Compile a set of two case-studies complete with motivating questions and datasets. Each case-study should include the following components:

* **Title:** Brief but descriptive case-study title
* **Overview**: A brief summary of the case-study
* **Driving Challenges**: List a set of 3-5 questions that have clear healthcare applications that might be addressed, or at least examined by, using the dataset
* **Meta-data**: Define all data elements, describe the dataset
* **Data**: Include the complete dataset. It could be observational, derived, or simulated data. It should be somewhat interesting (e.g., include at least 5 variables, one or more time-points and represent 30 + cases/subjects/instances, hopefully, hundreds of cases).
* **Provenance**: Include some references, URLs, PMCIDs, comments, etc. describing the provenance of these data.

Please upload all cases on Canvas and give them appropriate name providing a clue to the type of data/study this case presents. When you submit your HW electronically (on Canvas) please include the appropriate HW header and reference (by direct URL) the appropriate case-studies you uploaded.

See the example below

**A Case-Study of Parkinson’s Disease using Simulated Big Data**

* **Overview**: This case-study examines the associations between clinical, demographic, imaging and genetics variables for Parkinson’s disease. This is an example of Big Data for investigating important neurodegenerative disorders.
* **Driving Challenges**:
  + Are there relations between imaging, genetic and clinical covariates?
  + Can we predict subject diagnosis using model-based approaches?
  + Do exploratory data analytics provide clues to Parkinson’s disease?
  + If clinical PD Dx is solely based on UPDRS scores, is there value added by including other covariates?
* **Meta-data**:
  + **ID**: Case subject identifier
  + **Imaging Biomarkers**: ComputeArea = surface area of 3D brain region of interest, Volume= the 3D volume/size of the region of interest, region name is encoded in morphometry measure name (e.g., L\_putamen\_ComputeArea represents the left putamen surface area):

L\_caudate\_ComputeArea, L\_caudate\_Volume, R\_caudate\_ComputeArea, R\_caudate\_Volume, L\_putamen\_ComputeArea, L\_putamen\_Volume, R\_putamen\_ComputeArea, R\_putamen\_Volume, L\_hippocampus\_ComputeArea, L\_hippocampus\_Volume, R\_hippocampus\_ComputeArea, R\_hippocampus\_Volume, cerebellum\_ComputeArea, cerebellum\_Volume, L\_lingual\_gyrus\_ComputeArea, L\_lingual\_gyrus\_Volume, R\_lingual\_gyrus\_ComputeArea, R\_lingual\_gyrus\_Volume, L\_fusiform\_gyrus\_ComputeArea, L\_fusiform\_gyrus\_Volume, R\_fusiform\_gyrus\_ComputeArea, R\_fusiform\_gyrus\_Volume

* + **Demographics variables**: Sex, Weight, Age
  + **Diagnosis**: Dx, PD=Parkinson’s, HC=Healthy Control, SWEDD = (tremor associated clinical parkinsonism features) scans without evidence of dopaminergic deficit
  + **Genetics**: chr12\_rs34637584\_GT, chr17\_rs11868035\_GT
  + **Clinical**: UPDRS\_part\_I, UPDRS\_part\_II, UPDRS\_part\_III (Movement Disorder Society-Sponsored Revision of the Unified. Parkinson's Disease Rating Scale (MDS-UPDRS)), <http://www.movementdisorders.org/MDS/Education/Rating-Scales.htm> (Normal=0)
  + **Time**: VisitTime: four time-points (baseline (0), 6, 12, and 18 month follow-ups).
* **Data**: Complete data is in this XLSX table (HS853\_PD\_SimData.xlsx).
* **Provenance**: This case study only uses simulated data. The complete R-script generating the data is included below. The entire case study is CC-BY licensed and can be used, updated, refactored and expanded by the entire community.

**# Define number of subjects**

NumSubj <- 282

NumTime <- 4

**# Define data elements**

**# Cases**

Cases <- c(2, 3, 6, 7, 8, 10, 11, 12, 13, 14, 17, 18, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 34, 35, 37, 41, 42, 43, 44, 45, 53, 55, 58, 60, 62, 67, 69, 71, 72, 74, 79, 80, 85, 87, 90, 95, 97, 99, 100, 101, 106, 107, 109, 112, 120, 123, 125, 128, 129, 132, 134, 136, 139, 142, 147, 149, 153, 158, 160, 162, 163, 167, 172, 174, 178, 179, 180, 182, 192, 195, 201, 208, 211, 215, 217, 223, 227, 228, 233, 235, 236, 240, 245, 248, 250, 251, 254, 257, 259, 261, 264, 268, 269, 272, 273, 275, 279, 288, 289, 291, 296, 298, 303, 305, 309, 314, 318, 324, 325, 326, 328, 331, 332, 333, 334, 336, 338, 339, 341, 344, 346, 347, 350, 353, 354, 359, 361, 363, 364, 366, 367, 368, 369, 370, 371, 372, 374, 375, 376, 377, 378, 381, 382, 384, 385, 386, 387, 389, 390, 393, 395, 398, 400, 410, 421, 423, 428, 433, 435, 443, 447, 449, 450, 451, 453, 454, 455, 456, 457, 458, 459, 460, 461, 465, 466, 467, 470, 471, 472, 476, 477, 478, 479, 480, 481, 483, 484, 485, 486, 487, 488, 489, 492, 493, 494, 496, 498, 501, 504, 507, 510, 513, 515, 528, 530, 533, 537, 538, 542, 545, 546, 549, 555, 557, 559, 560, 566, 572, 573, 576, 582, 586, 590, 592, 597, 603, 604, 611, 619, 621, 623, 624, 625, 631, 633, 634, 635, 637, 640, 641, 643, 644, 645, 646, 647, 648, 649, 650, 652, 654, 656, 658, 660, 664, 665, 670, 673, 677, 678, 679, 680, 682, 683, 686, 687, 688, 689, 690, 692)

**# Imaging Biomarkers**

L\_caudate\_ComputeArea <- rpois(NumSubj, 600)

L\_caudate\_Volume <- rpois(NumSubj, 800)

R\_caudate\_ComputeArea <- rpois(NumSubj, 893)

R\_caudate\_Volume <- rpois(NumSubj, 1000)  
L\_putamen\_ComputeArea <- rpois(NumSubj, 900)  
L\_putamen\_Volume <- rpois(NumSubj, 1400)  
R\_putamen\_ComputeArea <- rpois(NumSubj, 1300)  
R\_putamen\_Volume <- rpois(NumSubj, 3000)  
L\_hippocampus\_ComputeArea <- rpois(NumSubj, 1300)  
L\_hippocampus\_Volume <- rpois(NumSubj, 3200)  
R\_hippocampus\_ComputeArea <- rpois(NumSubj, 1500)  
R\_hippocampus\_Volume <- rpois(NumSubj, 3800)  
cerebellum\_ComputeArea <- rpois(NumSubj, 16700)  
cerebellum\_Volume <- rpois(NumSubj, 14000)  
L\_lingual\_gyrus\_ComputeArea <- rpois(NumSubj, 3300)  
L\_lingual\_gyrus\_Volume <- rpois(NumSubj, 11000)  
R\_lingual\_gyrus\_ComputeArea <- rpois(NumSubj, 3300)  
R\_lingual\_gyrus\_Volume <- rpois(NumSubj, 12000)  
L\_fusiform\_gyrus\_ComputeArea <- rpois(NumSubj, 3600)  
L\_fusiform\_gyrus\_Volume <- rpois(NumSubj, 11000)  
R\_fusiform\_gyrus\_ComputeArea <- rpois(NumSubj, 3300)  
R\_fusiform\_gyrus\_Volume <- rpois(NumSubj, 10000)

**# Demographics variables**

Sex <- ifelse(runif(NumSubj)<.5,0,1)

Weight <- as.integer(rnorm(NumSubj, 80,10))

Age <- as.integer(rnorm(NumSubj, 62,10))

**# Diagnosis**:

Dx <- c(rep("PD", 100), rep("HC", 100), rep("SWEDD", 82))

**# Genetics**

chr12\_rs34637584\_GT <- c(ifelse(runif(100)<.3,0,1), ifelse(runif(100)<.6,0,1), ifelse(runif(82)<.4,0,1)) # NumSubj Bernoulli trials

chr17\_rs11868035\_GT <- c(ifelse(runif(100)<.7,0,1), ifelse(runif(100)<.4,0,1), ifelse(runif(82)<.5,0,1)) # NumSubj Bernoulli trials

**# Clinical** # rpois(NumSubj, 15) + rpois(NumSubj, 6)

UPDRS\_part\_I <- c( ifelse(runif(100)<.7,0,1)+ifelse(runif(100)<.7,0,1),

ifelse(runif(100)<.6,0,1)+ ifelse(runif(100)<.6,0,1),

ifelse(runif(82)<.4,0,1)+ ifelse(runif(82)<.4,0,1) )

UPDRS\_part\_II <- c(sample.int(20, 100, replace=T), sample.int(14, 100, replace=T),

sample.int(18, 82, replace=T) )

UPDRS\_part\_III <- c(sample.int(30, 100, replace=T), sample.int(20, 100, replace=T),

sample.int(25, 82, replace=T) )

**# Time**: VisitTime – done automatically below in aggregator

**# Data (putting all components together)**

sim\_PD\_Data <- cbind(

rep(Cases, each= NumTime), # Cases

rep(L\_caudate\_ComputeArea, each= NumTime), # Imaging

rep(L\_caudate\_Volume, each= NumTime),

rep(R\_caudate\_ComputeArea, each= NumTime),  
 rep(R\_caudate\_Volume, each= NumTime),  
 rep(L\_putamen\_ComputeArea, each= NumTime),  
 rep(L\_putamen\_Volume, each= NumTime),  
 rep(R\_putamen\_ComputeArea, each= NumTime),  
 rep(R\_putamen\_Volume, each= NumTime),  
 rep(L\_hippocampus\_ComputeArea, each= NumTime),  
 rep(L\_hippocampus\_Volume, each= NumTime),  
 rep(R\_hippocampus\_ComputeArea, each= NumTime),  
 rep(R\_hippocampus\_Volume, each= NumTime),  
 rep(cerebellum\_ComputeArea, each= NumTime),  
 rep(cerebellum\_Volume, each= NumTime),  
 rep(L\_lingual\_gyrus\_ComputeArea, each= NumTime),  
 rep(L\_lingual\_gyrus\_Volume, each= NumTime),  
 rep(R\_lingual\_gyrus\_ComputeArea, each= NumTime),  
 rep(R\_lingual\_gyrus\_Volume, each= NumTime),  
 rep(L\_fusiform\_gyrus\_ComputeArea, each= NumTime),  
 rep(L\_fusiform\_gyrus\_Volume, each= NumTime),  
 rep(R\_fusiform\_gyrus\_ComputeArea, each= NumTime),  
 rep(R\_fusiform\_gyrus\_Volume, each= NumTime),

rep(Sex, each= NumTime), # Demographics

rep(Weight, each= NumTime),

rep(Age, each= NumTime),

rep(Dx, each= NumTime), # Dx

rep(chr12\_rs34637584\_GT, each= NumTime), # Genetics

rep(chr17\_rs11868035\_GT, each= NumTime),

rep(UPDRS\_part\_I, each= NumTime), # Clinical

rep(UPDRS\_part\_II, each= NumTime),

rep(UPDRS\_part\_III, each= NumTime),

rep(c(0,6,12,18),NumSubj) # Time

)

**# Assign the column names**

colnames(sim\_PD\_Data) <- c(

"Cases",

"L\_caudate\_ComputeArea",

"L\_caudate\_Volume",

"R\_caudate\_ComputeArea",

"R\_caudate\_Volume",

"L\_putamen\_ComputeArea",

"L\_putamen\_Volume",

"R\_putamen\_ComputeArea",

"R\_putamen\_Volume",

"L\_hippocampus\_ComputeArea",

"L\_hippocampus\_Volume",

"R\_hippocampus\_ComputeArea",

"R\_hippocampus\_Volume",

"cerebellum\_ComputeArea",

"cerebellum\_Volume",

"L\_lingual\_gyrus\_ComputeArea",

"L\_lingual\_gyrus\_Volume",

"R\_lingual\_gyrus\_ComputeArea",

"R\_lingual\_gyrus\_Volume",

"L\_fusiform\_gyrus\_ComputeArea",

"L\_fusiform\_gyrus\_Volume",

"R\_fusiform\_gyrus\_ComputeArea",

"R\_fusiform\_gyrus\_Volume",

"Sex", "Weight", "Age",

"Dx", "chr12\_rs34637584\_GT", "chr17\_rs11868035\_GT",

"UPDRS\_part\_I", "UPDRS\_part\_II", "UPDRS\_part\_III",

"Time")

**# some QC**

summary(sim\_PD\_Data)

dim(sim\_PD\_Data)

head(sim\_PD\_Data)

**# Write out (save) the result to a file that can be shared**

write.table(sim\_PD\_Data, "C:\\Users\\Dinov\\Desktop\\data.csv", sep=",", row.names=FALSE, col.names=TRUE)

The transposed matrix looks like this:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cases | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 6 | 6 |
| L\_caudate\_ComputeArea | 597 | 597 | 597 | 597 | 604 | 604 | 604 | 604 | 580 | .. |
| L\_caudate\_Volume | 767 | 767 | 767 | 767 | 873 | 873 | 873 | 873 | 797 | .. |
| R\_caudate\_ComputeArea | 855 | 855 | 855 | 855 | 935 | 935 | 935 | 935 | 919 | .. |
| R\_caudate\_Volume | 968 | 968 | 968 | 968 | 1043 | 1043 | 1043 | 1043 | 1023 | .. |
| L\_putamen\_ComputeArea | 842 | 842 | 842 | 842 | 892 | 892 | 892 | 892 | 908 | .. |
| L\_putamen\_Volume | 1357 | 1357 | 1357 | 1357 | 1366 | 1366 | 1366 | 1366 | 1415 | .. |
| R\_putamen\_ComputeArea | 1285 | 1285 | 1285 | 1285 | 1305 | 1305 | 1305 | 1305 | 1264 | .. |
| R\_putamen\_Volume | 3052 | 3052 | 3052 | 3052 | 2920 | 2920 | 2920 | 2920 | 2995 | .. |
| L\_hippocampus\_ComputeArea | 1306 | 1306 | 1306 | 1306 | 1292 | 1292 | 1292 | 1292 | 1313 | .. |
| L\_hippocampus\_Volume | 3238 | 3238 | 3238 | 3238 | 3079 | 3079 | 3079 | 3079 | 3227 | .. |
| R\_hippocampus\_ComputeArea | 1513 | 1513 | 1513 | 1513 | 1516 | 1516 | 1516 | 1516 | 1541 | .. |
| R\_hippocampus\_Volume | 3759 | 3759 | 3759 | 3759 | 3827 | 3827 | 3827 | 3827 | 3791 | .. |
| cerebellum\_ComputeArea | 16845 | 16845 | 16845 | 16845 | 16698 | 16698 | 16698 | 16698 | 16480 | .. |
| cerebellum\_Volume | 13949 | 13949 | 13949 | 13949 | 14076 | 14076 | 14076 | 14076 | 13992 | .. |
| L\_lingual\_gyrus\_ComputeArea | 3268 | 3268 | 3268 | 3268 | 3243 | 3243 | 3243 | 3243 | 3331 | .. |
| L\_lingual\_gyrus\_Volume | 11130 | 11130 | 11130 | 11130 | 11033 | 11033 | 11033 | 11033 | 11093 | .. |
| R\_lingual\_gyrus\_ComputeArea | 3294 | 3294 | 3294 | 3294 | 3190 | 3190 | 3190 | 3190 | 3407 | .. |
| R\_lingual\_gyrus\_Volume | 12221 | 12221 | 12221 | 12221 | 12187 | 12187 | 12187 | 12187 | 12062 | .. |
| L\_fusiform\_gyrus\_ComputeArea | 3625 | 3625 | 3625 | 3625 | 3631 | 3631 | 3631 | 3631 | 3520 | .. |
| L\_fusiform\_gyrus\_Volume | 11087 | 11087 | 11087 | 11087 | 11116 | 11116 | 11116 | 11116 | 10890 | .. |
| R\_fusiform\_gyrus\_ComputeArea | 3232 | 3232 | 3232 | 3232 | 3302 | 3302 | 3302 | 3302 | 3328 | .. |
| R\_fusiform\_gyrus\_Volume | 10122 | 10122 | 10122 | 10122 | 10162 | 10162 | 10162 | 10162 | 9884 | .. |
| Sex | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | .. |
| Weight | 84 | 84 | 84 | 84 | 97 | 97 | 97 | 97 | 96 | .. |
| Age | 67 | 67 | 67 | 67 | 39 | 39 | 39 | 39 | 54 | .. |
| Dx | PD | PD | PD | PD | PD | PD | PD | PD | PD | .. |
| chr12\_rs34637584\_GT | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | .. |
| chr17\_rs11868035\_GT | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | .. |
| UPDRS\_part\_I | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | .. |
| UPDRS\_part\_II | 12 | 12 | 12 | 12 | 19 | 19 | 19 | 19 | 15 | .. |
| UPDRS\_part\_III | 1 | 1 | 1 | 1 | 22 | 22 | 22 | 22 | 19 | .. |
| Time | 0 | 6 | 12 | 18 | 0 | 6 | 12 | 18 | 0 | .. |