Case Study – Data Management

A group of three resident physicians were working with their attending physician on a retrospective research study to compare patient outcomes before and after a new therapy was implemented. The residents were responsible for abstracting data from medical records after IRB approval was obtained and were recording the data in a REDCap database to which they and the attending physician had access. The residents split up the task of abstracting the data and were collecting it over a 2 month period of time. From time to time, one of the residents would check to see if any new patients had the surgical procedure and if so, additional patient data was added to the REDCap database.

One of the residents downloaded the data from the database and performed the statistical analysis planned for the study. The data looked interesting and the resident provided it to the attending, who had already begun drafting a publication. The resident also wrote it up for an abstract, which was presented at a national scientific meeting. The manuscript was completed and submitted to a peer-reviewed journal for publication.

A few months later, the resident received comments from the journal's peer-review process, which were provided to the attending for review. Several comments involved the data and statistics presented in the manuscript. The attending asked the resident who performed the statistical analysis to provide the dataset used for analysis. The resident realized the dataset had not been saved, nor was the date of the original download from REDCap known. Neither the resident nor the attending could determine from which patients the data for the statistical analysis was obtained and the results could not be replicated.

What should the attending do next?

Because the patient data was stored in REDCap, which is a fully auditable system, the attending requested that every dataset downloaded by the resident be provided so that the statistical analysis could be repeated. Three different datasets were identified and re-analyzed, but the results could still not be fully replicated. Some of the statistical results were very close to the original, but others were not.

What possible explanations could there be for the inability to replicate the data?

With further review, the resident remembered that data from certain patients were removed from the dataset before analysis, based on criteria that had been discussed when the IRB protocol was submitted, but the criteria were not included in the protocol, were not written down by the residents anywhere, and the attending was not told what criteria had been applied. The resident who did the analysis was not 100% sure what the "final" criteria used were, but wrote down what s/he could remember. The attending repeated the statistical

analysis after removing patient data using those criteria, but was still unable to fully replicate the results submitted in the manuscript and decided to withdraw it.

What problems can you identify with how data was handled in this case?

Who was responsible for what happened?

What could be done in the future to prevent something like this from happening again?