The Math Behind the Mass: An Investigation of Obesity Dynamics

Team #16926
Modeling Megaminds
Modeling The Future Challenge 2024

March 3, 2024

Executive Summary

15 years ago, the global obesity rate was 23.9%. Today, the global obesity rate has risen to over 39%, affecting nearly 3.12 billion people worldwide[1]. This unprecedented growth in obesity rates shows its impacts in the growth of various risks. In the United States (US), for example, obesity-related cardiovascular disease deaths tripled from 2.2 per 100,000 to 6.6 per 100,000 between 1999 and 2020[2].While obesity is not the only cause behind these life-threatening diseases, it is still a substantial factor that significantly increases their prevalence in the world[3][4][5][6]. Our model (within the scope of the US) aims to identify the main causes of obesity, predict future trends of obesity in the US, and propose viable recommendations to address the problem.

To start, we first defined obesity as having a BMI of over 30[7]. We then used datasets from the CDC - most notably from the NHANES survey which combines questionnaires and physical examinations of a nationally representative sample. From our collected data, we divided our sets into 3 categories: demographics, bodily chemical makeup, and consumer behavior. In order to determine the relative importance of each cause of obesity, we utilized a random forest model to run a factor analysis on our data, creating a threshold model and a variation on the GAIN model to impute empty cells and clean/adjust our data. We discovered that increased age (demographic), increased c-reactive protein(CRP) level(body), decreased blood lead level(BLL)(body), increased red blood cell distribution width(RDW)(body), increased food stamp usage (consumer), and decreased heroin use (consumer) all contributed to an increase in BMI - an indication of obesity.

Moving forward with our Markov Chain analysis of obesity trends, we had to convert our cross-sectional NHANES data to a longitudinal subset. To do so, we utilized regressions to find how our important factors affected BMI, and utilized this to match people across our year datasets to find people who took the survey each year. With this longitudinal data, we ran the Markov model to obtain probabilities of shifting from various weight levels each year, thus creating our transition matrix. Employing this transitional matrix, we revealed the prevalence of obesity risk and model potential obesity counts in the following years. From this, we learned that if no changes are made, the US normal weight, overweight, and underweight population will decrease, while we predict for in 2067 (50 years in the future), the percentage of the US population that will be class I, II, or III obese will be 42.1%. Severe risks of obesity are known, such as cardiovascular disease. This paper highlights a method for hidden risks on our healthcare system, like obesity induced-cancer which contributes over $249 million of preventable costs.

To help mitigate the above risks and reduce the impacts of our most important factors, we outlined the following recommendations. For age, we simply recommend to follow the rest of our recommendations more closely. For CRP, we recommended a diet of anti-inflammatory foods, the use of various medications, and taking on activities that reduce stress. For BLL, we found that the potential obesity-related benefits were far outweighed by the risks of lead poisoning, and thus do not recommend considering this factor. For RDW, we recommend diets targeting iron, folate, and vitamin B12. For food stamps, we recommend implementing insurance policies, subsidizing food stamp programs, and mandating such programs to provide more nutritious foods. For heroin usage, while we cannot recommend the use of heroin due to its various risks, we instead recommend the use of medical marijuana. A common recommendation many factors shared was stopping alcohol use and smoking, and increased exercise.

Contents

1 Background

In 2013, 52-year-old John Alleman, the mascot of the Heart Attack Grill in Las Vegas, NV, died due to a heart attack [8]. While a devastating loss, its irony reveals a stark truth about America's eating habits: morbid. Despite numerous consumer deaths, Americans have championed the restaurant for its all-you-can-eat burgers, fries, milkshakes, and other calorie-dense items. The restaurant's target audience? Those who are already overweight: customers over 350 pounds eat for free. Its hospital theme mocks the American healthcare system in which treating obesity case-by-case is deemed more effective than devising a long-term solution to the rising epidemic. Fast food restaurants like Heart Attack Grill have skyrocketed in terms of popularity over the last decade. Offering a cheap and accessible source of calories, the fast food industry is expanding rapidly in this era of urbanization, hectic lifestyles, and a preference for convenience over nutrition. In fact, it is expected to grow at a compound annual growth rate of 5.0% from 2022 to 2029 [9], adding more victims to the obesity epidemic. However, fast food is just one component of the obesity equation. There are numerous indicators depicting just how vast this health crisis is and thus, it is imperative to uncover and tackle root issues immediately in order to halt the growth of obesity.

This is a problem affecting the whole world, causing ripples throughout not just the food industry, but the healthcare and service industries as well. Muti-billion dollar industries in the food and agricultural sectors are primarily impacted as they are the ones producing and supplying the malnourished supplements. The healthcare industry, meanwhile, suffers the brunt of the losses due to the risks of obesity as they seek to treat the temporary and chronic impacts of obesity. In the middle, the service industry is also scrutinized as they are responsible for the sale and distribution of foods. Right now, about 3.12 billion people (39% of the global population) have obesity. But 15 years ago, in 2008, the global obesity rate was only 23.9%, affecting 1.63 billion people [1]. Despite obesity being a global phenomenon, we will focus our paper on its impacts solely within the United States of America (US), where similar trends still show at a more manageable scale.

Obesity can be defined as having a Body Mass Index (BMI) of 30.0 or greater[7]. Obesity carries with it many risks that are necessary to mitigate. Excess weight or obesity boosts risk of death by anywhere from 22% to 91%[3]. This is the main risk of obesity - the high chances of developing life-threatening diseases such as diabetes, cancer, strokes, osteoarthritis, and many heart diseases. The number of U.S. adults who died of heart disease and whose death record cited obesity as a contributing factor was three times greater in 2020 than in 1999[4]. Even if one does not develop these mortal diseases, obesity can still cause many other uncomfortable health complications from sleep apnea to high blood pressure to even fertility problems that affect one's daily lifestyle[5]. Not only that, but these complications also often require large amounts of money to treat, accounting for increased burden. The direct and indirect medical costs attributable to obesity totaled $1.4 trillion in 2014, and appears to be on the rise[6]. The financial and health burdens one encounters due to obesity are immense, accounting for the great risks that should be mitigated.

We identify a decrease in food nutrition, driven by the widespread consumption of processed and low-nutrient-density foods, as a significant contributor to escalating obesity rates. A 2004 US

study found important nutrients in some garden crops (e.g. asparagus and spinach) are up to 38% lower than they were at the middle of the 1900s[10]. This is aided by the increased contamination in our food from things like pesticides and fertilizers that lead to a copious amount harmful chemicals in our bodies over the years, contributing to countless health complications. [11] Unhealthy dietary habits are also a focal point, demanding attention to reverse the adverse impact of poor nutritional choices. These habits may arise from a multitude of reasons, from a lack of proper nutritional and health education to the copious amounts of money spent on the food advertising agency. Not only that, a significantly higher proportion of ultra-processed food advertisements out of total food advertisements was identified in the low socioeconomic area[12]. This hints at the main targets of the cheap, processed food businesses: low income people who may find it difficult to purchase fresh, yet expensive, foods. A lack of exercise in our more sedentary lifestyles may also contribute to greater obesity rates, as do the increased stressors present in our world that lead to a lack of sleep[13][14].

Although the rate of increase in obesity seems to be declining in most high-income countries, it continues to rise in many low-income and middle-income countries and prevalence remains high globally[15]. This same pattern can be seen within the U.S., where areas with high poverty and low access to nutritional foods experience high levels of obesity. Genetics also play a substantial role in determining one's risk of developing obesity throughout their lives[16], leading to concerns about the plausibility of preventing obesity in all people. And even if these are the people most at risk from obesity, all people have some non-zero probability of becoming obesity, making this a problem that impacts everyone.

Some groundwork has been already paved for us to follow in the form of policies and solutions. Current strategies include improving the coverage of existing health and medical insurance, the alteration of in-school lunch policies and nutritional education to be more modern and encompassing, increasing the availability of gyms/exercise facilities, and implementing more genetically modified organisms (GMOs) and other agricultural techniques to improve the nutritional value of foods[17][18][19]. Researching the pros and cons of these policies will allow us to better target our risk analysis and recommendations to mitigate the risks of obesity.

In this paper, we seek to analyze the root causes of obesity to determine the most important factors that have led to this crisis, as this would allow us to create better targeted recommendations for the industries involved as well as the individuals at risk. We then will model the trend of obesity rates in the U.S. to determine the scope and extent of this crisis. Finally, we will use the results obtained from our models to create meaningful recommendations to mitigate the extensive, and often life-threatening, risks associated with obesity.

2 Data Methodology

Financial Burden of Cancer Care

Description and Type of Data: This dataset estimates national expenditures for cancer care by cancer site and year, essentially describing the financial burden of cancer.
Cleaning or Adjustments needed: No cleaning is needed as the data is already presented in a clean manner without any missing points.
Credibility of Data: This data comes from the National Cancer Institute, which is a reputable source from the US government.

Probability of Cancer Onset

Description and Type of Data: The dataset is a compilation of various cancers and the probability of their onset for the average person. *This data helps to quantify the frequency of risks so that we can compare the risk of a non-obese person to an obese person.*
Cleaning or Adjustments needed: The data is reported in cases of cancer per 100,000 people every year. We automatically computed the probability to scale them for the whole population.
Credibility of Data: The data is obtained by the CDC which is operated by the US Government.

CDC Obesity Rates by Age, Gender, Race, Education, Income and State

Description and Type of Data: Between 2011 and 2022, the CDC compiled the percentages of obese residents in each state, under the following definition of obese:

1. Over 18
2. BMI>30.0

Respondents of under 50 pounds, over 650 pounds, under 3 feet, and over 8 feet, were excluded. The data provided is under a range in the legend, leaving room for some vagueness in the exact percentages. Broken down by state, all areas have a value except for VI, which has unavailable data. *Additionally, due to the longitudinal nature of the study, this data is best used for projecting future trends.*
Cleaning or Adjustments needed: The data currently is in already complete, and merely needs transferring to excel spreadsheets.
Credibility of Data: This data of the US obesity rates is highly credible, as it is published by the Centers for Disease Control and Prevention, a US government run organization.

Obesity by State

Description and Type of Data: The data provided documents the percentage of residents in each state that are obese. The definition used for obese is not present. The data is updated, showing only the 2022 data. *Due to the cross-sectional nature of the data, this information is best used for running factor analysis in the US and specific states by separating outcomes.*
Cleaning or Adjustments needed: Cleaning is needed to compile the information, as it is in graphical depiction at the moment.
Credibility of Data: This data is credible, as it is sourced from data.gov, a US government based agency that collects reliable data.

National Health and Nutrition Examination Survey

Description and Type of Data: This dataset has information regarding the health and nutritional status of adults and children in the United States. The survey combines interviews and physical examinations in order to obtain the data. About 12,000 persons per 2-year cycle were asked to participate in NHANES. Response rates varied by year, but an average of 10,500 persons out of the initial 12,000 agreed to complete a household interview. Of these, about 10,000 then participated in data collection at the MEC. The survey includes demographic, socioeconomic, dietary, and health-related data, which can be used to determine the prevalence of major diseases and risk factors for diseases. *Not only do the number of factors help with separating outcomes, the longitudinal study helps to define historical trends and the sufficiently large database contributes to defining severity.*

Cleaning or Adjustments needed: Cleaning is required to compile all the relevant data into one neat spreadsheet in order to analyze, as currently the data is separated by year and combined with many data points irrelevant to our paper. Additionally, NHANES also conducted their own methodologies to ensure accurate data[20]. Firstly, they assigned weights to each individual according to their demographics in order to obtain a sample that more closely represented the entire US to prevent oversampling bias. They also ensured all conditions (from interviewing to laboratory testing) was standardized all around. Physical and Laboratory prescreenings are also conducted to remove potential outliers from the dataset and ensure standardized environments (for example, participants are required to fast before the laboratory tests).

Credibility of Data: This data is highly credible, as it comes from a CDC sponsored survey which surveys a nationally representative sample in mobile units. NHANES also follows all ethical guidelines for conducting surveys from informed consent to right to privacy, ensuring accurate data.

Note that as our data cleaning was dependent on the model we selected and required separate mathematical modeling, we will include it under the Math Methodology section.

3 Mathematical Methodology

3.1 Assumptions

1. **BMI is an accurate measurement of obesity.** The CDC (Center for Disease Control and Prevention) uses a Body Mass Index (BMI) to describe the body fat and health of an individual. As several reputable research centers and government organizations use BMI, it is a reliable measurement that we use in our paper as well.
2. **Obesity can be defined as having a Body Mass Index (BMI) of 30.0 or greater**[7]. It is necessary to have an official definition of what classifies as "obese" in order to conduct research. As such, we took to the most credible source in the U.S. regarding this subject - the CDC.
3. **If the ratio of factors vs. BMI matches between two individuals in differing year sets, among all identified important factors, they are the same person.** From the NHANES survey procedures, we are aware that many people are taking the survey each year, meaning correspondence is present throughout the yearly datasets. Due to human uniqueness, each important factor will effect BMI in a unique way. Thus, if 2 people in 2 differing year sets have identical factor ratios for each important factor, we assume that they are likely the same person due to the low probability of 2 people having extremely near identical bodily percentage compositions.
4. **Outliers can be eliminated in linear regression.** While NHANES data collection procedure did involve removing obvious outliers from fake responses, we recognize that the presence of any outliers significantly skews the data and lowers the r^2 value. Therefore, we underwent a more thorough outlier identification process by looking at points over 2 standard deviations away from the mean. While this removes 5% of data points, even if all of them were real points, being less than 0.05 proportion of our data set allows us to remove them safely without statistically significantly altering our results.
5. **The feature importance of various factors towards obesity does not change throughout the last 10 years.** Over the course of hundreds or even thousands of years, it is justifiable that the influence of various bodily factors will begin to affect humans differently, but due to the recent nature of our data, being the last 10 years, and the fact that our prediction will only look towards the next 10-25 years, we can assume that human biology will not change significantly and will not necessitate incorporation into our model.
6. **The sample present in the NHANES dataset is representative of the entire US population.** The NHANES survey methodology is noted in Data Methodology, and thus we assume that the proportions present in the sample can be extended and generalized to the US population.

3.2 Data Cleaning

3.2.1 Threshold Model

In our project, we seek to determine the most pressing factors contributing to obesity, predict the future trends of obesity in the US, and present estimated probabilities for individuals to move from various BMI levels over a period of time.

For the primary portion of our model, we decided upon using a random forest model. We did initially consider a wider range of models, yet through the specificity of data that we had, being a set from NHANES with ¿1500 factors possibly contributing to obesity, we narrowed down our final considerations to be between Random Forest, Permutation Importance, and Gradient Boosting Machines. At this point, we did determine Random Forest to be the model most capable of handling the large dataset we were equipped with, citing the following key advantages over the other two.

Advantage 1: Random Forest is significantly less computationally intensive than Gradient Boosting, due to the combined construction of the decision trees as opposed to Gradient Boosting's sequential construction. Permutation Importance is about equally resource intensive as Random Forest, yet would likely take longer to run due to its inherent inefficiency when working with high-dimensional data.
Advantage 2: Random Forest provides the strongest available estimates regarding factor importance given our total project time frame. Compared to Permutation Importance, Random Forest provides significantly more stable estimates of factor importance, due to its direct use of feature importance in the model training process. Additionally, while Gradient Boosting undeniably could provide higher predictive accuracy, it requires significantly more tuning and resources to prevent overfitting, due to our extremely high-dimensional and noisy data. With regard to our time split between this model and our following others, we determined it to be of best judgment to pursue Random Forest.

Consequently, Random Forest was solidified as our primary model. Using our main data source, being the NHANES dataset, we decided to combine all the data for each pairing of years, from 2003-2004 all the way to the 2017-2018 set. This was done under Assumption 5, allowing us to have a larger pool of data to draw from when formulating our Random Forest model. Following this comes a systemic cleaning of the data, due to a few various issues discovered within our large dataset.

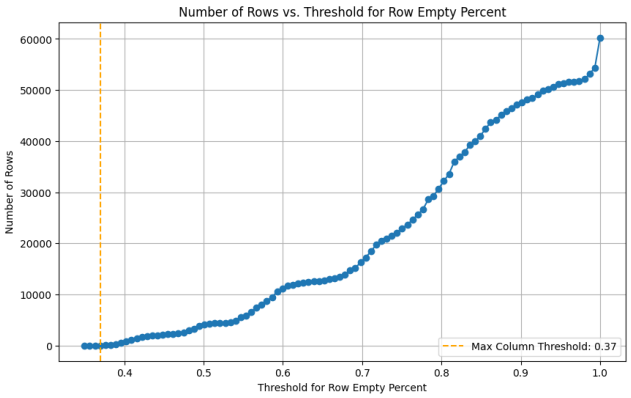
As is clearly visible upon scrolling through the data, there are quite a lot of empty cells in the datasheet, resulting from participants in the NHANES survey opting out of answering a question or participating in a laboratory test. Additionally, some of the factors are measured by a categorical encoding, where the actual value has no inherent meaning, rather only a connection to a provided table in which the numerical value typically indicates a word, such as food type. These categorically encoded factors are easy to take into consideration; due to the scarcity of them, we manually searched each file and identified the categorically encoded factors, which upon closer inspection were determined to be unnecessary to our final results. For example, one categorical variable was assigned to meal name, such as if a subject's identified caloric intake was eaten during Breakfast, Lunch, Brunch, Dinner, Almuerza, Desayuno, and various other names. To this end, we justified that the time of eating was likely minimally influential to BMI, and thereby removed it (citation needed). For the other categorical factors, we likewise pursued a similar method of researching possible influence and ruling them out. If interested, our attached full dataset contains these exact factors, which can be found when cross referencing with the NHANES variable name sheet.

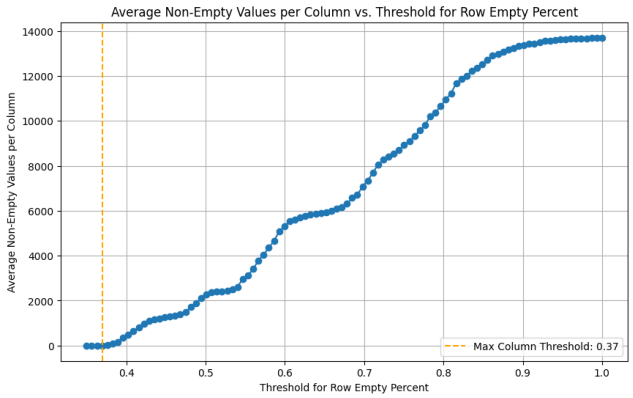
A larger issue was the blank cells due to unavailability of data in our spreadsheet. As our random forest model could not take in any blank cells, we decided upon pursuing artificially generated synthetic data to replace these empty cells, derived using a GAIN model which shall be explained shortly. Of course, the justification for employing synthetic data came through our research into the field. An MIT study concluded that 70% of synthetic data trained off portions of real data created models that were not statistically significantly different from models based off real data[21]. To this end, we concluded that upon generating synthetic data into our empty cells, 30% of this data would be incorrect and statistically distinguishable as compared to our real data. Citing the 5% rule, we determined that as long as the incorrect synthetic data in place for any one factor was under 5% of the total data points under that specific factor, we could safely conclude that the incorrect synthetic data would not be statistically significant to our final results from the Random Forest model. Running a quick calculation, we found that

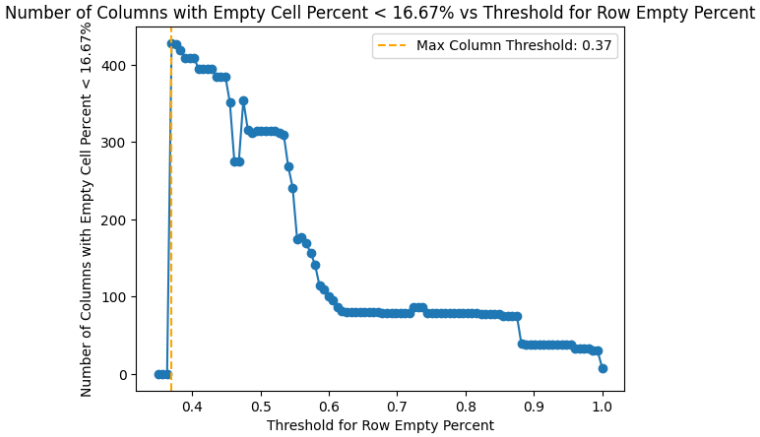
$$0.3 * S \leq 0.05S \leq 0.1666\ldots$$

This means that as long as the empty cells in one column, representing all data under one factor, is less than 16.6666...% of the total data points in that column, we could safely generate synthetic data to fill the gaps without compromising the integrity of our Random Forest model results.

However, as seen in the following graphic, the number of columns naturally under this percentage of empty cells were few. To resolve this, we looked into the dataset more closely, and realized that many of the rows consisted of mostly empty cells for each factor. Combining this observation with the NHANES survey instructions and procedures, we concluded that there was a large likelihood of many participants filling out a short piece of the full test, but refusing to continue with the rest due to lack of time. Therefore, the best path forward was to remove rows over a certain percentage of empty cells compared to the total cells in that row. Here, we define the Threshold $= T$ such that all rows with T% of their total cells empty will be eliminated from our dataset. From here, we looked to optimize this T, such that we get the maximum number of columns with less than 16.6666...% of the total data points in that column empty. Below are graphics depicting our findings.







From a glance, it is clear that the most optimal threshold without any constraints would be a value of 0.37. However, this brings to light another pressing issue; how many data points are fundamentally necessary in each column such that the Random Forest model still has sufficient data to find accurate results. Looking at employing our threshold of 0.37, note that this leaves only 8 data points in each column, which is a strikingly low number as compared to the total number of features, being 438. From here, we dive into the following calculations to determine what is the necessary amount of data points per feature in a Random Forest model to preserve its integrity.

To look into this, we analyze essentially a measure of what is necessary for a model such as Random Forest to effectively learn from the dataset provided. In a lecture from CalTech [22], it is proven that given the VC dimension of the hypothesis set, the necessary data points to let the model learn is

$$N \geq 10(VC)$$

where VC is the VC dimension of our hypothesis set, or in our specific case, the Random Forest model. To solve for this, we take a closer look at what a VC dimension is. A VC dimension, or Vapnik-Chervonenkis dimension, is a numerical value that looks at the capacity of a hypothesis class, or here our random forest model, to shatter a set of given points. Note that shattering a set of points is defined as separating perfectly any possible labeling of these points, essentially uniquely fixing each point or classifier in the hypothesis class in accordance with our labels. To find the VC dimension for our Random Forest model, we need to look at formulas for it regarding learning models, specifically of decision trees. In their textbook, Professor Shalev-Shwartz and Professor Ben-David suggest that the VC dimension of a given set H such that the set and model learns $|E|$ parameters in a decision tree can be solved as

$$O(|E| * \log_2 |E|)$$

where O denotes a function that grows up to but not faster than the interior function. [23] Of course, our Random Forest is not so simple; it consists of multiple decision trees, and we must adjust for this in accordance with the provided formula. To solve for —E—, the number of features that each decision tree learns from, recognize that for Random Forest models, each tree takes a subset of randomly chosen set from our dataset. Therefore, the by solving the VC dimension for a singular decision tree in a Random Forest, it will equal the total VC dimension for the whole Random Forest model, as each tree samples the same amount of data from the same size data set. Therefore, we merely look at how many parameters each decision tree in a Random Forest model

actually learns, which can be solved for as

$$|E| = \text{effective number of features used in each tree} = \sqrt{n}$$

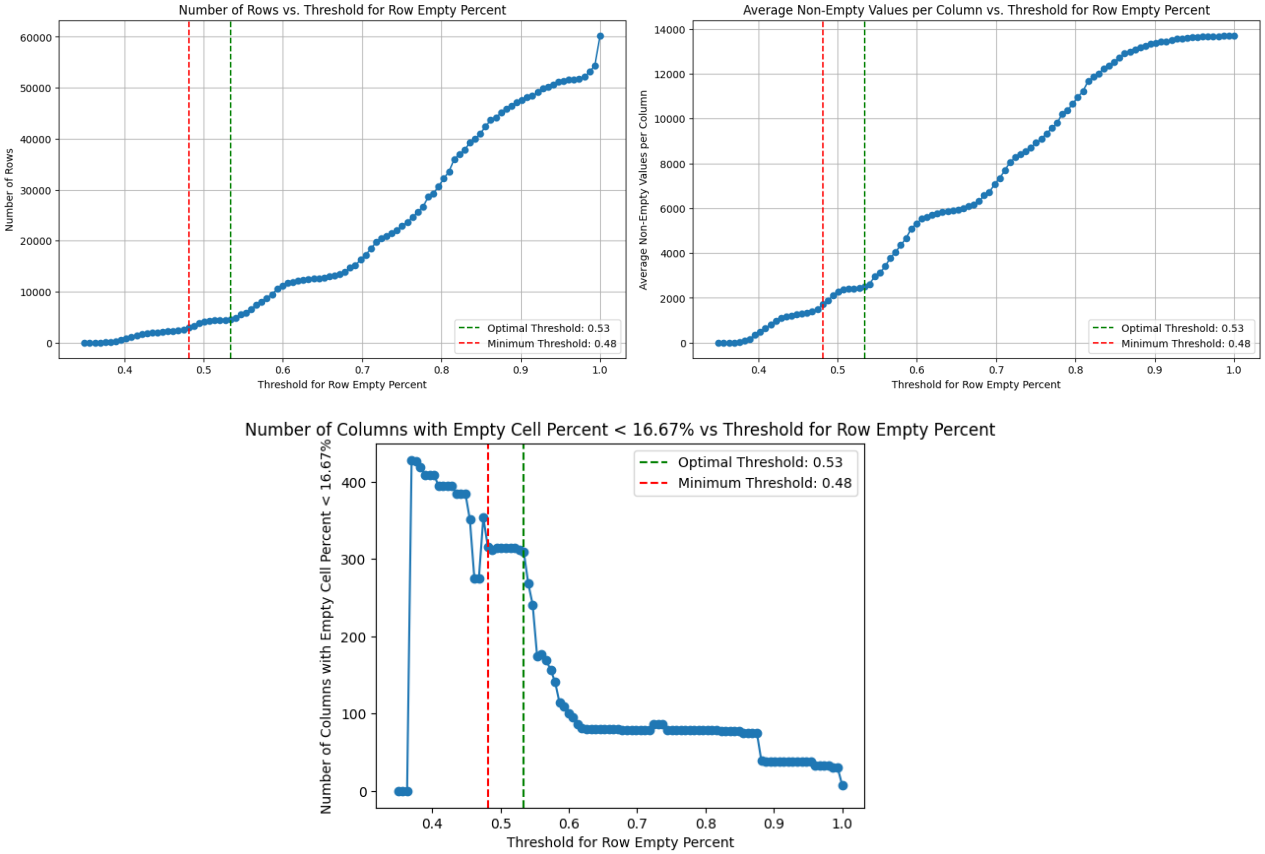
Therefore, placing this into our formula and taking the upper bound of the function O, as we are searching for minimum number of required data points to guarantee a successful Random Forest model, we find that

$$\text{VC Dimension} = O(\sqrt{900})\log_2(\sqrt{900}) = 147.206$$

From here, we merely look back to our initial formula relating VC dimension to needed data points to find that the minimum number of data points needed for each feature or column in our dataset is

$$\text{Minimum Data Points} = 10(\text{VC Dimension}) = 1472.06$$

Keep in mind that we did make a small approximation in stating that the effective features in a singular Random Forest decision tree would be root(n), as in reality it would be slightly higher as it requires a whole number of features. Therefore, our Minimum Data points would actually likely be a bit higher that 1472.06. From here, we make a slight assumption to set the Minimum Data points to be 1500, as through testing with smaller Random Forest models we found that typically it seemed the needed amount of data points was about 2% higher than the calculated value from the formula.



Going back to our graph, notice that utilizing Minimum Data points as 1500, the number of columns with under 16.666% do not change for as we shift our threshold right, all the way until the threshold is equal to 0.53, where it then begins a significant decrease. Recognizing that our minimum data point calculation is a minimum, and being comfortably over it would be preferable in case of miscalculation, we choose the higher threshold of 0.53 so as to have 2454 data points,

which is safely over the minimum we calculated. From here, we have a fully cleaned dataset, with no empty values, ready to go into the Random Forest Model. Of course, to generate the empty cells, we used the following model.

3.2.2 GAIN Model

To clean our data, we also utilized a variation of the GAIN model[24]. After our threshold model, we still have some missing data present in our dataset that require imputation. The generative adversarial imputation nets (GAIN) model we utilize is based off the GAN model and uses machine learning methods to deal with such missing data and allows our random forest to run smoothly by generating synthetic data (which is the missing value estimation). Synthetic data use was justified previously in the threshold model[21].

The GAIN model comprises a generator (G) and a discriminator (D). G receives real data vectors and fills in missing values based on observed data, generating completed vectors. D then discerns between real and synthesized elements within these completed vectors. To enhance G's learning of the desired distribution, D is aided by a hint vector conveying missing data quality information, enabling D to focus on imputation quality for specific missing values. This ensures G learns to generate data accurately. While not mandatory in GANs, convolutional neural networks (CNNs) often utilize convolution, a pivotal operation involving overlapping kernel application across data sections. The generator architecture typically includes the following layers:

1. Linear layer: The noise vector undergoes transformation in a fully connected layer, resulting in a reshaped tensor output.
2. Batch normalization layer: Normalization of inputs to zero mean and unit variance stabilizes learning, addressing issues like vanishing or exploding gradients, thereby facilitating gradient flow through the network.
3. Up sample layer: An alternative approach involving upsampling followed by a simple convolutional layer is mentioned instead of using a convolutional transpose layer for upsampling. Although convolutional transpose is still utilized in some cases.
4. Convolutional layer: Data is processed through a convolutional layer with specified parameters such as a stride of 1 and consistent padding, particularly crucial for learning from upsampled data.
5. ReLU layer: Integrated into the generator, ReLU activation aids in swift saturation and comprehensive coverage of the training distribution space.
6. TanH Activation: TanH activation contributes to faster model convergence, enhancing the speed of learning and convergence towards optimal solutions.

An image more clearly showing the process can be found in Appendix A.
With this, we created our fully cleaned dataset, and proceeded to run it in a Random Forest model.

3.3 Random Forest Factor Analysis

The first step in our model is to determine the relative importance of each cause of obesity, as this will help us better target our risk analysis and recommendations, as well as serve as the baseline for the rest of our model. As many factors play a role in the onset of obesity, in order to bring our

project down to a manageable scale, we select the following categories that we perceived as the most important and had sufficient data to focus on. In no particular order, they are:

1. Demographics
2. Consumer Behavior
3. Bodily Chemical Makeup

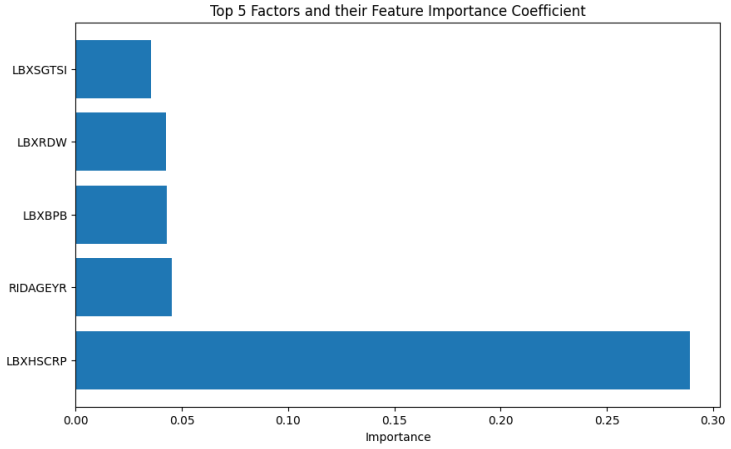
To complete this task, we choose to use a random forest model. Random forest is an ensemble learning model that uses multiple decision trees. Decision trees partition the feature space into regions until we are left with pure leaf nodes and find the best split by maximizing entropy gains. In Random Forests, each tree is trained on a random subset of the training data and a random subset of features. The number of features split is according to $\sqrt{p}$ where p is the total number of features, aiming to produce diversity among trees. Increasing the number of trees generally leads to better generalization, but comes at computational costs. Each individual decision trees contributes a vote to the overall result. The most populous vote takes precedence. This ensemble learning method reduces variance and improves performance over other learning models[25]. Employing this onto our cleaned data set, we produced the following results.

3.4 Results of Random Forest

Combining the demographic and laboratory data, we produce the following table and graph which consist of the code name used, the corresponding feature name, and the feature importance coefficient from the top 5 factors sorted in descending order by the feature importance coefficient.

Table 1: Top 5 Random Forest Factors

Code Name	Feature	Feature Importance Coefficient
LBXHSCRP	HS C-Reactive Protein (mg/L)	0.289222
RIDAGEYR	Age in Years at Screening	0.045957
LBXBPB	Blood Lead (ug/dL)	0.042760
LBXRDW	Red Cell Distribution Width (%)	0.041614
LBXSGTSI	Gamma Glutamyl Transferase (GGT) (IU/L)	0.035570



Factor 5 and beyond include very similar feature importance coefficient values, so we will only consider the top 4 factors: HS C-Reactive Protein, Age, Blood Lead, and Red Cell Distribution

Width. The low coefficient values are reasonable for such a model because we are considering a vast number of factors and obesity is caused by all kinds of bodily changes. We deduce that HS C-Reactive Protein is by far the most important factor to consider when thinking about chemical causes of obesity, because of its extraordinarily high feature importance coefficient.

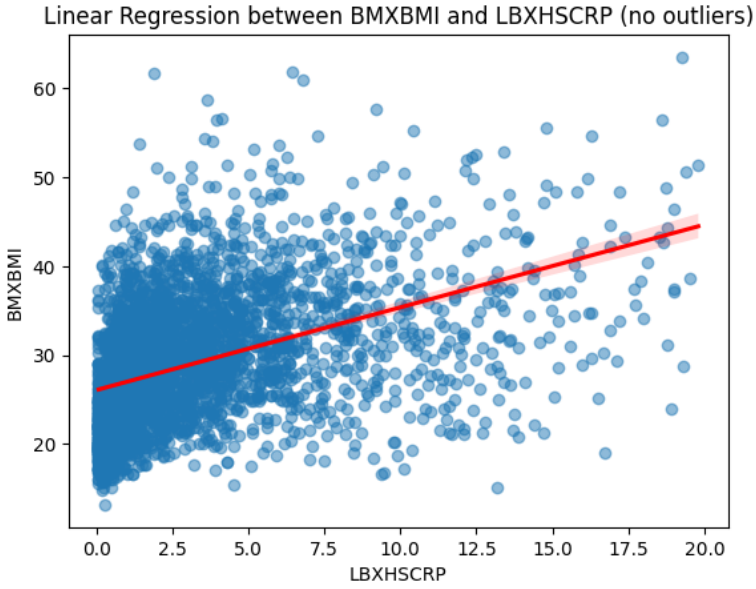
3.5 Markov Chain Prediction

The second part of our model involves predicting the probability someone in a BMI category fits into another category in a certain number of years. We define the categories as follows[26]:

Table 2: BMI Categories with State Numbers

BMI Range	**Category**	**State**
< 18.5	Underweight	1
$18.5 - 24.9$	Normal Weight	2
$25.0 - 29.9$	Overweight	3
$30.0 - 34.9$	Obesity Class I	4
$35.0 - 39.9$	Obesity Class II	5
≥ 40	Obesity Class III	6

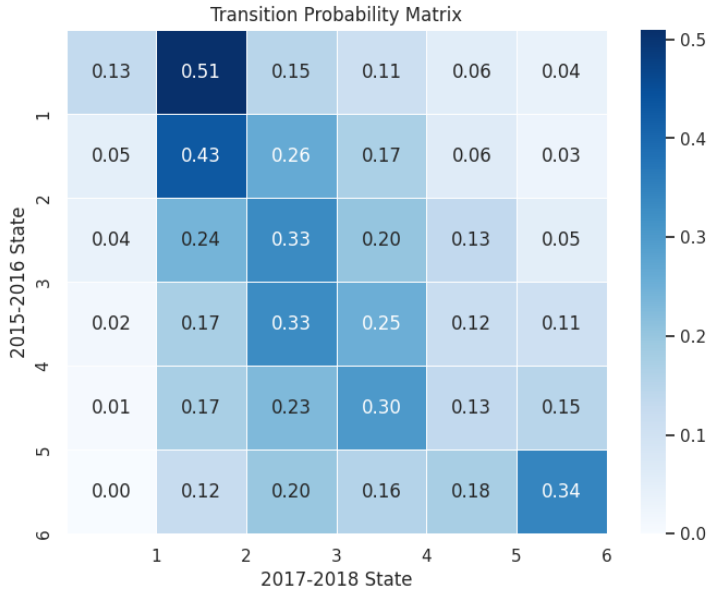
For simplification purposes, we label each category as a state 1-6. Our Markov Model implements our cleaned data as detailed previously. By tracking the progression of a person's BMI throughout the years, we can determine the holistic probability of a state transition between the years 2016-2018 (the most recent years). However, because a different label is given to each person every new year, we cannot simply match their unique identifier codes. Instead, we must utilize the important factors in our random forest model. We can determine the relationship between BMI and the factor using a simple linear regression. An example plot of BMI vs HS C-Reactive Protein is shown below, removing all outliers:



By multiplying the feature importance coefficient by the slope value of the feature, we can determine a general index that is indirectly correlated with BMI. We define this general index as follows:

$$I = \sum_{n=1}^{4} C_n * m_n$$

In this equation, n represents the factor, starting at the most important and ending at the 4th most important. C represents the corresponding feature importance coefficient and m represents the slope value determined via linear regression.
After defining this index value, we then denote $R = \text{BMI}/I$, which gives us the ratio that we use to find the same individual between years. All we have to do now is find an individual with the same R value in 2015-2016 to 2017-2018, and we can compare the BMI values. We can map each BMI value to a class to represent each state in our transition probability matrix. By determining individual transitions between classes based on the BMI value corresponding to individuals with R value closest to each other, we can fill out the transition probability matrix between the years 2015-2016 to 2017-2018.



The y-axis shows the previous state and the x-axis shows the newer state; the state numbers match those found in Table 2. As we can see, there is a general trend of obesity rising between these years, because the probability that one transitions into a higher state is greater than the probability that one transitions into a lower state. Therefore, we can conclude that obesity is on the rise. We will further explore this issue and trend in our Risk Analysis.

3.6 Consumer Behavior Model

Our consumer data presents itself as a classification problem instead of a regression problem like our other datasets, necessitating the use of a different methodology to analyze. While a large majority of the data comes from multiple-choice surveys, there still exists continuous data for some of our factors. To account for these discrepancies, we can utilize the most and least extreme answers to the questionnaires. First, we must take out those who have not answered or refused to answer the question, which is generally encoded with a series of 7s or 9s. After removing these values, we can see that the largest label for a class is the most extreme value. For example, a table may look something like this:
As we can see, the extremity only increases/decreases (depending on how you view the variable). While analyzing difference in median BMI between consecutive classification IDs may not present immediately significant results, comparing their extremities will reveal overarching influences in

Table 3: Time Since Used Marijuana Regularly (Classification Example)

Code	Description	Count
1	Days	348
2	Weeks	51
3	Months	110
4	Years	345
7	Refused	6
9	Don't Know	0

these classification-type factors. Because of this, we can assume that if we compare the largest and smallest code values, we will be able to determine the impact on BMI of each factor. For each factor, we determine the difference between the median BMI of all individuals who answered with the largest code value $\tilde{X_n}$ and the median BMI of all individuals who answered with the smallest code value $\tilde{X_1}$ for n such factors. Note that this already accounts for the few factors that have continuous data, because we are still taking the extreme values, regardless if the factor is a classification or regression problem. The following tables show the top 3 positively correlated factors with their code name, feature description, and corresponding median differences, and the same for the top 3 negatively correlated factors.

Table 4: Top 3 Positively Correlated Consumer Behavior Factors

Factor Code	**Feature**	**Median Difference**
FSD860	Amount in Food Stamps Recieved (increasing by amount of money)	15.80
DUQ400U	Last Time Inject Heroin (increasing by amount of time)	14.60
DBD900	Number of Meals From Fast Food/Pizza Place (increasing by number of visits)	10.80

Table 5: Top 3 Negatively Correlated Consumer Behavior Factors

Factor Code	**Feature**	**Median Difference**
DUQ320	Number of Days Used Heroin in the Last Month (increasing by days)	-15.05
RXDCOUNT	Number of Prescription Medicines Taken (increasing by amount)	-12.30
DBQ235A	Number of Days Drank Milk in the Last Week (increasing by days)	-8.50

From the tables above, we notice a trend in Heroin use and BMI. There seems to be a pretty strong difference in BMI when it comes to how often Heroin is used: using Heroin leads to lower levels of BMI. Additionally, relying on food stamps actually adds to the increase of BMI. Depending on fast food and pizza, a staple to many Americans also increases an individual's BMI, while drinking milk and using prescription drugs lead to a decrease in BMI. We will explore more about these factors in the following sections.

4 Risk Analysis

4.1 Assumptions

We will utilize the same assumption as enumerated in Math Methodology.

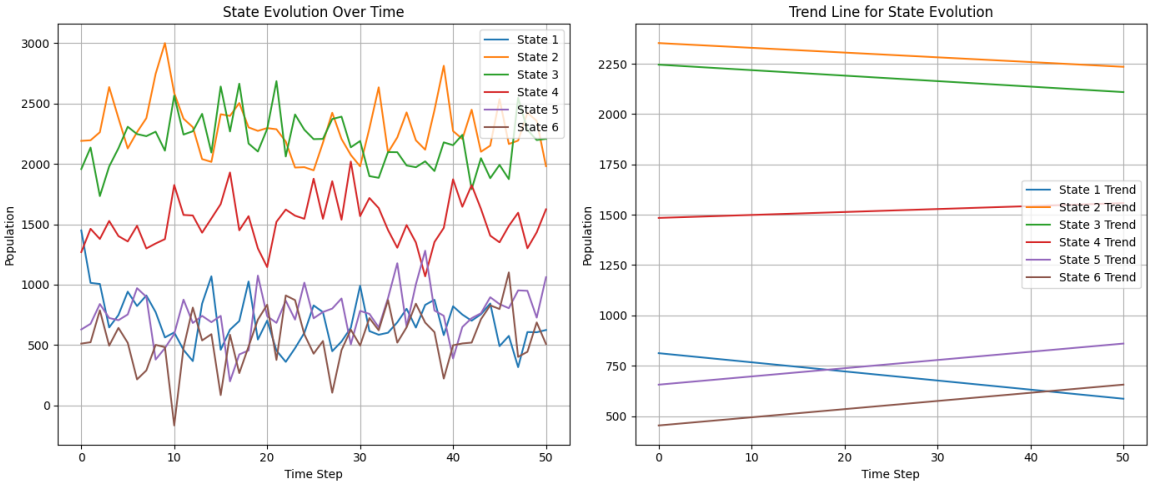
4.2 Trend of Obesity

Firstly, from our NHANES dataset and utilizing the assumption that the sample is representative of the US population, we can define the initial state (essentially the current population) of our transition matrix as follows:

Table 6: NHANES Data Separated Into BMI Categories

Number of People	Category	State
1449	Underweight	1
2191	Normal Weight	2
1957	Overweight	3
1269	Obesity Class I	4
628	Obesity Class II	5
511	Obesity Class III	6

Before applying our transition matrix onto this initial state, we first noted that for our transition matrix, it was reasonable for there to exist a maximum of 5% error in each probability value. Therefore, before proceeding with the iteration, we first set up a code function to generate a unique transition matrix derived from the original transition matrix, yet with random noise added of standard deviation 0.05. From this, we then multiplied the initial state by the random noise-influenced transition matrix, generated newly for each subsequent iteration, and received the following graph after 50 years from 2017-2018.



In the above images, the time step is defined with the units of years. The left image is the exact changes in number of people for each state of obesity. Clearly, we see frequent abrupt changes in count, which matches with our expectations. As time goes on, assuredly something will change in our transition matrix; the values we found are not perfect, and the assumption regarding unchanging factor influence on BMI may be roughly true, but might result in some slight deviations in reality

for 50 years. Therefore, this random noise graph aligns with expectations. Additionally, when each is fitted to a linear regression, we see clear signs of overall change in the following 50 years. Despite the random noise and variation, it is clear that on the current path the US is on, normal weight, overweight, and underweight numbers will decrease, while class I, II, and III obesity will gradually increase. Estimates obtained from our predictions is as follows:

Table 7: Prediction for 2027-2028

Category	People	Percent
Underweight	778	9.69
Normal Weight	2323	28.92
Overweight	2218	27.61
Obesity Class I	1501	18.69
Obesity Class II	715	8.90
Obesity Class III	498	6.20
All Obesity Classes	2714	33.8

Table 8: Prediction for 2037-2038

State	People	Percent
Underweight	731	9.21
Normal Weight	2130	26.84
Overweight	2189	27.58
Obesity Class I	1520	19.15
Obesity Class II	787	9.92
Obesity Class III	579	7.30
All Obesity Classes	2886	36.4

Table 9: Prediction for 2047-2048

Category	People	Percent
Underweight	686	8.92
Normal Weight	1937	25.20
Overweight	2160	28.10
Obesity Class I	1539	20.02
Obesity Class II	787	10.24
Obesity Class III	579	7.53
All Obesity Classes	2905	37.8

Table 10: Prediction for 2057-2058

State	People	Percent
Underweight	639	8.51
Normal Weight	1744	23.22
Overweight	2131	28.37
Obesity Class I	1557	20.73
Obesity Class II	822	10.94
Obesity Class III	619	8.24
All Obesity Classes	2998	39.9

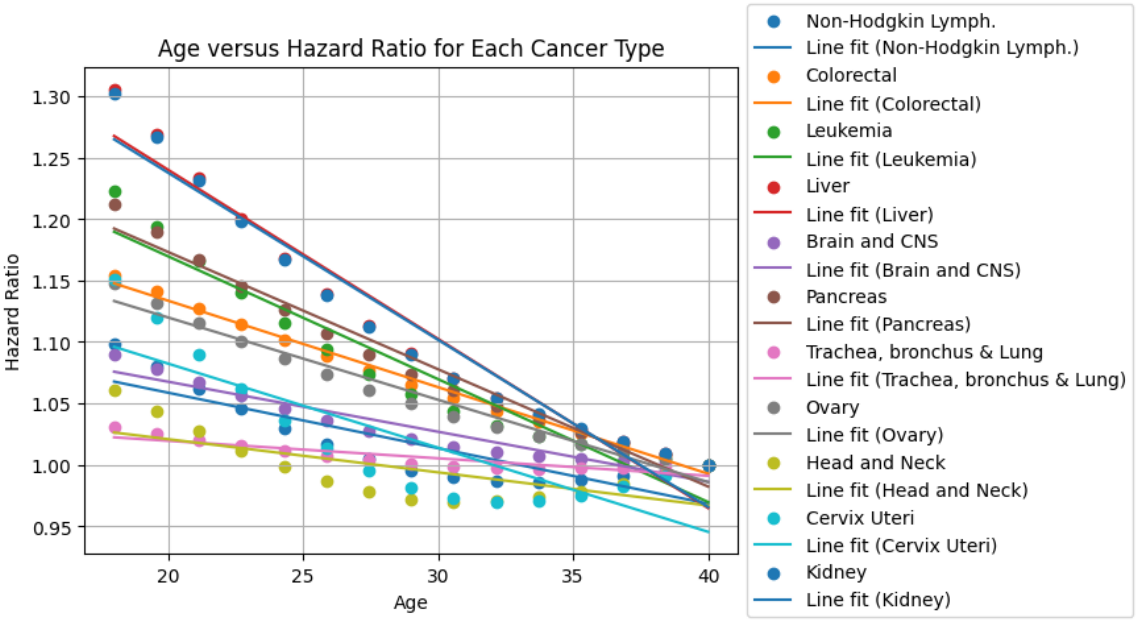
Table 11: Prediction for 2067-2068

Category	People	Percent
Underweight	593	8.08
Normal Weight	1551	21.13
Overweight	2102	28.64
Obesity Class I	1576	21.47
Obesity Class II	858	11.69
Obesity Class III	659	8.98
All Obesity Classes	3093	42.1

4.3 Characterization of Risks and Hazards

According to the CDC, more than 877,500 Americans die of heart disease or stroke every year, costing $216 billion annually in healthcare and $147 billion in lost job productivity [27]. While cardiovascular disease is extensively studied and linked to obesity [28], this report aims to shed light on lesser-known risks associated with other diseases such as cancer, which incurred an annual cost of $190.2 billion in 2015 on the healthcare system [29]. Our methodology can be extrapolated to explore additional diseases and risks.

In a study published by Nature [30], hazard ratios, H for the top 12 cancers were determined based on the age of onset of obesity. A hazard ratio represents the likelihood of developing cancer compared to the average person, with a ratio greater than 1 indicating a higher probability for individuals with obesity. This demonstrates the frequency of obesity causing the risk of cancer. Simple linear regressions were conducted for all 12 types of cancers, revealing a notable correlation: younger onset of obesity is associated with a higher risk of cancer.



Cancer Type	R^2 **Value**
Non-Hodgkin Lymph.	0.739
Colorectal	0.993
Leukemia	0.931
Liver	0.951
Brain and CNS	0.919
Pancreas	0.973
Trachea, bronchus & Lung	0.792
Ovary	0.968
Head and Neck	0.463
Cervix Uteri	0.681
Kidney	0.951

Table 12: R-squared values for different cancer types.

Notice that from the Age versus Hazard Ratio for Each Cancer Type graph, that any ages above 40 can be considered to have a Hazard ratio of 1. Moving forward, we only choose cancers with an $R^2 > 0.9$ for our risk calculations so that we are confident enough to use $\bar{H}$. For our calculations, we can approximate with $\bar{H}$ because from ages 18 to 40 each age has approximately the same amount of people. Thus, since we take cancer types with strong linear behavior, the area under the curve is approximately equal to if we simply took the mean of hazard ratios over ages 18 to 40 year-olds rather than taking each hazard ratio for each age. We combine these estimated with data from the National Cancer Institute [31] that shows the costs of a cancer per person, C, and the CDC [32] also has an estimate of the probability of getting a certain cancer for both genders

every year, P_C. In order to calculate our expected healthcare costs of cancer due to obesity per person, CCO, we set up an equation that mirrors the probability of a cancer occurring due to obesity multiplied by the cost of treatment of the cancer as follows:

$$P_C \cdot \bar{H} \cdot C = \text{CCO}.$$

Our results are summarized in the table below:

Cancer Type	$\bar{H}$	C in USD	P_C	**CCO** in USD
Colorectal	1.070	66,523.5	32.5 E-5	23.13
Leukemia	1.077	47,263.9	12.7 E-5	6.46
Liver	1.113	62,775.7	7.8E-5	5.45
Brain and CNS	1.030	139,813.8	6.1E-5	8.78
Pancreas	1.086	108,165.7	12.9E-5	15.15
Ovary*	1.058	79,120.3	9.2E-5	7.70
Kidney	1.112	41,121.7	15.8E-5	7.22
Total				70.04

Table 13: The expected healthcare cost of cancer per obese person.

*Note: Ovarian cancer costs were multiplied by 0.5 for the total CCO cost

In Section 4.2, we determined that 33.8 percent of the US population will be classified as obese by 2027-2028 year. We can multiply this by percent by the projected us population from the US Census [33]. We estimate an increased cost of $0.338 \cdot 105,335,000 \cdot \$70.04 =$ **$249 million** in the healthcare system attributed to obesity caused cancers. Since we only cancers with $R^2 > 0.9$, we see that this number is actually an underestimate of the true costs the healthcare system faces. There are still a plethora of other diseases linked to obesity that the healthcare system faces, which could be preventable with a healthier population.

5 Recommendations

From our mathematical and risk models, we were able to obtain relative importance of the various factors and obesity and their relationships. Thus, we determine that the following recommendations are most useful and necessary to help mitigate the various risks and losses associated with obesity:

5.1 Age

The most important demographic factor we found was **age**. Age is positively correlated with BMI, meaning as one gets older, their risk of obesity increases.
Many previous studies done into this field noted a noticable correlation between age and BMI, yet have not identified the exact root cause[34]. One study looked into specific chemical compositions in the human body that increase or decrease with aging, and suggested that insulin resistance and metabolism syndrome were seemingly the root drivers behind obesity. However, the paper does conclude that inconclusive results were found when testing specifics sub-phenotypes in the human body, suggesting future experiments as a possible follow-up to enhance their findings [35].
Thus, we cannot suggest a verifiable solution to this issue. While research into anti-aging related supplements and treatments - alongside metabolism boosters - exists, due to the lack of specificity in studies prior regarding aging and obesity, we are unable to directly support any treatment or recommendation. If the situation is relatively dire, it could be considered to taking various online anti-aging related treatments, as there is definitely a possibility of them aiding the situation, yet we cannot guarantee its success[36][37][38]. Our main recommendation would be for those older in age to attend to our following recommendations more closely as they are more at risk of obesity than those younger.

5.2 C-Reactive Protein (CRP) level

One important body factor is **CRP levels**. CRP level is positively correlated with BMI, meaning that a decrease in CRP would decrease one's risk of obesity.
This makes sense because CRP is an annular pentameric protein in which elevated serum levels of CRP are associated with increased body weight[39]. It is also widely recognized as a significant clinical indicator of inflammation. This aligns with obesity being considered a state of chronic, low-grade inflammation, where there is an increased production of pro-inflammatory cytokines compared to those with anti-inflammatory properties [40].
Thus, our approach towards this issue will center on modifying outcomes or changing behaviors on an individual scale as this issue applies to everyone on a case-by-case basis. Insurance strategies would not fit our goal as financial compensation would be the slowest acting and least effective method at solving the immediate problem of increased CRP levels.

1. **Modifying Outcomes:** We recommend the implementation of a diet that focuses on eating anti-inflammatory foods. This involves the avoidance of inflammatory foods like refined carbohydrates, fried foods, red meat, and processed meat, and the incorporation of fruits and vegetables like leafy greens, nuts, fatty fish and whole grains[41][42]. These foods have been found to reduce the frequency of elevated CRP levels. Additionally, many medications such as Statins, NSAIDs, Corticosteroids, Metformin, and herbal supplements may be prescribed by doctors and have an effect on controlling CRP levels[41].

2. **Changing Behaviors:** There are also many behaviors one can adopt to reduce CRP levels. In particular, we recommend individuals to stop smoking as smoking triggers inflammation and damages blood vessels, inducing elevated CRP levles[43]. Alcohol use is also recommended to be limited to one or two drinks a day as alcohol also promotes inflammation[44]. Another important habit one is recommended take on is exercising regularly. Even though intensive exercise may raise CRP levels in the short term, consistent exercise has been found in the long term to help control CRP levels and prevent their elevation[45]. Lastly, chronic stress also can raise inflammation[28]. We recommend individuals practice stress-relieving activities daily such as meditation, yoga, or even going to sleep early and consistently.

5.3 Blood Lead Level (BLL)

Another important body factor is **BLL**. BLL is negatively correlated with BMI, meaning that an increase in lead would decrease one's risk of obesity.

On first glance, our negative correlation between BLL and BMI seems to contradict what was found in another study on the topic[46]. The study mentioned found that for residents of China, their increase in BLL actually resulted in matched increases in BMI, likely due to the lead causing insulin resistance build up. However, upon closer examination, we note that the study conducted in China actually analyzed significantly higher levels of BLLs as compared to the ones provided by NHANES. As the study cites, they analyzed primarily chronic 0.05% lead exposure, which when conducted over the course of their study (21 weeks) led to BLLs of 56.25 +or- 7.47 ng/mg. This is the root cause of the stark difference in our results. In the NHANES dataset, the provided BLLs were significantly lower, with the maximum BLL sampled being 5.43 ng/mg. Due to the significantly higher range the China study analyzed, their results spoke to a much more extreme scenario, unlikely to occur commonly in the US, where policy dictates a mandatory maximum of 0.03% lead exposure for 30 days per year[47]. Therefore, we can reasonably conclude that, while our results were different from the study, both are likely correct results, and the difference stems from a disparity in extremity regarding the BLL test range. Additionally, a common side effect of lead poisoning is indeed weight loss[48][49][50]. Thus, we can conclude that even though a certain BLL may lead to slight increases in BMI, this increase likely immediately drops off, leading to the negative correlation we found.

However, this certain BLL is not worth pursuing. From common side effects like developmental delay, learning difficulties, irritability, and loss of appetite to more severe side effects like infertility, memory issues, and miscarriages[48][49][50], lead poisoning leads to a host of other serious issues. As high lead levels have contributed to 5.5 million adult cardiovascular disease deaths and 765 million lost IQ points among children younger than 5 years in 2019[51], it becomes clear that an increase in BLL is not a viable solution to recommend.

Thus, even though an increase in BLL may help decrease the risk of obesity, our analysis finds that the extraneous risks involved with this increase in lead levels far outweigh the potential obesity-related benefits it may bring. This result also informs us about the need to properly research the results of our model. As our model is unable to determine extraneous impacts each factor may have and only returns the relation each factor has with obesity, it is imperative to not take results at face value and to continue conducting the proper and necessary research to make meaningful recommendations.

5.4 Red Blood Cell Distribution Width (RDW)

Another important body factor is **RDW**. RDW is positively correlated with BMI, meaning that a decrease in RDW would decrease one's risk of obesity.
RDW represents the difference between the size of your smallest and largest red blood cells. Our findings makes sense because, as stated in our analysis of CRP in which obesity can be defined as a state of chronic low-grade inflammation, red blood cells are what physically cause this state of inflammation and have a high positive correlation with BMI[52][53]. Additionally, elevated RDW is an indicator of anemia, which is commonly associated with weight gain[54], among other effects.
Thus, our approach will center upon either modifying outcomes or encouraging behavior change on an individual scale, as this is a problem that mainly affects people on a case-by-case basis. Insurance strategies would not fit our goal as financial compensation would be the slowest acting and least effective method at solving the immediate problem of inflamed red cells.

1. **Modifying Outcomes:** One way to lower the outcomes of elevated RDW is to encourage the adoption of diets targeted at reducing deficiencies in iron, folate, and vitamin B12[55]. These specific chemicals were found to prevent extreme elevation in RDW, and thus are important for our goal of reducing RDW. Supplements or injections of the above chemicals work as well if one has issues absorbing nutrients from foods.
2. **Changing Behaviors:** There are also many behaviors one can adopt to reduce RDW. In particular, we recommend individuals to stop smoking and to avoid alcohol. Both of these habits damage red blood cells as well as prevent proper absorption of necessary vitamins like vitamin B12 that work to decrease RDW [55][56]. Another important habit one is recommended take on is exercising regularly[55], as for every one exercise session per day increase, the odds of having an elevated RDW (and thus an increased risk of obesity) was found to be reduced by 34%[57].

5.5 Food Stamp Usage

An important behavior factor is **food stamps**. The use of food stamps is positively correlated with BMI, meaning that the more food stamps one uses, the more at risk of obesity they will be. This makes sense because food stamp usage often is an indicator of low financial stability, leading to increased difficulty in purchasing nutritious or even just enough food. One particular food stamp program (the Supplemental Nutrition Assistance Program) conducted a study on their participants and found that use of their program did indeed positively correlate with overweight and obesity rates[58][59]. However, this is not to say that the physical use of food stamps is what leads to obesity. The food the food stamps provide is often not nutritious enough[60] due to things like lack of funds[61], leading to the same problems associated with lack of nutrition.
Thus, our approach will center upon insurance on an individual level (as those who suffer the losses are generally individuals) and modifying outcomes on a community and corporate scale (as this issue also impacts the groups who run such food stamp programs). Changing behaviors would not work as well as neither the food stamp programs nor the individuals themselves are not responsible for the losses they experience - these losses are due to external factors.

1. **Insurance:** We recommend the implementation of an insurance policy that covers a lack of nutritious food. This would work in a manner similar to food stamps but providing money

instead of food so that individuals may purchase foods to satisfy their missing nutrients. While this would be potentially more dangerous for the governmental bodies involved as there is no guarantee that the recipients will spend the money on food, this is only the worst case scenario. After all, this insurance policy only activates if individuals do not receive sufficient food/nutrition from their food stamps, and thus will need the money to physically survive. Ideally, this insurance will rarely activate, reducing the monetary burden on society.

2. **Modifying Outcomes:** We recommend that governmental bodies subsidize the food stamp programs. As loss and increased risk is generally found to be due to lack of funds to secure proper nutritious foods and transport, we can reduce the both the frequency and the severity of loss by providing more funds to such programs. Another way to reduce the severity of loss would be to adjust the food types that food stamps provide. By mandating them to provide healthier, more nutritious foods for impoverished families who might struggle to buy them on their own, the increase in food stamps usage could actually decrease obesity.

5.6 Heroin Usage

Another important behavior factor is **heroin usage**. The time an individual last used heroin is negatively correlated with BMI, meaning increased heroin usage decreases one's risk of obesity. This makes sense because heroin is an opioid and impacts people in many ways. For example, people using heroin may be neglecting to eat healthily or even at all, potentially leading to acute dehydration or imbalances in electrolytes. Furthermore, heroin affects the gastrointestinal tract, commonly resulting in symptoms such as constipation, nausea, and vomiting. These gastrointestinal issues can subsequently diminish appetite and contribute to weight loss[62]. Many studies also share these results and find that heroin users have a lower BMI than non-heroin users[63][64]. While heroin does seem to reduce BMI, its risks far outweigh the possible rewards that come with it. For long-time heroin users, employing heroin as a solution to obesity could feed into their addiction, cause them to relapse, or create an abusable, hard to regulate system possible providing heroin to addicts claiming to need them for obesity issues. Additionally, for people who have not used heroin prior, a first-time use to curb obesity is extremely dangerous. Not only are initial side effects possible nausea, vomiting, and severe itchiness, following its initial introduction to the user there is consistent evidence of heart function slowing, reduced breathing, and clouded mental function, all of which could potentially injure or permanently harm the patient [65][66]. However, while we cannot suggest heroin use for those wishing to decrease their risk of obesity, we instead recommend the use of less serious opioids such as marijuana. Medical marijuana shows similar results as heroin in reducing one's BMI and other risks of obesity[63]. This is a safer option for those still wishing to go down this route. However, we still recommend individuals consult with their doctor before utilizing various drugs to lower their risk of obesity.

5.7 Summary

Blood-lead and heroin usage were analyzed to be unfeasible factors to consider implementing to reduce obesity due to the extraneous risks associated with them. Age was found to be difficult to consider due to barriers in technology. Analyzing all the other important factors found from our methodologies, we strongly recommend the following actions to mitigate both the losses and the risks associated with obesity:

1. **Specialized Diets** - One main diets focuses on *anti-inflammatory foods* (reducing CRP levels) such as leafy greens, nuts, and whole grains. Another diet focuses on *reducing iron, folate, and vitamin B12 deficiencies* (decrease RDW).
2. **Reduce Alcohol Intake**
3. **Stop Smoking**
4. **Reduce Stressful Activities** - Some possible recommendations are to sleep more, and to practice yoga or mediation.
5. **Exercise More**
6. **Subsidize and Insure Food Stamp Programs**
7. **Mandate Nutritious Foods for Food Stamp Programs**
8. **Medical Marijuana usage** - We still recommend consultation with medical practitioners before usage.

And lastly, as one gets older, we recommend one to follow the above points more closely due to increased risk.

6 Acknowledgements

We sincerely thank the following for assisting us in this wonderful opportunity:

1. Our team coach: Paul Kim
2. Our mentor: Paul Heffernan
3. Members of the Actuarial Foundation for organizing and running this competition

References

[1] [Online]. Available: https://time.com/6264865/global-obesity-rates-increasing/

[2] [Online]. Available: https://newsroom.heart.org/news/obesity-related-cardiovascular-disease-deaths-tripled-between-1999-and-2020#

[3] [Online]. Available: https://www.colorado.edu/today/2023/02/23/excess-weight-obesity-more-deadly-previously-believed

[4] [Online]. Available: https://www.washingtonpost.com/wellness/2023/09/18/obesity-heart-disease-cardiac-death/

[5] [Online]. Available: https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/health-risks

[6] [Online]. Available: https://stop.publichealth.gwu.edu/sites/g/files/zaxdzs4356/files/2022-06/fast-facts-costs-of-obesity.pdf

[7] [Online]. Available: https://www.cdc.gov/obesity/basics/adult-defining.html

[8] [Online]. Available: https://lasvegassun.com/news/2013/feb/11/heart-attack-grill-spokesman-dies-heart-attack/

[9] [Online]. Available: https://www.grandviewresearch.com/industry-analysis/fast-food-market

[10] [Online]. Available: https://www.bbc.com/future/bespoke/follow-the-food/why-modern-food-lost-its-nutrients/

[11] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8199310/#:~:text=This%20happens%20due%20to%20advances,observed%20predominantly%20in%20food%20industries.

[12] [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-8090-5

[13] [Online]. Available: https://www.bcm.edu/news/how-stress-can-affect-your-sleep

[14] [Online]. Available: www.hsph.harvard.edu/obesity-prevention-source/obesity-causes/sleep-and-obesity/

[15] [Online]. Available: https://www.nature.com/articles/s41576-021-00414-z

[16] [Online]. Available: https://www.cdc.gov/genomics/resources/diseases/obesity/index.htm

[17] [Online]. Available: https://www.cdc.gov/obesity/strategies/index.html

[18] [Online]. Available: https://www.hsph.harvard.edu/obesity-prevention-source/obesity-prevention/

[19] [Online]. Available: https://www.nationalgeographic.co.uk/environment-and-conservation/2022/05/fruits-and-vegetables-are-less-nutritious-than-they-used-to-be

[20] [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/analyticguidelines.aspx#plan-and-operations

[21] [Online]. Available: https://news.mit.edu/2022/synthetic-data-ai-improvements-1103

[22] [Online]. Available: https://www.youtube.com/watch?v=Dc0sr0kdBVI&ab_channel=caltech

[23] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[24] [Online]. Available: https://www.mdpi.com/2078-2489/13/12/575#:~:text=3.1.-,System%20Model,and%20gives%20a%20completed%20vector.

[25] D. W. Gareth James, Trevor Hastie and R. Tibshirani, *An Introduction to Statistical Learning*. Springer, 2013.

[26] [Online]. Available: https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmi_dis.htm

[27] [Online]. Available: https://www.cdc.gov/chronicdisease/about/costs/index.htm#:~:text=Nothing%20kills%20more%20Americans%20than%20heart%20disease%20and,billion%20in%20lost%20productivity%20on%20the%20job.%203

[28] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476783/

[29] [Online]. Available: https://progressreport.cancer.gov/after/economic_burden

[30] [Online]. Available: https://www.nature.com/articles/s41467-023-39282-y#Sec15

[31] [Online]. Available: https://progressreport.cancer.gov/after/economic_burden

[32] [Online]. Available: https://gis.cdc.gov/Cancer/USCS/#/Demographics/

[33] [Online]. Available: https://www.census.gov/data/tables/2023/demo/popproj/2023-summary-tables.html

[34] [Online]. Available: https://www.science.org/doi/10.1126/sageke.2004.24.re4#:~:text=In%20sum%2C%20obesity%20appears%20to,to%20adiposity%20in%20young%20adults

[35] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5005878/

[36] [Online]. Available: https://www.news-medical.net/news/20230712/Researchers-develop-a-chemical-approach-to-reverse-aging.aspx

[37] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10373966/

[38] [Online]. Available: https://www.medicalnewstoday.com/articles/have-scientists-finally-found-the-way-to-the-fountain-of-youth#Can-these-findings-be-applied-to-humans?

[39] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7044181/

[40] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3752622/

[41] [Online]. Available: https://www.ondemand.labcorp.com/blog/what-is-high-c-reactive-protein-how-to-lower-crp#:~:

[42] [Online]. Available: https://www.personalabs.com/blog/how-to-reduce-c-reactive-protein-crp-levels-naturally/

[43] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3261116/

[44] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2842521/

[45] [Online]. Available: https://www.health.harvard.edu/staying-healthy/easy-ways-to-keep-inflammation-in-check

[46] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4677308/

[47] [Online]. Available: https://www.atsdr.cdc.gov/csem/leadtoxicity/safety_standards.html#:~:text=OSHA%20set%20a%20Permissible%20Exposure,than%2030%20days%20per%20year.

[48] [Online]. Available: https://www.mayoclinic.org/diseases-conditions/lead-poisoning/symptoms-causes/syc-20354717

[49] [Online]. Available: https://www.floridahealth.gov/environmental-health/lead-poisoning/adults.html

[50] [Online]. Available: https://www.atsdr.cdc.gov/csem/leadtoxicity/signs_and_symptoms.html

[51] [Online]. Available: https://jamanetwork.com/journals/jama/article-abstract/2809890#

[52] [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/23239558/

[53] [Online]. Available: https://link.springer.com/article/10.1007/s13300-020-00897-9

[54] [Online]. Available: https://www.healthline.com/health/does-anemia-cause-weight-loss-or-gain#:~:text=If%20your%20body%20doesn%27t,may%20contribute%20to%20weight%20gain.

[55] [Online]. Available: https://www.medicalnewstoday.com/articles/321568#how-to-lower

[56] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5281527/

[57] [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25602910/

[58] [Online]. Available: https://www.snaptohealth.org/snap/snap-and-obesity-the-facts-and-fictions-of-snap-nutrition/#:~:text=However%2C%20the%20same%20study%20found,obese%20by%202-5%25.

[59] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4580337/

[60] [Online]. Available: https://www.cato.org/briefing-paper/snap-high-costs-low-nutrition#

[61] [Online]. Available: https://www.cbpp.org/blog/end-of-snaps-temporary-emergency-allotments-resulted-in-substantial-benefit-cut

[62] [Online]. Available: https://www.bicyclehealth.com/opioid-education/heroin/impact-on-weight

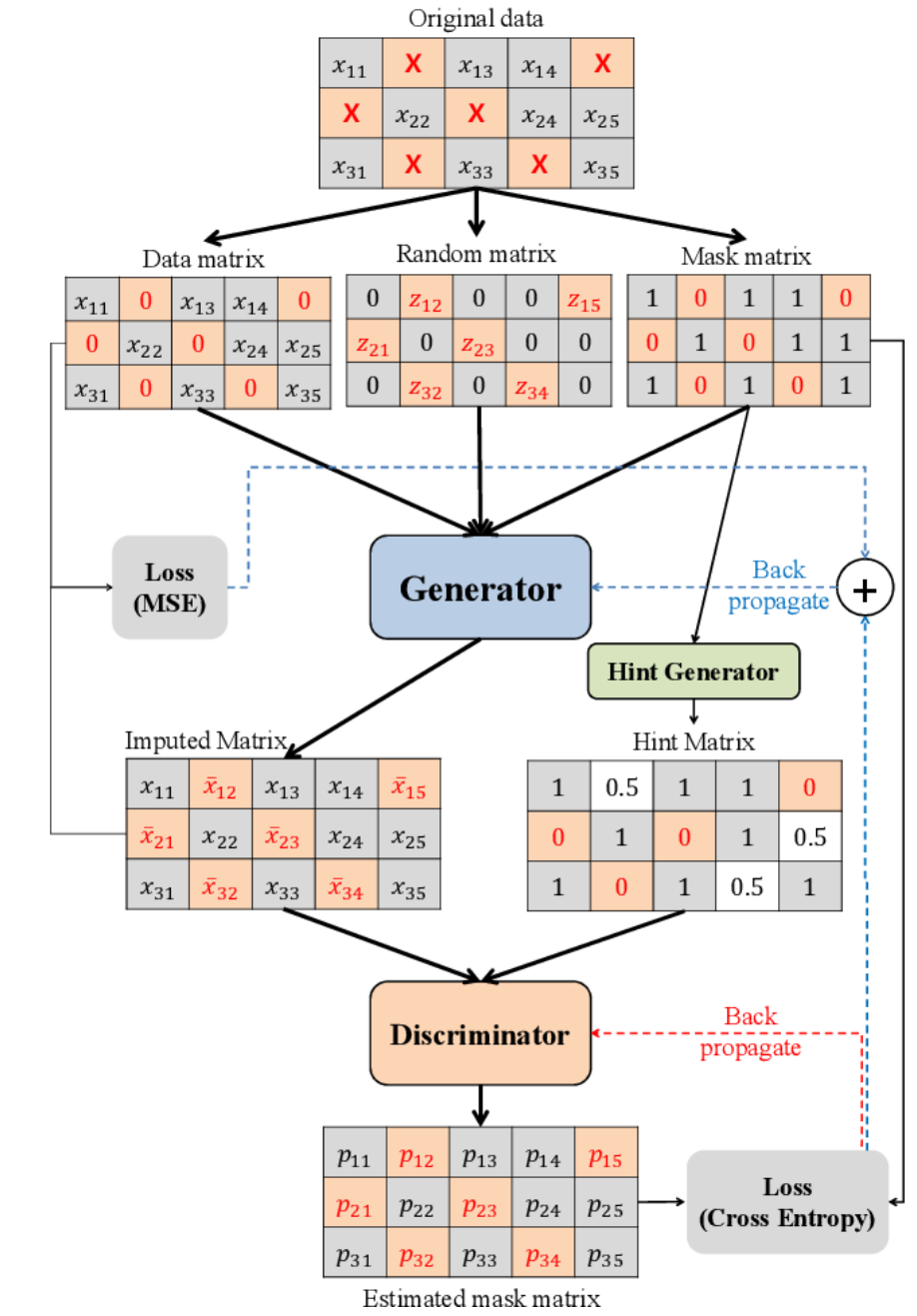
[63] [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4939607/

[64] [Online]. Available: https://academic.oup.com/nutritionreviews/article/79/6/627/5911317

[65] [Online]. Available: https://www.dea.gov/sites/default/files/2023-04/Heroin%202022%20Drug%20Fact%20Sheet.pdf

[66] [Online]. Available: https://nida.nih.gov/publications/research-reports/heroin/what-are-immediate-short-term-effects-heroin-use.

A GAIN Model


Original data
Data matrix
Random matrix
Mask matrix
Loss (MSE)
Generator
Back propagate
Hint Generator
Imputed Matrix
Hint Matrix
Discriminator
Back propagate
Loss (Cross Entropy)
Estimated mask matrix

B Code Used

```
#Necessary Packages
import os
import gspread
import pandas as pd
from oauth2client.service_account import ServiceAccountCredentials
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from sklearn.impute import SimpleImputer
import statsmodels.api as sm
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
import seaborn as sns
from sklearn.preprocessing import MinMaxScaler
import numpy as np
from scipy.integrate import simps
import warnings
warnings.filterwarnings("ignore")
import sys

#Connect To Colab
sys.path.append('/content/drive/MyDrive')
from google.colab import drive
drive.mount('/content/drive',force_remount=True)
file_path = '/content/drive/My Drive/NHANES Data'
json_path = '/content/drive/My Drive/NHANES.json'
BMI_df = pd.read_csv(file_path+"/BMI/BMI Composite.csv")

#Demographic Data Cleaning
def cleanDemo(folderName, factorsToInclude):
  all_years_demo_df = pd.DataFrame()
  csv_files = [file for file in os.listdir(folderName) if file.endswith(".csv")]
  print(csv_files)
  sheets_by_year = {}
  for dataSheet in csv_files:
    year = dataSheet[:4]
    if year not in sheets_by_year:
        sheets_by_year[year] = []
    sheets_by_year[year].append(dataSheet)
  for year, sheets in sheets_by_year.items():
    print(f"Year: {year}")
    year_df = pd.DataFrame()
    for dataSheet in sheets:
        df = pd.read_csv(os.path.join(folderName, dataSheet))
        common_columns = set(df.columns) & set(factorsToInclude)
        df = df[common_columns]
        if year_df.empty:
          year_df = df
        else:
```

```
          year_df = pd.merge(year_df, df, on="SEQN", how="outer")
    matching_columns = [col for col in BMI_df.columns if col.startswith(year[:4])]
    left_column = BMI_df.columns[BMI_df.columns.get_loc(matching_columns[0]) - 1]
    right_column = BMI_df.columns[BMI_df.columns.get_loc(matching_columns[-1]) + 1]
    bmi_year_df = BMI_df[[left_column, right_column]]
    bmi_year_df = bmi_year_df.rename(columns={left_column: "SEQN", right_column: "BMXBMI"})
    year_df = pd.merge(year_df, bmi_year_df, on="SEQN", how="outer")
    year_df['DMDEDUC3'].fillna(year_df['DMDEDUC2'], inplace = True)
    year_df.drop('DMDEDUC2',axis = 1, inplace = True)
    if 'INDHHIN2' in year_df.columns :
      year_df['INDHHINC'] = year_df['INDHHIN2']
      year_df.drop('INDHHIN2',axis = 1, inplace = True)
    all_years_demo_df = pd.concat([all_years_demo_df, year_df], axis = 0)
  return all_years_demo_df

factorsToInclude = ['SEQN','RIAGENDR','RIDAGEYR','DMDEDUC3','DMDEDUC2','INDHHIN2', 'INDHHINC', 'DMDHHSIZ', 'RIDRETH1']
all_year_demo_df = cleanDemo(file_path+"/Demographics",factorsToInclude)
all_year_demo_df = all_year_demo_df.dropna()

#Map Containing All Years and Corrsponding DataFrame
all_years_lab = {}

#Lab Cleaning
def runLab(folder, factorsToNotInclude, yearsToInclude):
  subfolders = [f.name for f in os.scandir(folder) if f.name[:4] in yearsToInclude]
  print(subfolders)
  for subfolder in subfolders:
    year = subfolder[:4]
    subfolder_path = folder+"/"+subfolder
    print(f"Contents of subfolder: {subfolder}")
    csv_files = [file for file in os.listdir(subfolder_path) if file.endswith(".csv") and file[:4] == year]
    year_df = pd.DataFrame()
    for csv_file in csv_files:
      csv_file_path = os.path.join(subfolder_path, csv_file)
      df = pd.read_csv(csv_file_path, index_col = 0)
      cols_to_use = df.columns.difference(year_df.columns).union(['SEQN'])
      if(year_df.empty):
        year_df = df
      else:
        year_df = pd.merge(year_df, df[cols_to_use], on="SEQN", how="outer")
      if 'Unnamed: 0_y' in year_df.columns:
        print(csv_file)
    print(year_df.shape)
    matching_columns = [col for col in BMI_df.columns if col.startswith(year[:4])]
    left_column = BMI_df.columns[BMI_df.columns.get_loc(matching_columns[0]) - 1]
    right_column = BMI_df.columns[BMI_df.columns.get_loc(matching_columns[-1]) + 1]
    bmi_year_df = BMI_df[[left_column, right_column]]
```

```
    bmi_year_df = bmi_year_df.rename(columns={left_column: "SEQN", right_column: "BMXBMI"})
    year_df = pd.merge(year_df, bmi_year_df, on="SEQN", how="outer")
    all_years_lab[year] = year_df

factorsToNotInclude = []
yearsToInclude = ['2005', '2009', '2011', '2013', '2015', '2017']
runLab(file_path+"/lab data v3", factorsToNotInclude, yearsToInclude)

#Combine to One DataFrame
all_years_lab_df = pd.DataFrame()
for year in yearsToInclude:
  all_years_lab_df = pd.concat([all_years_lab_df, all_years_lab[year]], axis = 0)
print(all_years_lab_df.shape)

#Plotting the Threshold Model
def testNan(df, lower_bound=0.35, upper_bound=1, nan_percentage_threshold = 16.67, num_steps=100, minimum_non_nan = 1472.06):
  thresholds = np.linspace(lower_bound, upper_bound, num_steps)
  column_counts = []
  row_counts = []
  avg_non_nan_values = []
  for threshold in thresholds:
      df_cleaned = df.dropna(thresh=len(df.columns) * (1 - threshold))
      nan_percentage = (df_cleaned.isnull().mean() * 100)
      columns_under_threshold = nan_percentage[nan_percentage < nan_percentage_threshold].count()
      column_counts.append(columns_under_threshold)
      row_count = len(df_cleaned)
      row_counts.append(row_count)
      avg_non_nan_per_column = df_cleaned.count().mean()
      avg_non_nan_values.append(avg_non_nan_per_column)

  max_column_count = max(column_counts)
  max_column_count_index = column_counts.index(max_column_count)
  print(max_column_count)
  print(row_counts[max_column_count_index])
  print(avg_non_nan_values[max_column_count_index])
  corresponding_threshold = thresholds[max_column_count_index]
  optimal_threshold_index = max(i for i, count in enumerate(column_counts) if count > 290)
  optimal_threshold = thresholds[optimal_threshold_index]
  minimum_threshold_index = next(i for i, avg_non_nan in enumerate(avg_non_nan_values) if avg_non_nan >= minimum_non_nan)
  minimum_threshold = thresholds[minimum_threshold_index]

  print(row_counts[optimal_threshold_index])
  plt.figure(figsize=(10, 6))
  plt.plot(thresholds, avg_non_nan_values, marker='o')
  plt.axvline(optimal_threshold, color='green', linestyle='--', label=f'Optimal Threshold: {optimal_threshold:.2f}')
  plt.axvline(x=minimum_threshold, color='r', linestyle='--', label=f'Minimum Threshold: {minimum_threshold:.2f}')
```

```
  plt.axvline(corresponding_threshold, color='orange', linestyle='--', label=f'Max Column Threshold: {corresponding_threshold:.2f}')
  plt.legend()
  plt.title('Average Non-Empty Values per Column vs. Threshold for Row Empty Percent')
  plt.xlabel('Threshold for Row Empty Percent')
  plt.ylabel('Average Non-Empty Values per Column')
  plt.grid(True)
  plt.show()

  plt.figure(figsize=(10, 6))
  plt.plot(thresholds, column_counts, marker='o')
  plt.xlabel('Threshold for Row Empty Percent')
  plt.ylabel(f'Number of Columns with Empty Cell Percent < {nan_percentage_threshold:.2f}%')
  plt.title(f'Number of Columns with Empty Cell Percent < {nan_percentage_threshold:.2f}% vs Threshold for Row Empty Percent')
  plt.axvline(optimal_threshold, color='green', linestyle='--', label=f'Optimal Threshold: {optimal_threshold:.2f}')
  plt.axvline(x=minimum_threshold, color='red', linestyle='--', label=f'Minimum Threshold: {minimum_threshold:.2f}')
  plt.axvline(corresponding_threshold, color='orange', linestyle='--', label=f'Max Column Threshold: {corresponding_threshold:.2f}')
  plt.legend()
  plt.show()

  plt.figure(figsize=(10, 6))
  plt.plot(thresholds, row_counts, marker='o', linestyle='-')
  plt.title('Number of Rows vs. Threshold for Row Empty Percent')
  plt.axvline(optimal_threshold, color='green', linestyle='--', label=f'Optimal Threshold: {optimal_threshold:.2f}')
  plt.axvline(x=minimum_threshold, color='red', linestyle='--', label=f'Minimum Threshold: {minimum_threshold:.2f}')
  plt.axvline(corresponding_threshold, color='orange', linestyle='--', label=f'Max Column Threshold: {corresponding_threshold:.2f}')
  plt.legend()
  plt.xlabel('Threshold for Row Empty Percent')
  plt.ylabel('Number of Rows')
  plt.show()
  df_cleaned = df.dropna(thresh=len(df.columns) * (1 - optimal_threshold))
  threshold_count = len(df_cleaned) * (nan_percentage_threshold / 100)
  df_cleaned = df_cleaned.dropna(axis=1, thresh=threshold_count)
  return df_cleaned
all_years_lab_df_cleaned = testNan(all_years_lab_df)

#More Cleaning
all_years_lab_df_cleaned = all_years_lab_df_cleaned.dropna(subset=['SEQN'])

#Normalizing Our Data
scaler = MinMaxScaler()
scaler.fit(all_years_lab_df_cleaned)
lab_df_normalized = pd.DataFrame(scaler.transform(all_years_lab_df_cleaned))

```

```
#Determine Classification Columns
threshold_unique_count = 3
class_columns = []
for column in lab_df_normalized.columns:
    unique_count = lab_df_normalized[column].nunique()
    if unique_count <= threshold_unique_count:
        class_columns.append(column)

#Create a Classification Map
column_unique_values_map = {}
for column in class_columns:
    unique_values = lab_df_normalized[column].dropna().unique()
    column_unique_values_map[column] = unique_values

#Import our GAIN Model
from gain import Gain

#Fit the GAIN Model To Synthesize Data With Vmaps as Classification Map
gain_imputor = Gain(parameters = {'iterations' : 100},names = list(
    lab_df_normalized.columns) , vmaps = column_unique_values_map)
gain_imputor.fit(lab_df_normalized)
imputed_data, new_names, vmaps  = gain_imputor.transform(lab_df_normalized)

#Export Our Data After Unnormalizing and Resetting Labels (Save a Copy)
imputed_data_df = pd.DataFrame(imputed_data, columns = new_names)
all_years_lab_df_cleaned.to_csv('nonimputed.csv', index = False)
lab_df_unnormalized = pd.DataFrame(scaler.inverse_transform(imputed_data_df),
    columns=all_years_lab_df_cleaned.columns)
lab_df_unnormalized.to_csv('imputedReal.csv', index = False)

#Reimport Our Data
lab_df = pd.read_csv('imputedReal.csv')

#Our Random Forest Model
def randomForest(df):
    X = df.drop(columns=['BMXBMI', 'SEQN'], axis=1)
    y = df['BMXBMI']
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
    random_state=33)
    rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
    rf_model.fit(X_train, y_train)
    y_pred = rf_model.predict(X_test)
    feature_importance = rf_model.feature_importances_
    feature_importance_df = pd.DataFrame({'Feature': X.columns, 'Importance':
    feature_importance})
    feature_importance_df = feature_importance_df.sort_values(by='Importance',
    ascending=False)
    return feature_importance_df

#Starting the Process of Merging With Demographic Data
lab_df = lab_df.drop_duplicates(subset=['SEQN'])
lab_df  = lab_df.drop('BMXBMI', axis = 1)
all_df = pd.merge(all_year_demo_df, lab_df, on="SEQN", how="outer")
```

```
all_df = all_df.dropna()
all_features = randomForest(all_df)

#Display Top 5
plt.figure(figsize=(10, 6))
plt.barh(all_features['Feature'].head(5), all_features['Importance'].head(5))
plt.xlabel('Importance')
plt.title('Top 5 Factors and their Importance')

#Get Only Necessary Features
specific_features_trimmed = all_features.head(4)

#Get Slope Values and Exclude Outliers Through Linear Regression
m_list = []
for index, row in specific_features_trimmed.iterrows():
    feature = row['Feature']
    importance = row['Importance']
    x = all_df[feature]
    y = all_df['BMXBMI']
    z_scores = (all_df[feature] - all_df[feature].mean()) / all_df[feature].std()
    outliers = abs(z_scores) > 2
    cleaned_df = all_df[~outliers]
    X_cleaned = sm.add_constant(cleaned_df[feature])
    model_cleaned = sm.OLS(cleaned_df['BMXBMI'], X_cleaned).fit()
    m = model_cleaned.params[1]
    m_list.append(m)

#Our Way of Determining BMI_Index
def BMI_Index(ls, bmi):
  total = 0
  for index, value in enumerate(m_list):
    total += value*ls[index]
  return total/bmi

#Create BMI_Index Column
for index, row in all_df.iterrows():
  row_values = [row[col] for col in specific_features_trimmed['Feature']]
  all_df.at[index,'BMI_Index'] = BMI_Index(row_values, row['BMXBMI'])

#Specific Year Distinction (Separate by Year)
all_df['SEQ'] = np.where((all_df['SEQN'] >= 83732) & (all_df['SEQN'] <= 93702), 8,
    9)

#BMI Class Thresholds
class_thresholds = [0, 18.5, 25.0, 30.0, 35.0, 40.0, float('inf')]
class_labels = [1, 2, 3, 4, 5, 6]

#Give a Classification Label to Each
all_df['BMI_Class'] = pd.cut(all_df['BMXBMI'], bins=class_thresholds, labels=
   class_labels, right=False)
all_df['BMI_Class'] = all_df['BMI_Class']

#Useful Conversion from Combination to Probability Matrix
```

```
def convertProb(transition_matrix):
    T = np.array(transition_matrix, dtype=float)
    row_sums = T.sum(axis=1, keepdims=True)
    P = T / row_sums
    return P

#First Year Classes
filtered_values = all_df.loc[all_df['SEQ'] == 8, 'BMI_Index'].tolist()

#Find Closest Matching BMI_Index in the Next Year
def closest_row(group, value):
    closest_index = (group['BMI_Index'] - value).abs().idxmin()
    return group.loc[closest_index]

#Create Matrix
transition_1 = np.zeros((6, 6), dtype=int)
for num in filtered_values:
  closest_rows = all_df.groupby('SEQ').apply(closest_row, value = num)
  transition_1[closest_rows['BMI_Class'].astype(int)[8]-1,closest_rows['BMI_Class'
    ].astype(int)[9]-1]+=1
transition_1 = np.nan_to_num(convertProb(transition_1), nan = 0)
print(transition_1)

#Visual
def plot_transition_matrix(transition_matrix, cmap="Blues"):
    sns.set()
    plt.figure(figsize=(8, 6))
    sns.heatmap(transition_matrix, annot=True, cmap=cmap, fmt=".2f", linewidths
    =.5)
    plt.xticks(ticks=[i + 1 for i in range(6)], labels=[str(i) for i in range(1,
    7)])
    plt.yticks(ticks=[i + 1 for i in range(6)], labels=[str(i) for i in range(1,
    7)])
    plt.xlabel('2017-2018 State')
    plt.ylabel('2015-2016 State')
    plt.title('Transition Probability Matrix')
    plt.show()
plot_transition_matrix(transition_1)

#Set Up Transition for Markov Chain
transition_matrix = transition_1
std_dev = 0.05
mean = 0
num_states = 50

#Initial Values Determined Through BMI Count in 2018
initial_state = []
BMI_df['BMI_class'] = pd.cut(BMI_df['BMXBMI'], bins=class_thresholds, labels=
    class_labels, right=False)
class_counts = BMI_df['BMI_class'].value_counts().sort_index()
for count in class_counts:
    initial_state.append(count)
initial_state = np.array(initial_state)
```

```
335
336 #Run Markov State Transitions With Noise
337 current_state = initial_state
338 states = [initial_state.flatten()]
339 for i in range(num_states):
340     noise = np.random.normal(mean, std_dev, size=transition_matrix.shape)
341     noised_transition_matrix = transition_matrix + noise
342     row_sums = noised_transition_matrix.sum(axis=1)
343     normalized_transition_matrix = noised_transition_matrix / row_sums[:, np.
    newaxis]
344     next_state = np.dot(current_state, normalized_transition_matrix)
345     states.append(next_state.flatten())
346     current_state = next_state
347 states = np.array(states)
348
349 #Plot Results
350 plt.figure(figsize=(14, 6))
351 plt.subplot(1, 2, 1)
352 for i in range(len(initial_state[0])):
353     plt.plot(states[:, i], label=f'State {i+1}')
354 plt.title('State Evolution Over Time')
355 plt.xlabel('Time Step')
356 plt.ylabel('Population')
357 plt.legend()
358 plt.grid(True)
359 plt.subplot(1, 2, 2)
360 for i in range(len(initial_state[0])):
361     x = np.arange(len(states))
362     y = states[:, i]
363     model = LinearRegression().fit(x.reshape(-1, 1), y)
364     trend_line = model.predict(x.reshape(-1, 1))
365     plt.plot(trend_line, label=f'State {i+1} Trend')
366 plt.title('Trend Line for State Evolution')
367 plt.xlabel('Time Step')
368 plt.ylabel('Population')
369 plt.legend()
370 plt.grid(True)
371 plt.tight_layout()
372 plt.show()
373
374 #Cancer Risk Analysis
375 df = pd.read_csv('cancers_age - Sheet1 (1).csv')
376 cancer_types = df['CancerType'].unique()
377
378 #Plot the Linear Regression for Multiple Cancer Types and Print R-Squared
379 fig, ax = plt.subplots()
380 for cancer_type in cancer_types:
381     cancer_data = df[df['CancerType'] == cancer_type]
382     age = cancer_data['Age']
383     hazard_ratio = cancer_data['HazardRatio']
384     model = LinearRegression()
385     model.fit(age.values.reshape(-1, 1), hazard_ratio)
386     ax.scatter(age, hazard_ratio, label=cancer_type)
```

```
    ax.plot(age, hazard_ratio_pred, label=f'Line fit ({cancer_type})')
    r_squared = r2_score(hazard_ratio, hazard_ratio_pred)
    print(f'R-squared value for {cancer_type}: {r_squared}')
ax.set_xlabel('Age')
ax.set_ylabel('Hazard Ratio')
ax.set_title('Age versus Hazard Ratio for Each Cancer Type')
ax.legend(loc='center left', bbox_to_anchor=(1, 0.5))
plt.grid(True)
plt.show()



#End of Code
```