Identifying the Leading Factors of Significant Weight Gains Using a New Rule Discovery Method

Mina Samizadeh\textsuperscript{1}, Jessica C Jones-Smith\textsuperscript{2}, Bethany Sheridan\textsuperscript{3} and Rahmatollah Beheshti\textsuperscript{1}

\textsuperscript{1} University of Delaware, Delaware, USA.\textsuperscript{2} University of Washington, Washington, USA.\textsuperscript{3} athenahealth, Inc., Massachusetts, USA.

Contributing authors: minasnmz@udel.edu; jjoness@uw.edu; bethany.sheridan.g@gmail.com; rbi@udel.edu;

Abstract
Overweight and obesity remain a major global public health concern and identifying the individualized patterns that increase the risk of future weight gains has a crucial role in preventing obesity and numerous subsequent diseases associated with obesity. In this work, we use a rule discovery method to study this problem, by presenting an approach that offers genuine interpretability and concurrently optimizes the accuracy (being correct often) and support (applying to many samples) of the identified patterns. Specifically, we extend an established subgroup-discovery method to generate the desired rules of type $X \rightarrow Y$, and show how top features can be extracted from the $X$ side, functioning as the best predictors of $Y$. In our obesity problem, $X$ refers to the extracted features from a very large and multi-site EHR data, and $Y$ indicates significant weight gains. Using our method, we also extensively compare the differences and inequities in patterns across 22 strata determined by the individuals’ gender, age, race, insurance type, neighborhood type, and income level. Through extensive series of experiments, we show new and complementary findings regarding the predictors of future dangerous weight gains.

Keywords: Subgroup Discovery, Interpretable Patterns Mining, K-association Mining, Obesity, Health equity

\*Our code is available on our \texttt{GitHub} repository.
1 Introduction

Obesity and overweight continue to prevail around the world. It is estimated that more than 30% of the world population has been overweight or obese in recent years [1], and the forecast trends do not show signs of dampening [2]. If the growth rates in body-weights stay similar, more than 41% of the world population are estimated to be overweight or obese by 2030 [3]. At the same time, the costs of the medical support for obesity management have been increasing constantly, for instance, from $212.4 to $315.8 billion during the 2005-10 period in the US, showing a 48.7% spike [4]. As a major risk factor of chronic diseases such as diabetes [5], cancer [6–8], and cardiovascular diseases [9, 10], as well as infectious diseases such as what the world experienced with the Covid-19 disease [11, 12], obesity puts a huge burden on the human societies. Several decades of research on obesity have shown that no single predictor of obesity may exist [13, 14]. Instead, obesity is the product of complex and non-linear interactions among many different factors including biological, environmental, and social ones.

Like other fields of biomedical research, the obesity field has also faced a potentially overhauling moment, by using large and diverse datasets to study this complex disease. Aligned with the precision medicine aims, applying advanced analytical methods on these large datasets can reveal the role of the complex risk factors of obesity based on the specific characteristics of individuals. One large and popular family of such analytical methods is the family of subgroup discovery (SD) methods [15], which is closely related to subtyping [16] and electronic phenotyping methods [17–19]. SD is a data mining technique that finds desired patterns in a dataset with respect to a certain target variable [15]. This method is generally used to generate if-then rules of type $X \rightarrow Y$, where $X$ is a set of input (independent) variables, and $Y$ is a fixed target (dependent) variable. In our study, $Y$ indicates an upward shift in the obesity classes or developing obesity (such as transitioning from obese class 1 to obese class 2, or normal weight to obese). We collectively refer to our target $Y$ variable, as the dangerous weight gain patterns. What makes SD methods different from the existing predictive (often classification) and descriptive (often clustering) methods in the field, is their ability to combine both predictive and descriptive inductions. Specifically, these methods can be used to optimize a supervised prediction task (e.g., optimize the sensitivity or specificity of a classification task) and ensure a high support for the identified rules (i.e., the identified rules apply to many of the samples).

In this study, we use an SD-based method to identify the individuals who experience significant body weight increases, using a longitudinal six-year dataset containing information from more than two million patients and around 13.5 million visits. Our dataset includes information related to demographic, socioeconomic, medication, and a series of major chronic conditions derived from adult population across various sites in the US. Specifically, we customize an established SD technique [20] to generate a pool of desired $X \rightarrow Y$ rules, as introduced earlier. Using the generated rules and based a mechanism
that we propose, we then determine the most important features of the individuals (from the features appearing on the left side or $X$) that can predict the dangerous weight gains. We further study the differences across these across six demographic and socioeconomic characteristics, capturing some of the important social determinants of health (SDOH). These six variables include gender, age, race, insurance type, neighborhood type, and income levels. This way, the specific contributions of our study are:

- We extend an interpretable SD method, used for identifying the predictive rules, to identify the key features that can predict the outcome of interest. While other SD methods exist in the literature, our method for using the extracted subgroups to identify the important predictive features is new and applicable to any problem of this type.
- Our study identifies the roots of dangerous weight gains by analyzing a large, semi-nationally representative, and longitudinal dataset. We show the top features predicting dangerous weight gains and compare those features in various stratified groups based on demographic and socioeconomic factors. Our results also demonstrate the inequities with respect to dangerous weight gains.

2 Related Work

**Similar work to study obesity** — Electronic health records (EHRs) have demonstrated a great potential for studying health problems, and obesity is no exception [21] (a few of these types of studies are reviewed by Baer et al. [22]). Considering the common properties of EHR datasets (such as having large volumes and missing elements), machine learning and data mining methods have been among popular choices to work with these datasets, commonly to predict future obesity trends [23–27]. From this latter category, a closely related set of studies to our work are the studies that have used association rule mining methods, which like our method involves generating *if-then* rules (but not necessarily with a fixed right side). Examples of such studies include the work by Kim et. al to predict the diseases related to obesity [28] and the work of Hu et. al studying the composition of the gut microbial community in normal and obese adolescents [29]. Among the studies directly comparable to ours, the work by Roth et. al uses multinomial logistic regression to identify community-level factors associated with overweight and obese BMI levels [30]. In a similar study, Rifat et al. have used a naive Bayes classifier to identify the obesity risk factors using cross-sectional data from 259 individuals [31]. Besides the predictive models in the field, another group of studies, closely related to our work, use descriptive methods for extracting the characteristics of an outcome of interest. Clustering-based methods are common in this group. To name a few examples, clustering was used to find the groups of patients who could maintain the weight they had lost [32], to cluster individuals based on lifestyle risk factors in studying the links between the clusters and BMIs.
[33], and in another work to study the characteristics of the clusters of people with obesity [34].

**SD applications** – The goal of SD methods is detecting the best subsets of a dataset having a certain property concerning a target feature. Because of having desirable properties such as interpretability, they have been widely used in biomedical domains. Example applications demonstrating the wide scope of their applications include detecting risk groups with coronary heart disease [35], extracting useful information in diagnosis and prevention of brain diseases [36], finding important features contributing to cancers [37, 38], characterizing the patients who visit emergency departments [39], and interpreting the medical imaging results [40].

### 3 Dataset

In this study, we use a deidentified dataset from athenahealth, which is a large provider of network-enabled services for hospital and ambulatory clients in the US. Patients visiting athenahealth providers are broadly representative of the outpatient visits in the US, when compared to national benchmarks provided by the National Ambulatory Medical Care Survey (NAMCS) [41]. The dataset is collected from the EHRs of the patients aged 20+ years who had a visit to an athenahealth primary care provider from 2012–2017. Our study was approved by a local institutional review board at the University of Delaware.

As EHRs generally contain very diverse and heterogeneous data types (such as demographics, measurements, medications, and tests), to reduce the dimension of the dataset, only a subset of variables (that are known to have relationships with obesity) has been selected. These variables included 1) patient demographics (sex, age, race, and ethnicity), 2) socioeconomic status (type of insurance coverage, urban or rural residence, and median household income in the zip code), 3) provider type (MD, NP, PA, and RN), 4) measurements (systolic and diastolic blood pressure, hemoglobin A1c, and LDL cholesterol), 5) selected medications (from the five categories of antihypertension, antihyperlipidemic, antidepressant, antiobesity, and antidiabetic medications), and 6) selected diagnoses (prior and incident diagnosis of 18 major chronic conditions). Specifically, the 18 chronic conditions included 1) hypothyroidism, 2) stroke, 3) Alzheimer’s or dementia, 4) anemia, 5) asthma, 6) heart failure or ischemic heart disease or acute myocardial infarction (AMI), 7) benign prostatic hyperplasia (BPH), 8) chronic kidney disease (CKD), 9) cancer (breast, colon, prostate, endometrial, or lung), 10) depression, 11) diabetes, 12) hip or pelvic fracture or osteoporosis, 13) hyperlipidemia, 14) hypertension, 15) obesity, 16) rheumatic arthritis or osteoarthritis, 17) Atrial Fibrillation (AFib), and 18) Chronic Obstructive Pulmonary Disease (COPD). To further reduce the dimensionality of the variables and prepare the dataset for extracting interpretable rules, all numerical values were categorized (and one-hot encoded) in our dataset by following the standard levels used for each variable in the field. Specifically, age was categorized into 10-year buckets, and
the income-level into low, medium, and high, as determined by the American Community Survey (ACS) for each zip code. Three categories of low, normal, and high were defined for the systolic blood pressure (low as \(\leq 98\), high as \(\geq 166\), and normal as in between), for diastolic blood pressure (low as \(\leq 58\), high as \(\geq 100\), and normal as in between), for hemoglobin A1c (low as \(\leq 5\), high as \(\geq 11.7\), and normal as in between), and for LDL cholesterol (low as \(\leq 44\), high as \(\geq 189\), and normal as in between). BMI values were categorized into six categories based on the CDC (Centers for Disease Control and Prevention) categorization, as underweight \(< 18.5\) (underweight), normal \(< 25\) (normal), overweight \(< 30\) (overweight), and class 1–3 obesity \(< 35, < 40, \text{ and } 40 \leq\) (obesity class 1–3).

Visit-level (temporal) data was aggregated by calculating the mode of the measurements and replacing the chronic condition incidences with zero (if never recorded) and one (if recorded at any visit). Additionally, since prior BMIs are strongly correlated with future BMIs, we omitted features relating to BMI to avoid identifying trivial patterns in our analysis.

Following the preprocessing steps described above, we ended up with 210 variables (including 128 medication variables) per patient. From this dataset, we define the final cohort by including those patients who have at least two BMI recordings spanning over a minimum of two years. We then assign labels to the patients in the cohort, by considering anyone who develops (new) obesity or shifts upward in the obesity classes as class positive (indicated by \(\text{class} = 1\)) and rest as class negative (\(\text{class} = 0\)). We follow the standard definitions by CDC for classifying adult body weight status [42]. This way of defining the positive class is specifically targeting the risks of obesity, either in the form of the new incidence or disease exacerbation, which we refer to as the dangerous weight gains and formally define as:

\[
\text{Class} = \begin{cases} 
1, & \text{if } (BMI_s < 30 \text{ AND } \exists BMI_{s'} \geq 30) \\
1, & \text{if } (30 \leq BMI_s < 35 \text{ AND } \exists BMI_{s'} \geq 35) \\
1, & \text{if } (35 \leq BMI_s < 40 \text{ AND } \exists BMI_{s'} \geq 40) \\
0, & \text{Otherwise.} 
\end{cases}
\]

(1)

where, \(BMI_s\) refers to the first available (start) BMI recorded for a patient, and \(BMI_{s'}\) refers to “a” recorded BMI, following a minimum two-year gap that belongs to a higher obesity class. If such BMI exists, we label that patient as positive and use their data from \(BMI_s\) to (the first observed) \(BMI_{s'}\). If none of the above conditions are met, the patient is labeled as negative, and their entire available data is used. This way, those who maintain their weights (by staying in the same body-weight class) are not considered in the positive class. We study the dangerous weight gain patterns across the entire cohort and in separate 22 strata determined by six demographic and socioeconomic status categories (sex, age, race/ethnicity, insurance type, residence type, and income). Table 1 shows the number of patients with positive and negative cases in our cohort and each of these strata.
Table 1: The number of individuals in the positive and negative classes across different demographic and socioeconomic variables in our cohort.

| Variable    | Strata          | # of positive cases | # of negative cases |
|-------------|-----------------|---------------------|---------------------|
| Gender      | Women           | 10,331              | 205,683             |
|             | Men             | 7,533               | 152,269             |
| Race        | Latino          | 1,048               | 2,284               |
|             | White           | 13,650              | 264,839             |
|             | African-American| 1,442               | 33,741              |
|             | Other Races     | 437                 | 10,203              |
|             | Race Unavailable| 1,287               | 26,885              |
| Age         | Under-Thirty    | 1,307               | 14,907              |
|             | Thirties        | 1,934               | 28,864              |
|             | Forties         | 3,515               | 57,898              |
|             | Fifties         | 4,493               | 89,946              |
|             | Sixties         | 3,897               | 89,396              |
|             | Seventies       | 2,718               | 76,941              |
| Neighborhood| Metro           | 13,572              | 272,156             |
|             | Metro-Adjacent  | 3,134               | 60,945              |
|             | Rural           | 1,158               | 24,851              |
| Income      | Low-Income      | 9,017               | 183,946             |
|             | Med-Income      | 5,389               | 105,210             |
|             | High-Income     | 3,458               | 68,796              |
| Insurance   | Medicare        | 6,376               | 156,113             |
|             | Medicaid        | 1,477               | 24,606              |
|             | Commercial      | 9,444               | 167,135             |
|             | Self-Pay        | 540                 | 9,601               |
| Total       | All Patients    | 17,864              | 357,952             |

4 Method

In this work, we adopted the CN2 SD algorithm [20] for the initial extraction of rules. CN2 generates a random pool of $X \rightarrow Y$ rules (in our case, $Y$ is the patients with $class=1$), and iteratively continues to improve those rules with respect to a certain quality measure using a heuristic search algorithm called beam search. Considering each generated rule as a graph node, beam search generates new rules to explore in a breadth-first-search manner. We refer the readers to other references for in-depth discussions of the concepts related to CN2 and relevant SD methods [15]. For the quality measure, we use the weighted relative accuracy ($WRAcc$), as defined in Equation 2,

$$WRAcc = \frac{\text{Support}}{\text{dataset size}} \times (\text{Confidence} - \text{Expected Confidence})$$  \hspace{1cm} (2)

$$\text{Support} = \frac{\text{subgroup size}}{\text{dataset size}}$$  \hspace{1cm} (3)

$$\text{Confidence} = \frac{\text{positive # in subgroup}}{\text{subgroup size}}$$  \hspace{1cm} (4)

$$\text{Expected Confidence} = \frac{\text{positive # in subgroup}}{\text{dataset size}}$$  \hspace{1cm} (5)
where \( WRAcc \) (of a rule \( r \)) is defined using the support and confidence quality measures. Considering the general format of \( X \rightarrow Y \) for the rule \( r \), one can consider support as \( P(X) \) (showing the probability of \( X \)), confidence as \( P(Y \mid X) \), and expected confidence as \( P(Y) \). What makes \( WRAcc \) especially relevant for our study is its ability to combine both generalizability (support) and accuracy (Confidence – Expected Confidence; the right side of Equation 2) that facilitates capturing both descriptive and predictive patterns across the dataset. The descriptive patterns are captured by the support (by ensuring that the patterns apply to many), and the predictive patterns are captured by the right side of Equation 2. This right side of the equation compares the accuracy of the rule with respect to a naive classification which considers all the cases as positive. This part normalizes the bias between the classes, since it depends on the proportion of the class that is being analyzed to the size of whole dataset. The higher value for \( WRAcc \) is interpreted as a higher balance between the generality of the pattern and its confidence, and as a result, a higher quality of the subgroup (identified by a rule). This formulation is also well suited for imbalanced datasets, such as ours, as it directly considers the size of each of the classes and adjusts the final value accordingly. After ranking all possible rules (with a certain length) using \( WRAcc \), we keep those rules with the \( WRAcc \) more than a threshold, set empirically in our main experiments. From these rules, we then identify the most important features appearing on the left side (of the \( X \rightarrow Y \) rules), using an average quality approach. In this approach, for each feature, we calculate the weighted average \( (A_{W}) \) of the \( WRAcc \) of the remained rules (those with \( WRAcc \) above the threshold) in which the feature has appeared, as shown in Equation 6.

\[
A_{W}(f_i) = \frac{\sum_{\text{rules with } f_i} WRAcc}{\# \text{ of rules with } f_i} \quad (6)
\]

where \( A_{W} \) shows the weighted average and \( f_i \) is the \( i \)th feature. A higher weighted average quality can indicate the higher importance of a feature in predicting \( Y \) (in our case, dangerous weight gains). A schematic representation of the method we use for generating the rules and identifying the important features from those is presented in Fig 1.
5 Experiments and Discussion

In our experiments, we first study the differences in the identified dangerous weight gain patterns across the entire study population, and then compare those versus the patterns obtained from any of the six strata indicated by the six demographic and socioeconomic characteristics. We refer to the 22 subgroups identified by these six characteristics as “strata” (instead of subgroups), to avoid confusion with the term subgroups used in the SD method. The notion of being a “top feature” in our study combines accuracy (features that most accurately predict weight gain) and support (features that apply to a large portion of samples). Choosing only accuracy may lead to identifying patterns that apply to few numbers of individuals and choosing only support may lead to identifying patterns that are not accurate enough. We run our experiments using three maximum lengths for the rules (referring to the $X \rightarrow Y$ rules, as $Y$ is fixed in all cases) with values of three, four, and five; and three beam search widths (the number of generated rules in iterations) of 2,000, 5,000, and 10,000. Each bar in the results presented in this section (Fig 2 to Fig 4) corresponds to the average value across the nine experimental settings that we have used in our main experiments (three different width of beam search by three different numbers of generated rules). Similarly, the error bars in the figures show the variation (95% confidence interval) across the experimental settings. We use $5.0e^{-4}$ (or 0.0005), as the $WRAcc$ threshold, which we empirically identified in our experiments. Besides the nine settings used here, we present more in-depth sensitivity analyses in Appendix A. Also, in Appendix B, we present the results related to the similar experiments to those reported here, but by further combining the 128 medication variables into five categories of antihypertension, antihyperlipidemic, antidepressant, antiobesity, and antidiabetic medications. When studying the whole cohort (Fig. 2), having a commercial type of insurance is found to be the top factor predicting dangerous weight gains. This is in line with several studies reporting the overall importance of social determinant of health (SDOH) in determining the risks of developing obesity [43, 44]. The second top feature is taking selective serotonin reuptake inhibitors (SSRIs), which are generally used as antidepression medications, and gaining weight is known to be among their major side effects [45–48]. The third top feature is being white. While obesity is known to be more prevalent among non-Hispanic blacks [49], identifying the white race as a top feature in predicting dangerous weight gains may indicate complementary patterns. We note that our method (through $WRAcc$) already accounts for imbalanced nature of samples. For instance, while white race is the dominant race in our samples, appearing on top of the important features list is not solely due to its higher frequency. Living in metro areas is identified as the fourth feature, which is in line with the studies reporting more sedentary lifestyle and consuming less healthy foods in metropolitan areas [50, 51]. A similar pattern to the third feature (race=white) can be also observed for the fifth feature (gender=man), where obesity is known to be more prevalent among women than men in the US [49, 52]. Being in the 40s, under 30s, and
Fig. 2: Top 10 features with highest weighted average quality across the whole cohort. (Dx=diagnosis, SSRIs= “Selective Serotonin Reuptake Inhibitors”, an antidepressant medicine, Ins=insurance, Metro=living in a metro area)

30s are placed in sixth, seventh, and ninth features, aligned with the studies reporting that the middle aged groups (40s and 50s) suffer more from obesity [49, 52]. Depression is ranked eighth (the Dx prefix indicates the diagnosis of a disease). Like SSRIs, the positive association between obesity and depression is reported by many studies [53–57]. The tenth feature is having low income, which is another major SDOH, widely known as a major risk factor in developed countries (a phenomena sometime called “reverse gradient”) [58].

Following the top features across the whole cohort in Fig. 2, the six charts in Fig. 3 show where these “same top 10 features” stand across the 22 separate strata. While most patterns observed in Fig. 3 remain unchanged compared to the whole cohort as shown in Fig. 2, many new patterns can be also observed. In following, we cover some of the more prominent patterns. In Fig. 3.A, ranking of the top features in Fig. 2 are shown across different races. Here, it can be seen that taking SSRIs and depression have higher ranks in Black-Americans.

In Fig. 3.B (stratified by insurance type), one can observe that SSRIs, Under-30s, 30s, and 40s have values equal to zero in the Medicare strata. This is reasonable, as Medicare is only offered to 65+ yr individuals. In Fig. 3.C (stratified by income levels), in the middle-income category the Men feature has a zero value, which can be compared to the studies showing that low- and high-income men have higher risks of obesity [58, 59]. Fig. 3.E (stratified by gender) shows that in men, depression is not among the top 10 features and SSRIs are ranked fourth. Fig. 3.F (stratified by age), shows that a commercial insurance (which was found to be the most important in entire cohort analysis) loses its importance as the age goes up to the extent that it is not among the top 10 factors in the category of Age_70s. Additionally, the importance of the features related to depression, such as SSRIs and Dx_Depression, decreases as the age increases. Specifically, Dx_Depression feature is not an important factor in age categories of 60s and 70s.

Similar to our first experiment, we also identify the top features in any of the 22 strata as identified by the six different demographic and
Fig. 3: Comparing where the top 10 features shown in Fig. 2 (shown by the yellow bars here) stand in 22 strata determined by six demographics and socioeconomic factors: A) race, B) insurance type, C) income level, D) neighborhood type, E) gender, and F) age.

socioeconomic characteristics (Fig. 4). Note that this time not the fixed 10 features determined from the whole cohort are used. Fig 4.1 shows the important features across different races. Among noticeable patterns, in Fig 4.1.B (Black-American stratum), hypertension-related features (including PriorDx_Hypertens, Dx_Hypertens) and Hyperlipidemia medications (Ace Inhibitors, Ace Inhibitors and combination), CCB (Calcium Channel Blockers), and CCBs- Dihydropyridines are important factors in this cohort. Hypertension prevalence is higher among minority groups in comparison to
Fig. 4: Top 10 features with highest weighted average quality across 22 different strata indicated by: 1) race, 2) insurance, 3) income, 4) neighborhood, 5) gender, and 6) age.
whites [60], and the association of hyperlipidemia and obesity have been reported in other studies [61–64], as well as a higher prevalence in Black-Americans in comparison to the whites [65]. Hyperlipidemia is a condition in which patient has high levels of lipids (bad cholesterol and triglycerides) in their blood [66]. The lipid abnormalities include higher serum triglycerides, VLDL, apolipoprotein B, and non-HDL-C levels [67]. In Fig 4.1.C (Latino stratum) Med-income and Medicaid insurance are showing up among the top 10 important factors. Fig 4.2, shows the top features across different insurance types. Here, SSRI, Dx_Depression, and PriorDx_Depression are not important in Medicare insurance category, unlike the other insurance types. Instead, 60s, 50s, Priordx_Hyperlipid, Dx_Hyperlipid, Priordx_Hypertens, and Dx_Hypertens are ranked high among patients having the Medicare insurance type. An interesting pattern with respect to Fig 4.2.B (self-pay stratum) is the larger uncertainty ranges compared to other figures (as shown by the error bars), possibly due to the greater heterogeneity in the characteristics of self-pay patients. Fig 4.3 (income-level strata) shows that the features related to hyperlipidemia (such as Priordx_Hyperlipid, and Dx_Hyperlipid are among the top features in the high-income stratum. In Fig 4.4 (neighborhood strata), having the Medicaid insurance is identified as a more important factor for the patients in rural areas. Fig 4.5 (gender strata), shows that Dx_Depression and PriorDx_Depression are among the important factors in women and SSRI in the both strata. This pattern is conforming with recent studies reporting a positive association of depression with obesity in general population and a higher association among women [68–70]. The results also indicate Priordx_Hyperlipid and Dx_Hyperlipid are among important features in the men strata which is consistent with other studies [71–73]. In Fig 4.6 (age strata), a noticeable pattern relates to decreasing the rank of taking antidepression medications, as the age increases, which relates to the reports showing higher prevalence of depression among younger people specifically those in 24-30 [74]. Having a commercial insurance is not among the top for the 60s and 70s strata. On the other side, PriorDx_Hyperlipid, Dx_Hyperlipid, and antihyperlipidemic class of medications (such as statins and antihyperlipidemic – HMG CoA Reductase Inhibitors) are important in older patients (those in their 50s, 60s, and 70s). The ranking of hyperlipidemia and antihyperlipidemia medication also increases with the patient’s age, aligned with the higher prevalence of hyperlipidemia in older adults [75, 76]. Other noticeable features are the anti-hypertension CCBs (Calcium Channel Blockers) and dihydropyridines (also used as antihypertension) are ranked high in the age 60s stratum. Medicare is also an important feature among those in their 60s and 70s. Dx_Obesity appears as top feature only in the under 30s and 70s strata stratum. More details about the specified feature across different categories are available in Appendix C.

Lastly, we note some limitations of our study. We only study the weight gain patterns within a two-year frame based on the considerations about not choosing a too short period and not excluding too many samples. The two-year
A period may still offer a reasonable gap for engaging in timely interventions before seeing the adverse outcomes of excessive weight gain. Our approach does not consider all the concerning patterns of dangerous weight gain such as transitioning from the normal to overweight. Still, our work captures some of the most concerning weight patterns. The way we define our positive class also captures weight cycling patterns (also known as the yo-yo effect) [77]. Our results are based on the analysis in one dataset. Despite this, our study is among the largest of this kind. We also tried to address many of the concerns of this type by running significant sensitivity analyses demonstrating the robustness of our results.

6 Conclusion

In this study, aiming to identify the most important factors that lead to dangerous weight gains, we generated a series of $X \rightarrow Y$ (if-then) rules predicting such patterns using a very large longitudinal dataset. To do this, we presented a new subgroup discovery method inspired by existing data mining methods. In our method, we first rank the generated rules using the $WRAcc$ (weighted relative accuracy) value for each rule, and then determine the most important features on the left-side that can predict the right side. We identify these important features by calculating the weighted average $WRAcc$ of the top $n$ rules that each feature appears in. Through a series of sensitivity analysis experiments, we show that our results are robust to the parameter choices. Applying this method to our weight gains problem reveal various patterns in our large dataset and across 22 different strata (determined by different demographic and socioeconomic factors). Specifically, we show how the most important features predicting dangerous weight gains differ across individuals with different sex, age, race/ethnicity, insurance type, residence type, and income-level. Our findings may inform more effective interventions of obesity. Moreover, our application agnostic rule-discovery method can be applied to other similar problems.

References

[1] for Disease Control, C., Prevention, et al.: Adult obesity facts Overweight & obesity (2020)

[2] Dobbs, R., Sawers, C., Thompson, F., Manyika, J., Woetzel, J.R., Child, P., McKenna, S., Spatharou, A.: Overcoming obesity: An initial economic analysis (2014)

[3] Shaw, J.E., Sicree, R.A., Zimmet, P.Z.: Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes research and clinical practice 87(1), 4–14 (2010)
[4] Biener, A., Cawley, J., Meyerhoefer, C.: The high and rising costs of obesity to the US health care system. Springer (2017)

[5] Dandona, P., Aljada, A., Bandyopadhyay, A.: Inflammation: the link between insulin resistance, obesity and diabetes. Trends in immunology 25(1), 4–7 (2004)

[6] Vucenik, I., Stains, J.P.: Obesity and cancer risk: evidence, mechanisms, and recommendations. Annals of the New York Academy of Sciences 1271(1), 37 (2012)

[7] De Pergola, G., Silvestris, F.: Obesity as a major risk factor for cancer. Journal of obesity 2013 (2013)

[8] Barb, D., Pazaitou-Panayiotou, K., Mantzoros, C.S.: Adiponectin: a link between obesity and cancer. Expert opinion on investigational drugs 15(8), 917–931 (2006)

[9] Nakamura, K., Fuster, J.J., Walsh, K.: Adipokines: a link between obesity and cardiovascular disease. Journal of cardiology 63(4), 250–259 (2014)

[10] Ritchie, S., Connell, J.: The link between abdominal obesity, metabolic syndrome and cardiovascular disease. Nutrition, Metabolism and cardiovascular diseases 17(4), 319–326 (2007)

[11] Dietz, W., Santos-Burgoa, C.: Obesity and its implications for covid-19 mortality. Obesity (Silver Spring) 28(6), 1005 (2020)

[12] Korakas, E., Ikonomidis, I., Kousathana, F., Balampanis, K., Kountouri, A., Raptis, A., Palaiodimou, L., Kokkins, A., Lambadiari, V.: Obesity and covid-19: immune and metabolic derangement as a possible link to adverse clinical outcomes. American Journal of Physiology-Endocrinology and Metabolism 319(1), 105–109 (2020)

[13] Malik, V.S., Willett, W.C., Hu, F.B.: Global obesity: trends, risk factors and policy implications. Nature Reviews Endocrinology 9(1), 13–27 (2013)

[14] Mokdad, A.H., Ford, E.S., Bowman, B.A., Dietz, W.H., Vinicor, F., Bales, V.S., Marks, J.S.: Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. Jama 289(1), 76–79 (2003)

[15] Herrera, F., Carmona, C.J., González, P., Del Jesus, M.J.: An overview on subgroup discovery: foundations and applications. Knowledge and information systems 29(3), 495–525 (2011)

[16] Wang, S., Sparks, L., Xie, H., Greenway, F., De Jonge, L., Smith, S.:
Subtyping obesity with microarrays: implications for the diagnosis and treatment of obesity. International Journal of Obesity 33(4), 481–489 (2009)

[17] Blundell, J.E., Dulloo, A.G., Salvador, J., Frühbeck, G., et al.: Beyond bmi-phenotyping the obesities. Obesity Facts 7(5), 322–328 (2014)

[18] Hong, N., Wen, A., Stone, D.J., Tsuji, S., Kingsbury, P.R., Rasmussen, L.V., Pacheco, J.A., Adekkanattu, P., Wang, F., Luo, Y., et al.: Developing a fhir-based ehr phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries. Journal of biomedical informatics 99, 103310 (2019)

[19] Banda, J.M., Seneviratne, M., Hernandez-Boussard, T., Shah, N.H.: Advances in electronic phenotyping: from rule-based definitions to machine learning models. Annual review of biomedical data science 1, 53–68 (2018)

[20] Clark, P., Niblett, T.: The cn2 induction algorithm. Machine learning 3(4), 261–283 (1989)

[21] Bailey, L.C., Milov, D.E., Kelleher, K., Kahn, M.G., Del Beccaro, M., Yu, F., Richards, T., Forrest, C.B.: Multi-institutional sharing of electronic health record data to assess childhood obesity. PloS one 8(6), 66192 (2013)

[22] Baer, H.J., Cho, I., Walmer, R.A., Bain, P.A., Bates, D.W.: Using electronic health records to address overweight and obesity: a systematic review. American Journal of Preventive Medicine 45(4), 494–500 (2013)

[23] Gupta, M., Phan, T.-L.T., Datto, G., Bunnell, T., Beheshti, R.: An interpretable prediction model for obesity prediction using ehr data

[24] Gupta, M., Phan, T.-L.T., Bunnell, T., Beheshti, R.: Obesity prediction with ehr data: A deep learning approach with interpretable elements. arXiv preprint arXiv:1912.02655 (2019)

[25] Dugan, T.M., Mukhopadhyay, S., Carroll, A., Downs, S.: Machine learning techniques for prediction of early childhood obesity. Applied clinical informatics 6(03), 506–520 (2015)

[26] Zhang, S., Tjortjis, C., Zeng, X., Qiao, H., Buchan, I., Keane, J.: Comparing data mining methods with logistic regression in childhood obesity prediction. Information Systems Frontiers 11(4), 449–460 (2009)

[27] Campbell, E.A., Qian, T., Miller, J.M., Bass, E.J., Masino, A.J.: Identification of temporal condition patterns associated with pediatric obesity
incidence using sequence mining and big data. International Journal of Obesity 44(8), 1753–1765 (2020)

[28] Kim, S., An, Y.-J., Lee, K.-H., Jung, S.-H., Cho, W.-S.: Using association analysis to find diseases related to childhood obesity. In: 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), pp. 847–850 (2017). IEEE

[29] Hu, H.-J., Park, S.-G., Jang, H.B., Choi, M.-G., Park, K.-H., Kang, J.H., Park, S.I., Lee, H.-J., Cho, S.-H.: Obesity alters the microbial community profile in korean adolescents. PloS one 10(7), 0134333 (2015)

[30] Roth, C., Foraker, R.E., Payne, P.R., Embi, P.J.: Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis. BMC medical informatics and decision making 14(1), 1–8 (2014)

[31] Hossain, R., Mahmud, S.H., Hossin, M.A., Noori, S.R.H., Jahan, H.: Prmt: Predicting risk factor of obesity among middle-aged people using data mining techniques. Procedia computer science 132, 1068–1076 (2018)

[32] Ogden, L.G., Stroebelbe, N., Wyatt, H.R., Catenacci, V.A., Peters, J.C., Stuht, J., Wing, R.R., Hill, J.O.: Cluster analysis of the national weight control registry to identify distinct subgroups maintaining successful weight loss. Obesity 20(10), 2039–2047 (2012)

[33] Schuit, A.J., van Loon, A.J.M., Tijhuis, M., Ocké, M.C.: Clustering of lifestyle risk factors in a general adult population. Preventive medicine 35(3), 219–224 (2002)

[34] Green, M., Strong, M., Razak, F., Subramanian, S., Relton, C., Bissell, P.: Who are the obese? a cluster analysis exploring subgroups of the obese. Journal of public health 38(2), 258–264 (2016)

[35] Gamberger, D., Lavrac, N.: Generating actionable knowledge by expert-guided subgroup discovery. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 163–175 (2002). Springer

[36] Gamberger, D., Lavrač, N., Krstačić, A., Krstačić, G.: Clinical data analysis based on iterative subgroup discovery: experiments in brain ischaemia data analysis. Applied Intelligence 27(3), 205–217 (2007)

[37] Tan, P.-N., Steinbach, M., Kumar, V.: Data mining introduction. Bei Jing: The people post and Telecommunications Press (2006)

[38] Trajkovski, I., Zelezny, F., Lavrac, N., Tolar, J.: Learning relational descriptions of differentially expressed gene groups. IEEE Transactions
on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 38(1), 16–25 (2007)

[39] Carmona, C.J., González, P., del Jesus, M.J., Navío-Acosta, M., Jiménez-Trevino, L.: Evolutionary fuzzy rule extraction for subgroup discovery in a psychiatric emergency department. Soft Computing 15(12), 2435–2448 (2011)

[40] Schmidt, J., Hapfelmeier, A., Mueller, M., Perneczky, R., Kurz, A., Drzezga, A., Kramer, S.: Interpreting pet scans by structured patient data: a data mining case study in dementia research. Knowledge and Information Systems 24(1), 149–170 (2010)

[41] for Disease Control, C., Prevention, et al.: National Ambulatory Medical Care Survey. https://www.cdc.gov/nchs/data/ahcd/NAMCS_30A_2010. pdf Accessed 2021-09-10

[42] for Disease Control, C., Prevention: Overweight and Obesity, Defining Adult Overweight and Obesity. https://www.cdc.gov/obesity/adult/defining.html Accessed 2021-06-07

[43] Yusuf, Z.I., Dongarwar, D., Yusuf, R.A., Bell, M., Harris, T., Salihu, H.M.: Social determinants of overweight and obesity among children in the united states. International Journal of Maternal and Child Health and AIDS 9(1), 22 (2020)

[44] Bryant, P.H., Hess, A., Bowen, P.G.: Social determinants of health related to obesity. The Journal for Nurse Practitioners 11(2), 220–225 (2015)

[45] Ferguson, J.M.: Ssri antidepressant medications: adverse effects and tolerability. Primary care companion to the Journal of clinical psychiatry 3(1), 22 (2001)

[46] Cascade, E., Kalali, A.H., Kennedy, S.H.: Real-world data on ssri antidepressant side effects. Psychiatry (Edgmont) 6(2), 16 (2009)

[47] Patten, S.B., Williams, J.V., Lavorato, D.H., Brown, L., McLaren, L., Eliasziw, M.: Major depression, antidepressant medication and the risk of obesity. Psychotherapy and psychosomatics 78(3), 182–186 (2009)

[48] Dannon, P.N., Iancu, I., Lowengrub, K., Gonopolsky, Y., Musin, E., Grunhaus, L., Kotler, M.: A naturalistic long-term comparison study of selective serotonin reuptake inhibitors in the treatment of panic disorder. Clinical neuropharmacology 30(6), 326–334 (2007)

[49] Fryar, C.D., Carroll, M.D., Afful, J.: Prevalence of overweight, obesity,
and severe obesity among adults aged 20 and over: United states, 1960–1962 through 2017–2018. NCHS Health E-Stats (2020)

[50] Bodor, J.N., Rice, J.C., Farley, T.A., Swalm, C.M., Rose, D.: The association between obesity and urban food environments. Journal of Urban Health 87(5), 771–781 (2010)

[51] Ewing, R., Schmid, T., Killingsworth, R., Zlot, A., Raudenbush, S.: Relationship between urban sprawl and physical activity, obesity, and morbidity. American journal of health promotion 18(1), 47–57 (2003)

[52] Fryar, C.D., Kruszan-Moran, D., Gu, Q., Ogden, C.L.: Mean body weight, weight, waist circumference, and body mass index among adults: United states, 1999–2000 through 2015–2016 (2018)

[53] Roberts, R.E., Deleger, S., Strawbridge, W.J., Kaplan, G.A.: Prospective association between obesity and depression: evidence from the alameda county study. International journal of obesity 27(4), 514–521 (2003)

[54] Luppino, F.S., de Wit, L.M., Bouvy, P.F., Stijnen, T., Cuijpers, P., Penninx, B.W., Zitman, F.G.: Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Archives of general psychiatry 67(3), 220–229 (2010)

[55] Dong, C., Sanchez, L., Price, R.: Relationship of obesity to depression: a family-based study. International journal of obesity 28(6), 790–795 (2004)

[56] Faith, M.S., Matz, P.E., Jorge, M.A.: Obesity–depression associations in the population. Journal of psychosomatic research 53(4), 935–942 (2002)

[57] Preiss, K., Brennan, L., Clarke, D.: A systematic review of variables associated with the relationship between obesity and depression. Obesity Reviews 14(11), 906–918 (2013)

[58] Bentley, R.A., Ormerod, P., Ruck, D.J.: Recent origin and evolution of obesity-income correlation across the united states. Palgrave Communications 4(1), 1–14 (2018)

[59] Ogden, C.L., Fakhouri, T.H., Carroll, M.D., Hales, C.M., Fryar, C.D., Li, X., Freedman, D.S.: Prevalence of obesity among adults, by household income and education—united states, 2011–2014. MMWR. Morbidity and mortality weekly report 66(50), 1369 (2017)

[60] Fei, K.: Racial and ethnic subgroup disparities in hypertension prevalence, new york city health and nutrition examination survey, 2013–2014. Preventing chronic disease 14 (2017)
[61] Sullivan, P.W., Ghushchyan, V.H., Ben-Joseph, R.: The impact of obesity on diabetes, hyperlipidemia and hypertension in the united states. Quality of Life Research 17(8), 1063–1071 (2008)

[62] Bozkurt, B., Aguilar, D., Deswal, A., Dunbar, S.B., Francis, G.S., Horwich, T., Jessup, M., Kosiborod, M., Pritchett, A.M., Ramasubbu, K., et al.: Contributory risk and management of comorbidities of hypertension, obesity, diabetes mellitus, hyperlipidemia, and metabolic syndrome in chronic heart failure: a scientific statement from the american heart association. Circulation 134(23), 535–578 (2016)

[63] Robbins, I.M., Newman, J.H., Johnson, R.F., Hennes, A.R., Fremont, R.D., Piana, R.N., Zhao, D.X., Byrne, D.W.: Association of the metabolic syndrome with pulmonary venous hypertension. Chest 136(1), 31–36 (2009)

[64] Klop, B., Elte, J.W.F., Cabezas, M.C.: Dyslipidemia in obesity: mechanisms and potential targets. Nutrients 5(4), 1218–1240 (2013)

[65] Morris, A., Ferdinand, K.: Hyperlipidemia in racial/ethnic minorities: differences in lipid profiles and the impact of statin therapy. Clinical Lipidology 4(6), 741–754 (2009)

[66] Havel, R.J., Rapaport, E.: Management of primary hyperlipidemia. New England Journal of Medicine 332(22), 1491–1498 (1995)

[67] Feingold, K.R.: Obesity and dyslipidemia. Endotext [Internet] (2020)

[68] De Wit, L., Luppino, F., van Straten, A., Penninx, B., Zitman, F., Cuijpers, P.: Depression and obesity: a meta-analysis of community-based studies. Psychiatry research 178(2), 230–235 (2010)

[69] Simon, G.E., Ludman, E.J., Linde, J.A., Operskalski, B.H., Ichikawa, L., Rohde, P., Finch, E.A., Jeffery, R.W.: Association between obesity and depression in middle-aged women. General hospital psychiatry 30(1), 32–39 (2008)

[70] Li, L., Gower, B.A., Shelton, R.C., Wu, X.: Gender-specific relationship between obesity and major depression. Frontiers in endocrinology 8, 292 (2017)

[71] Deng, B., Luo, T., Huang, Y., Shen, T., Ma, J.: Prevalence and determinants of hyperlipidemia in moderate altitude areas of the yunnan-kweichow plateau in southwestern china. High altitude medicine & biology 13(1), 13–21 (2012)

[72] Brown, C.D., Higgins, M., Donato, K.A., Rohde, F.C., Garrison, R.,
Obarzanek, E., Ernst, N.D., Horan, M.: Body mass index and the prevalence of hypertension and dyslipidemia. Obesity research 8(9), 605–619 (2000)

[73] Kawada, T.: Body mass index is a good predictor of hypertension and hyperlipidemia in a rural Japanese population. International journal of obesity 26(5), 725–729 (2002)

[74] Cuijpers, P., Karyotaki, E., Eckshtain, D., Ng, M.Y., Corteselli, K.A., Noma, H., Quero, S., Weisz, J.R.: Psychotherapy for depression across different age groups: A systematic review and meta-analysis. JAMA psychiatry 77(7), 694–702 (2020)

[75] Pencina, M.J., Navar-Boggan, A.M., D’Agostino Sr, R.B., Williams, K., Neely, B., Sniderman, A.D., Peterson, E.D.: Application of new cholesterol guidelines to a population-based sample. N Engl J Med 370, 1422–1431 (2014)

[76] Navar-Boggan, A.M., Peterson, E.D., D’Agostino Sr, R.B., Neely, B., Sniderman, A.D., Pencina, M.J.: Hyperlipidemia in early adulthood increases long-term risk of coronary heart disease. Circulation 131(5), 451–458 (2015)

[77] Brownell, K.D., Rodin, J.: Medical, Metabolic, and Psychological Effects of Weight Cycling. Archives of Internal Medicine 154(12), 1325–1330 (1994). https://doi.org/10.1001/archinte.1994.00420120035004

Mina Samizadeh is a Ph.D. student at University of Delaware. She works on the application of Machine Learning in solving health related problems.
Jessica Jones-Smith is an associate professor at University of Washington who investigates socioeconomic causes and correlates of obesity risk in both high- and low/middle-income countries.

Bethany Sheridan was leading athenahealth’s Research and Insights team, a group of researchers and data scientists dedicated to expanding knowledge on clinical operations, care patterns, and public health trends from athenahealth’s unique data asset.

Rahmat Beheshti is an assistant professor at University of Delaware. He works in the area of Applied Machine Learning and Health Data Science and directs the healthy IAife lab at UD.
Appendix A

(1) Race
(2) Insurance

(3) Income
(4) Neighborhood

(5) Gender
Fig. A1: Sensitivity Analysis of Weighted Average Quality of important features in low-level (medication-level) data contributing to obesity across 6 different groups of 1) Race, 2) Insurance, 3) Income, 4) Neighborhood, 5) Gender, and 6) Age in 9 different experimental settings. The vertical (Y) axis demonstrates WA, the X axis demonstrate the setting of experiment (width of the beam search_max length of the rules), and the Z axis shows the name of features.
Appendix B

Fig. B2: Comparison between average weighted quality of visit level (high-level) features with the highest summation across different categories of different groups consisting of: A) Race, B) Insurance, C) Income, D) Neighborhood, E) gender, F) Age.
Appendix C

Fig. C3: A comparison of weighted average quality of features association to obesity with highest summation across different categories of different groups including A) Race, B) Insurance types, C) Income-Level, D) Neighborhood areas, E) Genders, F) Age categories