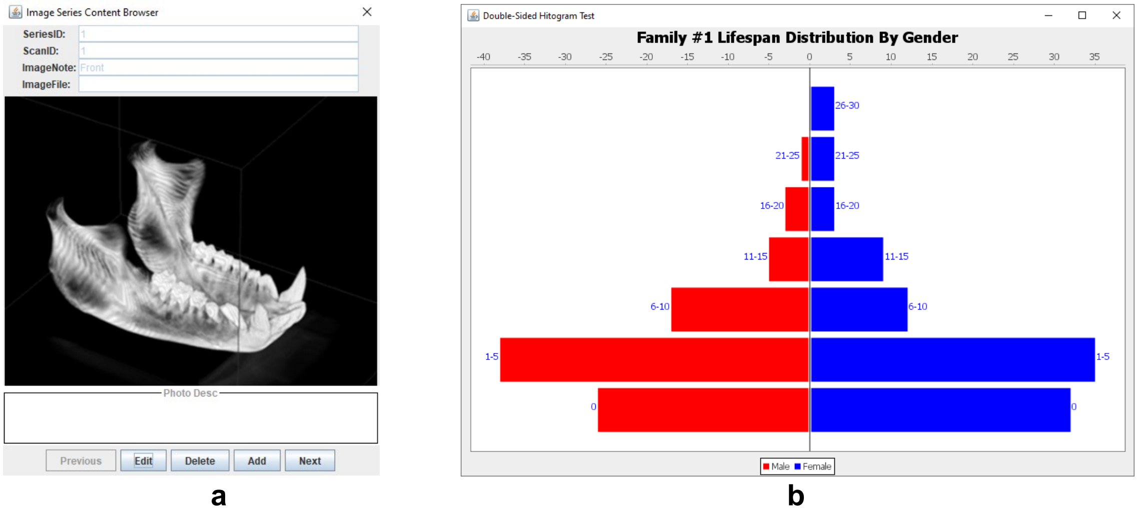

SeriesID: 1
ScanID: 1
ImageNote: Front
ImageFile:

Figure 7 Screenshots of Additional User Interfaces . a. Dialog Box for Manipulating Entries in the ImageSeries Table. b. b. Additional Window for Showing Data Analytics Results.

Fig. 7b. illustrates a screenshot of a separate window showing a dual histogram of subject count by gender using data from one rhesus family. Only subjects with a known death year value (i.e. not removed or still alive) and whose gender is not labelled as unknown (U) are included. Lifespan values are calculated using the difference between death year and birth year data, and categorized in bins for 0, 1-5, 6-10, and so on. An open source Java charting API JFreeChart (jfree.org) is used in generating the two-sided bar chart.

A SQL Server instance has been set up on Amazon Web Services (AWS) cloud-based facility to build the simplified database to concept-prove the possibility of exposing the proposed database to the research community after it is built. Other cloud-based technologies to make the data accessible will be explored for the web-based application to be developed later. Integrated data and information could be extracted in Excel, comma separated values (CSV), or other forms for easy datamining.

CONCLUSIONS & FUTURE WORK

With careful design and data collection, we will build a web-based interface and make it useful for future scientific queries and studies with proper privacy protection mechanisms. Hypotheses regarding sex-based difference, ageing, geographic adaptations, and impacts of natural disasters such as hurricanes could be tested to examine the correlation between natural and/or independent factors and anato-physiological features for developmental, evolutionary, and biomedical studies.

ACKNOWLEDGEMENT

The CPRC Skeletal Collection has been supported by National Institutes of Health NIH contracts NIH 5P40OD012217. This project is supported by NSF grants to M.Q.Z., L.K, D.G.-S. and Q.W. (NSF #1926402, 1926481, 1926528, 1926601). Dr. Melween I. Martinez Rodriguez (Current CPRC Director), Mr. Bonn V Aure Liong, Dr. Angelina Ruiz-Lambides, and other CPRC staff members for their support and help. Co-author Terry B. Kensler, who currently manages the collection, is thanked for her excellent curatorial skills. Dr. Li Sun is thanked for her help, support, and patience. Special thanks go to Mr. Jesse Sowell for setting up the database servers, and students at the Mercer's Computer Science Department for their contributions. We also thank the Editor and reviewers for very constructive comments.

REFERENCES

Dunbar DC. 2012. Physical anthropology at the Caribbean Primate Research Center: Past, present, and future. In: Wang Q, editor. Bones, genetics, and behavior of rhesus macaques: *Macaca mulatta* of Cayo Santiago and beyond. New York: Springer. p 1-35.

Kohn LAP, Bledsoe Z. 2012. Genetic and group influences on postcranial morphology in rhesus macaques (*Macaca mulatta*) of Cayo Santiago. IN Q. Wang (ed). Bones, Genetics, and Behavior of Rhesus Macaques: *Macaca mulatta* of Cayo Santiago and Beyond. Springer Verlag: New York. p. 117 – 129.

JFree.org, 2020. The most widely used chart library for Java (Accessed on May 10, 2020).

Kessler MJ. 1989. (Editor). Proceedings of the Meeting to Celebrate the 50th Anniversary of the Cayo Santiago Rhesus Monkey Colony. Puerto Rico Health Science Journal. 8(1). p. 1-200.

Kessler MJ, Rawlins RG. 2016. A 75-year pictorial history of the Cayo Santiago rhesus monkey colony. American Journal of Primatology 78:6-43.

Kessler MJ, Wang Q, Cerroni AM, Grynpas MD, Velez ODG, Rawlins RG, Ethun KF, Wimsatt JH, Kensler TB, Pritzker KPH. 2016. Long-term effects of castration on the skeleton of male rhesus monkeys (*Macaca mulatta*). American Journal of Primatology. 78:152-166.

Li H, Luo W, Feng A, Tang ML, Kensler TB, Maldonado E, Gonzalez OA, Kessler MK, Dechow PC, Ebersole JL, Wang Q. 2018. The odontogenic abscess in Rhesus macaques (*Macaca mulatta*) from Cayo Santiago. American Journal of Physical Anthropology. 167:441-457.

Microsoft Corp. 2020. SQL Server technical documentation, https://docs.microsoft.com/en-us/sql/sql-server/?view=sql-server-ver15 (Accessed on May 10, 2020).

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE , Haendel MA. 2012. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012 Jan 31;13(1):R5.

Turnquist JE, Hong N. 1989. Current status of the Caribbean Primate Research Center Museum. PR Health Sci J 8:187-189.

Rawlins RG, Kessler MJ. (Editors). 1986. The Cayo Santiago Macaques; Albany: State University of New York Press.

Sade DS, Chepko-Sade B, Schneider J, Roberts SS, Richtsmeier JT. 1985. Basic demographic observations on free-ranging rhesus monkeys. New Haven, Human Relations Area Files Press. p1-98.

Seo D, Lee S, Lee S, Jung H, Sung WK.2013. Construction of Korean Spine Database with Degenerative Spinal Diseases for Realizing e-Spine, KSII. The 8th Asian Pacific International Conference on Information Science and Technology (APIC-IST) 2013, Jeju, Republic of Korea.

Wang Q. 2012. (Editor). Bones, Genetics, and Behavior of Rhesus Macaques: *Macaca mulatta* of Cayo Santiago and Beyond. New York: Springer.

Wang Q, Dechow PC, Hens SM. 2007. Ontogeny and diachronic changes in sexual dimorphism in the craniofacial skeleton of rhesus macaques from Cayo Santiago, Puerto Rico. Journal of Human Evolution. 53:350-361.

Wang Q, Kessler MJ, Kensler TB, Dechow PC. 2016a. The mandibles of castrated male rhesus macaques (*Macaca mulatta*): The effects of orchidectomy on bone and teeth. American Journal of Physical Anthropology. 159:31-51.

Wang Q, Opperman LA, Havill LM, Carlson DS, Dechow PC. 2006a. Inheritance of sutural pattern at the pterion in rhesus monkey skulls. Anatomical Record. 288A:1042-1049.

Wang Q, Strait DS, Dechow PC. 2006b. Fusion patterns of craniofacial sutures in rhesus monkey skulls of known age and sex from Cayo Santiago. American Journal of Physical Anthropology. 131:469-485.

Wang Q, Turnquist JE, Kessler MJ. 2016b. Free-ranging Cayo Santiago rhesus monkeys (*Macaca mulatta*): III. Dental eruption Patterns and dental pathology. American Journal of Primatology. 78:127-142.

Zhao MQ. 2010. Knowledge representation and reasoning for impact/threat assessment in cyber situation awareness systems. Final Report to AFRL/RI, Rome, NY, June 2010.

Zhao MQ. 2012. Analysis tool development for quantifying the SITA System. Technical Report to AFRL/RI, Rome, NY, August 2012.

Zhao MQ. 2018. A first course in database systems using SQL Server. Published by Linus Learning, Ronkonkoma, NY, 2018.