**Research Data Management Teaching Case: Combining Data from 10 Years of Research for Retrospective Studies on the Effects of Exercise and Diet on the Risk of Diabetes**

**Summary of Teaching Points**

**Module 1: Overview of Research Data Management**

* Challenge in pooling together data retrospectively from several large projects
* Lack of oversight of data management activities in lab
* Graduate students taking data upon leaving institution

**Module 2: Types, Formats & Stages of Data**

* Most of the projects’ data has been stored on Excel spreadsheets which can easily be saved in other formats
* Types of instrumentation and measurements have remained the same, but different instrument models may have been used. Is there a possibility that some of the project data could be in an obsolete proprietary format?

**Module 3: Contextual Details**

* Researchers want to regroup responses from previous studies by classifying them under tissue response rather than singular responses of different tissues
* Lab does not use standard naming conventions
* No established standards for file structures
* Names and models of instrumentation are not recorded with the data

**Module 4: Data Storage, Backup, and Security**

* Data from past studies left in the lab’s online file storage system account maintained by the university
* Oversight of keeping data secure in lab has been inconsistent: students take data with them when they leave institution
* No standardized practice for where data should be stored (found in paper notebooks, iPad hard drives, individual researcher’s hard drive or personal University networked drive
* Inconsistent backups for data kept on personal hard drives

**Module 5: Legal and Ethical Issues**

* Ownership: graduate students taking project data with them when they leave
* Privacy: data is anonymized
* Consent forms kept in locked cabinet

**Module 6: Data Sharing and Re-use**

* No restrictions on data sharing
* Data sharing is done through university’s online file storage system: what about sharing with those outside the university?
* Requests for data sharing handled by PI

**Module 7: Plan for Archiving and Preservation of Data**

* No plan
* Researchers unaware of data repository in their field

**Research Data Management Teaching Case: Combining Data from 10 Years of Research for Retrospective Studies on the Effects of Exercise and Diet on the Risk of Diabetes**

**UMass Amherst**

The head of an Energy Metabolism Laboratory in a Department of Kinesiology wants to combine data from 15-30 large projects, spanning 10 years of studies, so that the data can be used for retrospective studies. The studies included in this research were mostly funded through the National Institutes of Health and the American Diabetes Association. These original projects studied the effects of different regimes of exercise and diet on preventing the development of type 2 diabetes. They studied the effects of

* a single bout of exercise, OR
* several days of exercise and diet manipulation, OR
* several months of exercise training

on aspects of metabolism that indicate the likelihood of developing type 2 diabetes. Common metabolic measurements recorded were

* fasting and post-meal blood sugar and insulin,
* cholesterol,
* free fatty acids/triglycerides,
* and other markers of cardiovascular and metabolic health.

The researchers are in the early stages of planning for the retrospective studies, but a general thesis is that the response to exercise/diet affects the tissues involved in regulating blood sugar differently based on the dose of exercise. Typically, the past studies have focused on exercise and the singular responses of the different tissues (liver, muscle, etc.). The researchers would like to group the responses of all previous studies by tissue response, rather than by the exercise group.

The current project involves a Principal Investigator to oversee conclusions drawn, a doctoral student to the P.I., and a few undergraduate students recruited to help reduce and analyze the data. The past studies were completed with this same P.I. providing oversight and graduate students performing much of the research. These Masters and Ph.D. students have executed 15-30 projects as part of this research.

The lab produces enough data to justify a lab coordinator, but the lab is not large enough to fund this position. As a consequence, there has not been consistent oversight of the data over time. As the students have completed their degrees, the data from these studies were left in the Lab’s University-maintained online file storage system account.

The data are existing experimental data of blood glucose levels and other metabolic measures obtained from observational instrumentation. Types of instruments and measurements have remained the same, but over the years, different instrument models may have been used. Data were recorded by graduate students either on iPads or in paper lab notebooks (which the students took with him/her upon graduation). Data were entered in Microsoft Excel files and analysis was performed using “R” statistical program.

The data types are numerical evaluations of key metabolic variables, which can be interpreted alone or as “change” (how the initial value changes after exercise). An example is insulin action (a numerical assessment of the body’s response to insulin used in virtually all of the studies) both at baseline and after a bout of exercise or across multiple standardized and controlled conditions.

Multiple files were associated with larger projects. Some of these files group data collected under different conditions (e.g. single exercise bout versus exercise training) or based on the grouping of specific hormone/metabolite/responses such as insulin sensitivity or blood glucose to measure blood sugar response, or a triglycerides/free fatty acids spreadsheet (for fat response), or leptin/ghrelin (for appetite regulation).

The names and models of instrumentation are not recorded with the data. The researchers understand which equipment was used when they refer to their datasets. The lab does not use standard file naming conventions. File structures are not standardized and are established by the preference of the researcher. Data are in the Excel file format, which is proprietary, but can easily be saved in other formats. Specific information about data collected can be found in associated published papers.

During research projects, data are stored in paper notebooks, on iPad hard drives, or on an individual researcher’s hard drive or personal University networked drive. Some researchers and students have personal files backed-up, but there is no standardized practice.

After each research project is completed, data are stored on the University-maintained online file storage system. All data on this drive are stored on a network attached storage (NAS) device that uses a snapshot technology. The data are captured several times each day, and are asynchronously replicated from the primary, "production" NAS device to a secondary NAS device. Nightly, all data from the secondary NAS is written to a tertiary disk-based storage device. Weekly, two full copies of all data from the online file storage system is written to tape. One set of tapes is stored in a fire safe and the other is sent to an off-campus location. More than 100 sheets of data in Excel files on the megabyte scale comprise the post-project data. The plasma and serum collected from past studies has been kept at a controlled temperature so that the studies could be replicated, or so that the samples could be used in new studies.

The Institutional Review Board approved all past research projects. The data is anonymized; thus subjects’ names and dates of birth are not tracked. Characteristics such as age, height, weight are tracked and associated with an identification number. Once the characteristics are attached to the identification number, the subject names are no longer used. Print informed consent forms are kept in a locked cabinet, separate from current and past studies. The researchers desire a safe for privacy and security purposes.

One goal of the researchers is to share and combine their data with colleagues’ data. There are currently no restrictions on data sharing and re-use of analyzed data, and requests are handled on a by-request system to the P.I. The researchers do not report institutional review board or intellectual property issues related to sharing data. Files are shared through the University’s online file storage system. The researchers are open to sharing post-project data through a repository, but they are unaware of a data repository in their field. To view the data, Excel or another spreadsheet application is required.

There are currently no plans for archiving and long-term preservation of the data.

**Discussion Questions**

Module 1: Overview of Research Data Management

What issues need to be addressed in this project related to the 7 segments of the data management plan recommendations?

Module 2: Types, Formats, and Stages of Data

1. What types of data are being collected for this study?
2. How will data management responsibilities be delegated among the members of the research team?
3. How should the project team plan to regroup responses of previous studies by tissue response?
4. What format would be the best choice for the combined data from all the past projects?
5. How will you ensure that all research staff use the same data sources and data definitions?
6. What details should be added about how the data was collected?
7. How will the project team name files of response data from past projects that have been regrouped by tissue response rather than by exercise response?

Module 3: Contextual Details Needed to Make Data Meaningful to Others

1. What metadata is needed to enable the project data to be reused?
2. What standards (e.g. for entering time or geographic data) should be used for metadata fields?
3. Can you think of a metadata standard that could be used in this project?

Module 4: Data Storage, Backup, and Security

1. How can the lab improve upon maintaining consistency in data storage and avoid loss of data?
2. Should data from project be kept on individuals’ hard drives?
3. What measures should be taken to keep data secure?
4. This case discusses backup strategy for the data upon the completion of a project, but how often and where will the data be backed up during the course of the project?

Module 5: Legal and Ethical Considerations

1. Do all members of the research team understand who owns the data?
2. Do the consent forms from past projects state that data from the projects may be repurposed or reused in future projects?
3. Aside from names and dates of birth, is there any other data that could potentially reveal subjects’ identities?
4. How will the data from past studies be cited?

Module 6: Data Sharing and Reuse Policies

1. How will this project data be shared? (upon request, data repository, institutional repository, via publisher)
2. What policies should be checked before sharing project data with external researchers? (funders, institution, IRB, publisher)
3. What format for the data would facilitate re-use of the data by external researchers?
4. What quality assurance measures should be taken before sharing data to external researchers?

Module 7: Archiving and Preservation

1. What would be the benefit of preserving and archiving this project data?
2. Is there an ideal repository for this data? (e.g. institutional, publisher’s, disciplinary)