# Spin #3 Retrospective Document

**Describe choice of descriptive statistics and visualization (EDA) based on variable's scales of measurement & audience expectations.**

Descriptive statistics will be used to examine general trends, patterns, and correlations in the data. Visualization tools such as histograms, scatterplots, and correlation maps and plots are used to identify patterns, validate hypotheses from the literature review, and ensure data quality.

**Discuss any potential data bias and other constraints**

The main bias in the data arises from the gender-specific nature of the MrOS dataset, which only includes male participants. Hence, the results and predictive model can't be directly extrapolated to women. Another potential issue is the presence of null values, which must be properly handled.

**Discuss "big data" issues related to volume/variety/veracity and planned solutions for remedy**

The data was collected from multiple clinical sites, over time, and across diverse test methods, adding complexity to the data analysis. The data is stored in a Postgres database, and the team plans to continuously refine the database structure as EDA proceeds and new supplemental datasets may be needed.

One issue is the overall data is split into numerous sets based on visit and measurements collected. We have split those sets even further by form for EDA purposes. We will need to merge the data back together appropriately by participant ID.

**Discuss plan for additional data, data shaping and carpentry**.

Working on analyzing the forms used during the clinical interviews to appropriately handle all missing values. Many forms have a lot of missing values, but not missing information. The missing values often correspond to chained questions which have to be appropriately dealt with. This may not be of large concern for modeling purposes, as we plan to use models that can handle robust missing values, but for EDA purposes we are analyzing the actual clinical forms to gain insight on missing value reasons.

Renaming columns for interpretability.

One-hot encoding some categorical columns for data analysis.

**Discuss changes regarding data storage location structure, use of database, and data curation and management strategy**

The raw data is initially stored in the team’s shared folder. It's then uploaded to a Postgres database for further processing, analysis, and storage. This week, raw data file V1 has been split into separate forms by category.

**Discuss changes related to data provenance and assessment**

The MrOS (Osteoporotic Fractures in Men) dataset is a comprehensive and detailed database specifically gathered to understand the factors influencing bone health in older men. This dataset was originally collected through a research study conducted from 2000-2002 at six clinical sites across the United States, and its primary aim was to determine how various lifestyle factors, diseases, and health histories impact bone health, particularly the risk of fractures in men. The V1 dataset within the MrOS dataset was subsetted into csv forms and added to the team folder for team members to analyze.

**Discuss changes to data acquisition and characteristics for data sets used in project**

The raw data file V1 has been split into separate forms for further analysis by group members. The V1 data file looks at current and past health history, and lifestyle factors, which have each been further broken down into sub-categories in forms so that team members can assess these factors independently.

**Discuss changes to scope and description including analytic type (e.g. descriptive, diagnostic, predictive, prescriptive)**

We are currently working to compile concise descriptive health data with enough data points to be used for predictive analysis when assessed with other data sets which we will analyze in the near future. This week, our team analyzed the V1 dataset to determine the prevalence of current health, health history and lifestyle factors. The V1 dataset was divided into forms so that the time needed and size of the data sections was more manageable. The more invasive analysis that we are performing this week will permit us to home in on the most relevant factors pertaining to our final story.

**Discuss changes to research/domain question(s) to answer or problem(s) to solve**

What specific current and past health-related issues or lifestyle factors are related to low bone mineral density or increased fracture risk?

**Discuss changes to statement of project goals and measurable objectives, constraints**

*Project Goals:*

Further refine the identity of lifestyle factors, prevalent diseases, and aspects of health history that significantly contribute to low bone mineral density and the subsequent risk of fractures. The goal of this project is to achieve high accuracy, precision, and recall in predicting fractures.

Our goal is to create a machine learning algorithm that can accurately predict fracture risk based on health and lifestyle factors. To reach this goal, we first need to determine which of these factors are the most common .

Measureable Objectives:

First we will narrow down the lifestyle factors, diseases, and health history aspects most commonly affecting patients in our dataset. Work will be divided equally among team members. Frequent communication on slack will allow each team member to provide/recieve feedback to make sure we are on track with one another.

Constraints:

The project relies heavily on the MrOS dataset which exclusively contains male records, meaning our findings cannot be extrapolated to females.

Another constraint in this dataset is the frequency or missing or incomplete data.

**Discuss changes to Relevant Key Performance Indicators**

1. Bone Mineral Density (BMD) measurements:
   * Average BMD: Calculated as the mean BMD across all participants, providing an overall measure of bone health.
   * BMD by age group: Examining BMD variations across different age groups to understand age-related changes.
   * BMD by site (e.g., hip, spine, wrist): Analyzing BMD variations at specific bone sites to assess regional differences.
2. Fracture Incidence and Risk:
   * Fracture rate: Calculating the number of fractures per participant or per unit of time, providing an indication of fracture occurrence.
   * Fracture risk score: Utilizing a validated risk scoring system to assess the likelihood of experiencing a fracture based on various factors.
3. Lifestyle Factors:
   * Exercise level: Measuring the frequency, duration, and intensity of exercise to determine its impact on BMD and fracture risk.
   * Diet quality: Assessing the adherence to a bone-healthy diet, such as one rich in calcium, vitamin D, and other nutrients.
   * Alcohol intake: Investigating the relationship between alcohol consumption and BMD/fracture risk, considering both moderate and excessive intake.

KPI’s affecting the inclusion into the V1 dataset:

* ability to walk without the assistance of another
* absence of bilateral hip replacements
* ability to provide self-reported data
* residence near a clinical site for the duration of the study
* absence of a medical condition that would result in imminent death
* ability to understand and sign an informed consent

**Discuss changes to vision for final data story including intended audience and format/style medium of data story/product**

As we progress through the data, our final story is changing through the elimination of health and lifestyle factors shown to be irrelevant or without enough collected data to contribute.

**Weekly individual team member accountability/contribution assessment/evaluation included**

Josh

* Initial review of forms
* Creation of CSV files from forms
* Started and completed preliminary analysis on all assigned (2) Forms
* Contributed to SpIn #3 Retrospective

David

* Uploaded Form tables to Postgres
* Started and completed preliminary analysis on all assigned (2) V1 Forms
* Created “Start Here” notebook
* Contributed to SpIn #3 Retrospective

Tyler

* Completed first pass updates on SpIn #3 Retrospective
* Review of annotated forms used for clinical interview
* Started and completed preliminary analysis on all assigned (4) V1 Forms

Karen

* Added majority of content to SpIn #3 Retrospective
* Started and completed preliminary analysis on all assigned (3) V1 Forms
* Continued literature review efforts

**Project work success status evaluated and future work/tasks discussed**

Our goal for this week was to divide and conquer across the various Forms within the V1 dataset that were relevant. We all had an opportunity to retrieve the data, review the shape of the data, understand the completeness of it (how many NAs) and begin analysis. We met this goal and are happy with our progress.

During the next week, we’ll identify the best way to use the insights we’ve gained from this analysis to continue solving our domain questions. Upcoming work for this week includes EDA on endpoint data with a goal of beginning work on principal component analysis, feature selection, and multivariate analysis between target variable and baseline variables. Getting these tasks completed would allow us to continue on to model building.

**.Jupyter Notebook(s) fully internally documented**

Yes

**All paths to data files map to DSA team shared folders**

Yes

**Current project SpIn artifacts (notebooks) are located in the TeamArtifacts\SpIn\_3\_Artifacts folder (provide link (within Europa) to the first notebook in the pipeline)**

Yes

<https://europa.dsa.missouri.edu/user/ejm301/notebooks/su23CaseStudy_Team03/TeamArtifacts/SpIn_3_Artifacts/SpIn3-StartHere.ipynb>

**Jupyter Notebook(s) execute without exceptions**

Yes – data is retrieved from the database for all workbooks. Executing workbooks will require a valid username/password but ETL can be completed without any impact to subsequent EDA as long as no table locks prevent data insertion.

**Link to Mentor recorded mentor meeting and key meeting takeaways provided**

Our mentor was out of town this week and had no internet access so we do not have a recording with our mentor for this week. However, we did meet as a group at our regular time at which point, we discussed our progress and future direction for the week.

Weekly meeting notes: [2023-06-21 DSA Weekly Meeting](onenote:https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/Documents/SU23_DSA8080%20Casestudy/SU23_DSA8080%20Casestudy/Group%20Meetings.one#2023-06-21%20DSA%20Weekly%20Meeting&section-id={22F8CF55-8C19-4BDF-B116-5B29E397A512}&page-id={EC92D5DC-BB38-419E-8A2E-0C785AAB90C4}&end) ([Web view](https://mailmissouri-my.sharepoint.com/personal/jwj8c8_umsystem_edu1/_layouts/OneNote.aspx?id=%2Fpersonal%2Fjwj8c8_umsystem_edu1%2FDocuments%2FSU23_DSA8080%20Casestudy%2FSU23_DSA8080%20Casestudy&wd=target%28Group%20Meetings.one%7C22F8CF55-8C19-4BDF-B116-5B29E397A512%2F2023-06-21%20DSA%20Weekly%20Meeting%7CEC92D5DC-BB38-419E-8A2E-0C785AAB90C4%2F%29))