In-class exercise

*Work in your Project 1 groups on this exercise during class. After working on the exercise, each group will present orally the components they were responsible for.*

The ARIC study, which is run out of the Collaborative Studies Coordinating Center (CSCC), is beginning to collect data using the electronic medical record from hospitalizations. These data are collected by sending a spreadsheet with items requested to the hospitals’ data warehouses, and then receiving the extracted data back from the hospital in a delimited file. The data contain both structured data (e.g., weight) and notes from healthcare workers (e.g., “no evidence of chest pain”). To extract data for analysis from these records, a specialized, non-CSCC team is needed with expertise in natural language processing, apart from the established team, procedures, and security privelages of CSCC employees. Thus, responsibilities for data management and analysis are split across two teams with unequal access to hardware, software, and knowledge of methodology.

*Design a data management plan for a 5-year study of the prevalence of undiagnosed hypertension and diabetes among hospitalized patients at these hospitals. Records from 800 randomly sampled hospitalizations will be collected every year across 10 different hospitals for the 3 middle study years.*

For this particular exercise, cover these items:

1. Scope of work and roles required to do it
2. Data management system, software, and hardware
3. Data sources, flow, and frequency of transfers
4. Data validation
5. Data security

library(tidyverse)  
set.seed(123)  
tibble(  
 group = seq(1, 10),  
 item = sample(x = c(1, 1, 2, 2, 3, 3, 4, 4, 5, 5), size = 10)  
) %>%  
 knitr::kable()

|  |  |
| --- | --- |
| group | item |
| 1 | 2 |
| 2 | 5 |
| 3 | 1 |
| 4 | 4 |
| 5 | 3 |
| 6 | 5 |
| 7 | 1 |
| 8 | 4 |
| 9 | 3 |
| 10 | 2 |

There are “data sciency” things I’m sure you can think of here. I encourage you to do so.