**Case: Outcomes from Orthopedic Implant Surgery**

**Summary of Teaching Points**

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

* the challenges in conducting a multiyear research project with changing personnel each year

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

* Proprietary software using file naming conventions that were not clear and not under control of investigator and software stored data in its own application database that needed to be exported to a common format for storage
* The need to update proprietary software as new releases become available to maintain support from vendor and keep data formats current

**Module 3: Contextual Details**

* No validity checks on data entry for patient survey data

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

* No plan for storage of survey source documents
* Security and backup plan in place

**Module 5: Legal and Ethical Issues**

* Use of novel instrumentation with single license proprietary software and need to use it on multiple PCs
* Need to de-identify patient data when doing research with human subjects
* Need for IRB application and informed consent form

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

* Informed consent restrictions

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

* None

**Research Data Management Case: Outcomes from Orthopedic Implant Surgery**

Dr. X wrote a 5 page proposal for funding for a study to use a novel monitor with proprietary software to assess patient outcomes 2 years after orthopedic implant surgery. This prospective longitudinal study would determine the rate of sub-optimal outcomes based on specialized analysis using the proprietary software that accompanied the monitor. The study was funded and the research resident working with the PI prepared the IRB application that received approval. With a clearly defined research hypothesis, innovative monitor technology, and IRB application and consent form complete, the goal was to collect the same measures over 3 years. The resident on the project began to enroll patients and collect data on his office PC. At the end of the training year, the resident handed the study to the next resident whose responsibility was to continue enrollment and collect the one year follow-up data on the initial cohort. At the end of study year 2, the third resident continued to enroll more patients, collect 2 year outcomes from the first cohort and 1 year outcomes on the second cohort. A very large volume of data had been collected and the new research resident was responsible for integrating and analyzing the data in preparation for publication the following year. She encountered a series of data issues that were not documented or clear to her. While the PI had originally defined the data to collect, she had not been directly involved in the data collection and could not answer the questions. The first resident who started the project had completed training and left the institution.

The same patients were followed for three years so it required tracking them down to have them come in to allow for collection of data via multiple sources: patient surveys, accelerometer measurements, and surgeon notes from the physical exams. The Principal Investigator had HIPAA authorization to use the patient’s name, Med record #, and telephone/address to contact them for follow-up. However, the data base was organized by unique study ID assigned to each patient.

The study was complex due to the need to collect and integrate data from these three different sources:

1) Patient-generated data regarding their demographics and their symptoms, the amount of pain and disability. Patients filled out a hand-developed paper survey at baseline and annually for 3 years. The core outcome measures were the same from the survey each year, but the basic demographics questions were not repeated each year. Much of it was based on pre-existing standardized forms so there were already some data definitions for some responses. The format was a mix of these standardized questions (well tested and validated responses) and some new questions with uncertain responses (open-ended response option). Data were hand-entered into an Excel spread sheet; so there was no application of data quality checks (number range, etc) as data were entered. Survey data were entered by various people into an excel spreadsheet and the source documents were stored in multiple locations. Eventually the patient surveys were moved onto a direct computer data entry system to avoid the validation problem. The data were captured in survey software that could be downloaded into a spreadsheet/data file for analysis.

2) The second source was measurements from an accelerometer that did 24 hour tracing of patients’ steps and walking rate annually. This novel monitor came with proprietary software that produced bulk summary statistics on an excel spreadsheet. However, the study required individual patient records that had to be exported for analysis. We exported the data on each patient from the software to a data file. The monitor analytic software was on a lab PC originally. It was a proprietary software package that could be loaded only on one computer and it had to be handed off as residents changed. The rest of the data from other sources were on the research assistant’s PC. We bought another monitor software license to get it off the original PC because the monitor analytic data were housed there and we then put it on a laptop. The specialized monitor analysis software used naming conventions that were not clear and data were stored in the proprietary software. The software itself was updated across the 3 years.

3) The third source was a surgeon note in the EMR and there was no standard for this surgeon note resulting in varied styles of documentation. Residents read the charts every month related to patients in the study to identify any follow-up MD office visits and to extract physical exam measures which were inserted into a structured database with data definitions for each measure.

The data from these multiple sources needed to be integrated for a biostatistician to apply longitudinal modeling software. ACCES was used as the final data base and was used to house the total data set and integrate data (through a flat file) from all the sources. Data sub-sets were imported to STATA software for particular analyses, as needed. Data were stored on a server solely for research that was password protected, backed up nightly, and protected by institutional firewalls, etc. (not on a computer). STATA software was used for data analysis such as linear and logistic multi-variate models. Backup was done nightly through the institutional IS procedures for data stored on their research servers. Security measures such as passwords, limited access, firewall, etc. were used to safeguard the data.

**Module 1 (Overview module) discussion question:**

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

**Discussion Questions for Other Modules:**

1. Types of data
   1. What types of data are being collected for this study?
   2. How will you ensure all research assistants/residents used the same data sources and data definitions?
   3. What would be needed in a data management plan to describe use of novel equipment?
   4. What needs to be in the plan related to the patient survey data capture and the capture of surgeon notes?
   5. What analytical methods and mechanisms will be applied to your data either prior to or post integration
   6. What type of outcome data will be generated?
2. Contextual details
   1. What file formats and naming conventions will be used for the separate data sources and for the integrated file used for analysis?
   2. What impact would the naming conventions, proprietary software, and software updates have on later data access?
   3. What other contextual details would you specifically need to document to make your data meaningful to others?
   4. In what form will you capture these details?
3. Data Storage, Backup, Security
   1. Where and on what media will the data from each data source be stored?
   2. How, how often and where will the data from each source be backed up?
   3. How will you manage data security across research assistants/residents on the study for each data source?
   4. How long following the completion of your study will you store the data?
4. Data protection/privacy
   1. How are you addressing any ethical or privacy issues?
   2. What mechanism are you using to identify individual patients?
   3. Who will own any copyright or intellectual property rights to the data from each source?
   4. How will the dataset be licensed if rights exist?
   5. How will the data be associated with a study ID?
5. Policies for reuse of data
   1. How will you create a de-identified copy of the data?
   2. Will a new patient consent be required for subsequent re-use of data collected specific to the purpose of this study?
   3. Will the data be restricted to be re-used only for certain purposes or by specific researchers?
   4. Are there any reasons not to share or re-use data?
6. Policies for access and sharing
   1. Will some kind of contribution or fee be charged for subsequent access to this data?
   2. What process should be followed to gain future access to your study data?
7. Archiving and preservation
   1. What is the long-term strategy for maintaining, curating and archiving the data?
   2. What data will be included in an archive?
   3. Where and how will it be archived?
   4. What other contextual data or other related data will be included in the archive?
   5. How long will the data be kept beyond the life of the project?