Homework 7

Choose one of the Case studies and based on that fill in ‘initial DMP’ using https://dmponline.dcc.ac.uk/ tool.

Case studies are taken from **Frameworks for a Data Management Curriculum**: Course plans for data management instruction to undergraduate and graduate students in science, health sciences, and engineering programs (link).

Research Data Management Case A: 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 database 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 spreadsheet; 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.

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.

3) The data from these multiple sources needed to be integrated for a biostatistician to apply longitudinal modelling software. ACCES was used as the final database 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 multivariate 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.

Research Data Management Case B: Regeneration of Functional Heart Tissue in Rats

The goal of the study is to try to regenerate functional heart tissue in a rat. Unlike other organs and tissues which regenerate themselves, the heart does not have the ability to regenerate, so we intend to regenerate it by delivering stem cells to the heart. The hope is that in generating heart tissue, we generate tissue that is actually functioning and contracting and doing mechanical work.

Two days before we operate on the rat, we take adult stem cells and incubate them for 24 hours with our marker for cells [fluorescent nanoparticles]. We then put them in a solution and inject them into a tube that has a biological suture in it, so the cells sit down on the outside of the biological suture. We incubate it for 24 hours, and then do the surgery. During the surgery, we open up the thoracic cavity of the rat and create a myocardial infarction by occluding the left anterior descending coronary artery. At this point it is ischemic; we keep it ischemic for 1 hour, not letting any blood flow go through, and then we reperfuse it and let the blood go back. About a minute after that, we put the biological suture with the cells on it through the infarcted region. We then close the rat up and put it back in the cage for a week. We go back a week later, open the rat up again, and use our camera system to acquire images of the heart. We take images with two cameras simultaneously and we'll also have a pressure transducer that syncs automatically with the pictures inside the left ventricle cavity to measure left ventricle pressure. Then we reposition our cameras and take another data set and we usually do that about 4 or 5 times to look at different regions around that infarct.

Then we euthanize the animal. We isolate the heart. We fix the heart in a fixative and then put it in the freezer for about 24 hours. Then we start cutting sections of the heart and putting them onto slides – about 3 sections of the rat heart per slide. We generate about 200 slides per rat heart. At any time, some tissue that was sectioned and on slides may be in one freezer, and some tissue that had not been sectioned yet but was embedded and ready to be sectioned is in another and still other tissue that

may be sitting in a container someplace in another freezer. It should be entered into the excel spreadsheet saying what was done and where it is, but that doesn't always happen. Then we stain some slides and then sometime after – anywhere from a day to a couple of months - we stain some of them with trichrome. That tells us what tissue is dead. We stain some of them for specific markers in looking to find out exactly where the stem cells are in that cross-section. Then we take images on our microscope, which is an epifluorescent microscope and, if we are happy with the staining and the way they look, then we make an appointment to use the confocal microscope which takes much better quality pictures and take those images on the confocal. At the same time, we also look at the data we acquired and use our home-grown custom software to track particles on the surface of the heart to see how far and how fast those particles are moving. The software was written in C and MATLAB (C runs the code faster, but MATLAB is easier to work with; usually we develop the code in MATLAB and then convert it to C so that it runs faster). We use this software to analyze the optical images of the heart. That tells us what the function is like in that region of the heart. We do that for several heartbeats in different data sets. And then we save that data. That is everything we do for one heart.

Data sets:

1) We have the optical images after the first surgery to insert the cells – on average for one experiment we probably have about 10,000 images. We acquire images of the heart at about 250 frames/second. We acquire 4 seconds worth of data, so we have 1,000 images for each data set. The images are initially stored on the hard drive of the acquisition computer, then are transferred to a Drobo backup system and the hard drive of a network computer that is backed up by the institution.

2) The second data set is where we measure the left ventricle pressure at the same time we are acquiring those images so that we know that image # 127 correlates with the pressure at time point # 127 milliseconds. For this measure we use an analog to digital (A/D) board and a Millar pressure transducer. Both the camera and the acquisition system are computer-controlled to synchronize them. These data are stored in the same way as the optical images, although they are separate files.

3) Electronic data sets are used to acquire images from the different stained tissue sections after the second surgery. We may stain on average 4-5 different markers and we will have different data sets for the different stains. So we will have images taken from the epifluorescent scope, and, usually in a cross-section of heart, we may take some high-resolution images in zoomed-in regions and some low magnification images. On average for each section, we take about 20 pictures with the epifluorescent and then with the confocal, we probably take about another 20; if we take a z-stack with

the confocal, that can be an additional 200 images. They are both taking the same images except one is a much better resolution than the other. Long term storage for these datasets is on the Drobo and DVD backups.

Our naming convention is that we name our files EXP (for experiment) and then usually 4- or 5-digit codes like 2001, and then we have several data sets so it is DS # and then it could be image 1. Then we need to link these data sources together via an excel spreadsheet. The sections are all linked by the same experiment number. They are linked with the digital images just based on what number section they are.

Multiple research staff may be analyzing the same heart, and one person will be doing the mechanical function of the heart, one will be doing the trichrome staining, another will be doing the actinin staining and maybe another will be doing the imaging. The data sets should all be linked in the excel spreadsheet. There could easily be up to 10 people involved in data analysis, and we have not yet found a good way to link all the data. We have an excel spreadsheet basically, and it says in DS1 – the mechanical function in this area was xyz, in DS2 it was this. In DS1 the tissue section showed this, and we try to link them all up together, but the tissue sections are on conventional microscope slides that are stored someplace. Even that – the location of where they are stored is a problem. We have the usual places where we store things but we have 3 or 4 freezers and if it is not in 1, we look in 2, and so on. The slide box is labelled with the experiment number and the individual slides are labelled with the slide number & experiment number.

The types of data we use are mostly images and numeric measures in addition to the lab notebook which may have some observational notes. Some of it is number-crunching but a lot of it is images.

The content of a lab book relates to a particular experiment and is used by all staff working on that experiment. There is a format they are all supposed to follow, which they don't always do. There could be on average 5-6 people using the notebook. The paper lab notebook basically performs the function of being an index into the actual datasets and it should record all the information the PI specifies. We also have a paper surgical log that is kept with the animal and whatever project staff writes down in that surgical log should be transferred into the lab notebook – so it has to be in 2 places. It has to be down there in case there is a problem with the animal, but the PI also needs it in the lab notebook to be able to write papers. The older lab notebooks are in the PI's office, but the ones that are currently in use are in the lab. Older Lab notebooks are only in the PI's office of the lab with no backup. The lab notebook has to be in pen on specific paper because this paper is supposed to be good for 100 years.

We back up the data sets on an external hard drive someplace. The optical and electronic images are both backed up. The current backup system we are using truncates the data set name to 6 digits then puts a tilde sign and number starting from 001.

The files are not password protected or anything. The lab notebook is either in the PI's office or (most often) in the lab, which has key card access (although it is an open floor plan).

Research Data Management Case C: Improving End-of-Life Care for African Americans

An MD applied for grant funding to do a qualitative study focusing on how to improve physician communication with African Americans (AA) and their relatives when their patients were receiving end-of-life care.

This qualitative study was conducted to expand knowledge about AA experiences and opinions about end-of-life care. Multiple-meeting focus groups were held to build trust and allow time for full participation. Following a review by a Community Advisory Board (CAB), protocols were approved by the University‟s Institutional Review Board. Participants were AA adults who had experienced at least one death of a significant other or family member. Convenience sampling by staff and CAB members was used to recruit participants, and flyers were distributed at neighbourhood activities. Participants were screened for eligibility and assigned to one of two focus groups. Focus group 1, which met for four sessions, was comprised of AAs with family members who had died at home. Focus group 2 met for three sessions and included AAs with family members who had died in the hospital. An average of five individuals attended each session. Three participants worked in health care, and their observations reflected experiences with a dying family member, as well as experiences with caring for terminally ill AA patients.

Data collection All participants gave informed consent. An open-ended interview script stimulated discussion about (1) positive and negative experiences of participants related to end-of-life care in the hospital or at home, (2) preferences for treatment by health care providers, (3) communication issues, and (4) end-of-life decision making pertaining to living wills and advance directives. An AA member of the project staff moderated the focus groups.

Each session was audio-taped taped. Unlabeled tapes were mailed to a transcriptionist in their plastic cases which were labelled. During the mailing process the

package was damaged and the plastic tape cases broke and were no longer associated with the tapes for which the cases had been labelled. The tapes, however, were not damaged. The transcriptionist transcribed the tapes and the transcripts were sent back to the project team for identification of which focus group and which session should be used to identify each transcript. Focus Group Participants‟ comments were identified on the transcript by either Miss, Mrs. or Mr. plus the first initial of their first name. The transcripts were also reviewed for accuracy by the project team.

Data analysis Transcripts were reviewed for themes through a continuous process of text data segment comparison based on qualitative research techniques. After reading the transcripts several times, a codebook was developed defining themes and subthemes, and a numeric theme code was assigned to each particular category of text responses. Microsoft Word was used to create transcript tables of participant responses which then could be sorted by theme code. Participants‟ responses were coded and sorted accordingly into differing categories which were then summarized to capture the richness and range of data within each theme code. The analysis was systematic and involved triangulation of data from the two focus group sources. Within-focus group set analyses were performed, as well as cross-focus group set analyses to develop a set of themes/recommendations for how end-of-life care communications might be conducted to improve the process for all concerned.

Resulting Data:

In a subsequent publication, the results were published as follows: Analysis of the transcripts revealed five major theme groupings. These groupings contained text data related to:

1. Communicating about dying and end-of-life care

2. Choice about dying at home or in the hospital

3. Dying in the hospital

4. Dying at home

5. Other end-of-life care issues

Additionally the implications for clinical care were summarized as follows:

• Be mindful of the diversity of preferences and needs within any population subgroup

• Recognize that many AAs have very strong religious and spiritual beliefs about dying and that their words often reflect that the patient is preparing to leave his or her earthly home

• Empower dying AAs and their family members by speaking respectfully, using lay terminology, and checking for understanding. Encourage the patient to be the primary decision-maker and ensure that the dying person is not infantilized.

• Determine whether the dying person and/or caretaker has adequate assistance. Since awareness of home and hospice services is low, facilitate getting necessary support and resources, including connections with social services.

• Encourage patients to decide how the family should be informed about prognosis and provide assistance in telling the family if requested.

• Determine in advance who the primary family contact is and where to contact him or her in the final hours if the patient is hospitalized. If possible, ensure that the family has the opportunity to spend the last hours with the patient. The “gathering of the family” is very important during this phase of life.

• For patients dying in the hospital, treat patients the way you want to be treated with nurturing, compassion, dignity, love, touch, and careful listening. Diligent monitoring of the patient‟s medical status, needs, and cleanliness is imperative.

The tapes were eventually destroyed and the transcripts and other files generated during the analysis remained with the analyst who was not part of the project team and was affiliated with another medical school. The analyst was very involved with the drafting of the publication. Excerpts from the transcripts were later re-used as examples for a qualitative analysis class taught by the analyst; however, for the reuse, all participant IDs were changed to P1, P2, etc.