Kunal Shukla

Lauren Siewny

Simran Singh

Ifrah Sohail

Qi Tan

MSQM: HA Group 2

July 31, 2024

An Investigation of the Influence of Genetic Predisposition on the Development of Cardiometabolic Conditions

Introduction

In the rapidly evolving field of healthcare analytics, our research team set out to investigate new identifiable risk factors for coronary heart disease (CHD). As CHD is the leading cause of death in the United States and worldwide, targeting measures that prevent its development is of critical clinical value [1]. It is well known that there are a variety of cardiometabolic risk factors, such as high body mass index (BMI), that are risk factors for the development of CHD [2]. What is not well known, however, is if there are identifiable genetic risk factors for the development of these diagnoses. In other words, what role does genetic predisposition play in the development of cardiometabolic conditions, therefore putting the patient at risk of the development of CHD? Or does CHD itself have a genetic risk factor?

Our theory is that if genetic factors do play a role, an area of focus could be early detection and intervention in the population with genetic risk. In doing so, care could be individualized and targeted to the patient's specific risk factors as well as social needs. More globally, this project will leverage advanced data analytics to contribute to the broader understanding of cardiovascular health, ultimately aiding in the identification and treatment of a culturally and economically impactful disease. This research aims to take a step in the direction of shedding light on the complex interplay between genetics and health, paving the way for personalized medicine approaches and lifestyle management.

Data Management

The proposed project utilized the Coherent Dataset, a synthetic collection of approximately 290,000 Electronic Health Records (EHRs). This dataset is designed to emulate the diversity and complexity of real-world EHRs, encompassing a comprehensive array of patient care data. It included patient demographics, clinical conditions, observations, medications, encounters, procedures, care plans, payers, and payer transitions. This dataset was obtained from Synthea's repository, ensuring compliance with all privacy and security regulations (e.g., HIPAA) given it is synthetic data. Upon acquisition, the individual components of the Coherent Dataset were in CSV format, segmented by different aspects of patient demographics and care. The first step involved loading these CSV files into Python using pandas, a versatile data manipulation library.

We then merged these files into a unified dataset using patient ID as the primary key, ensuring comprehensive data linkage across all relevant data points.

To facilitate meaningful analyses, we conducted any necessary data cleaning and transformation procedures. This includes checking for missing values or lost values in the merging of data and understanding the social demographics of the patients in the dataset. Python's capabilities in data preprocessing and transformation were leveraged to streamline these processes efficiently. Quality assurance measures were implemented throughout the data management process. We validated data integrity at each stage, conducting thorough checks for accuracy, completeness, and consistency.

The integrated and cleaned Coherent Dataset was securely stored on a password-protected shared drive. Access was granted only to our MSQM: HA Group 2 involved in the research project, ensuring compliance with institutional data management policies and regulatory requirements. A comprehensive data management plan was developed and implemented to document all aspects of data acquisition, integration, cleaning, transformation, quality assurance, storage, and security to ensure transparency, reproducibility, and ethical handling of the data throughout the project life cycle.

Study Design

Our retrospective study design focuses on a cohort of distinct patients with recorded BMI values and at least one of the following conditions: CHD, hyperlipidemia, hypertension, or prediabetes. Patients not meeting these criteria were excluded from the analysis. Initial exploratory data analysis (EDA) assessed the demographic distribution of the patient population, revealing a diverse representation across age, gender, marital status, and race. The relationship between BMI and CHD was investigated, with the highest recorded BMI value per patient used to mitigate potential bias.

In order to establish the link between BMI and CHD, we started by calculating the average and middle BMI values for patients with and without CHD. We then measured the correlation coefficient between BMI and CHD diagnosis to assess the strength and direction of the relationship. Subsequently, we used the matplotlib and seaborn libraries to create visual representations of the data to aid in our interpretation of the connection between BMI and CHD. We then focused on demonstrating the interaction between genetic predisposition and the listed cardiometabolic syndromes. Specifically, our aim was to explore associations between genetic variations in the genes LDLR and PON1 and different cardiometabolic conditions.

Our initial steps in data analysis involved tallying the genetic variations by chromosome and identifying the chromosome with the highest number of variants. We then pinpointed the top genes linked to pathogenic risk factors. We filtered the data for variations in the identified chromosome of interest and delved into the demographics and clinical characteristics of the affected patients. Our comprehensive statistical analysis encompassed conducting a chi-square test to evaluate associations between genetic variations and the cardiometabolic conditions of interest. Next, we performed a logistic regression to further investigate these associations, generating odds ratios and confidence intervals. To visualize these odds ratios for cardiometabolic disorders based on genetic variations, we created a forest plot. Finally, we

constructed an ROC curve and calculated area under the curve (AUC) scores to evaluate the predictive performance of the model for each cardiometabolic condition. Overall, our study leveraged a comprehensive dataset to analyze specific clinical risk factors such as BMI and their connection to CHD. We also employed statistical analysis and visualization techniques to explore relationships between genetic variations and cardiometabolic disorders, thus offering insights into potential associations and predictive capabilities.

Demographic Distribution

Demographic	Category	Count	Percentage
Gender			
	M	1978.0	55.89%
	F	1561.0	44.11%
Marital Status			
	M	2604.0	80.20%
	S	643.0	19.80%
Race			
	white	2978.0	84.15%
	black	316.0	8.93%
	asian	233.0	6.58%
	native	9.0	0.25%
	other	3.0	0.08%
Age Group			
	0-10	0.0	0.00%
	11-20	0.0	0.00%
	21-30	216.0	8.53%
	31-40	204.0	8.06%
	41-50	229.0	9.04%
	51-60	252.0	9.95%
	61-70	247.0	9.76%
	71-80	297.0	11.73%
	81-90	422.0	16.67%
	91-100	665.0	26.26%

Table 1

Results

Our analysis began by thoroughly examining the demographics within the Coherent Dataset, revealing a diverse range of ages, genders, marital statuses, and races. The dataset included patients aged from 21 to over 100 years, with the majority (approximately 26%) falling within the 91-100-year age range. Gender distribution was relatively balanced, with 44% female and 55% male patients. Marital status data showed that a significant portion of the patients were married (approximately 80.2%), while the remaining 19.8% were single, divorced, or widowed. The racial composition of the dataset included White (84.2%), African American (8.9%), Asian (6.6%), Native Americans (0.25%), and other races (0.08%) (see Table 1).

In our exploration of the relationship between BMI and CHD, we initially observed no significant difference in BMI between patients with CHD (N = 562 unique patients with 10,152 total encounters) and the general patient population (N = 3,403 unique patients with 598,650

total encounters). The mean BMI among all patients was 28.5, and the median BMI was 28.0. Among patients with CHD, the average BMI was 28.2, and the median BMI was 27.9. This initial finding suggested that BMI may not have a straightforward association with CHD.

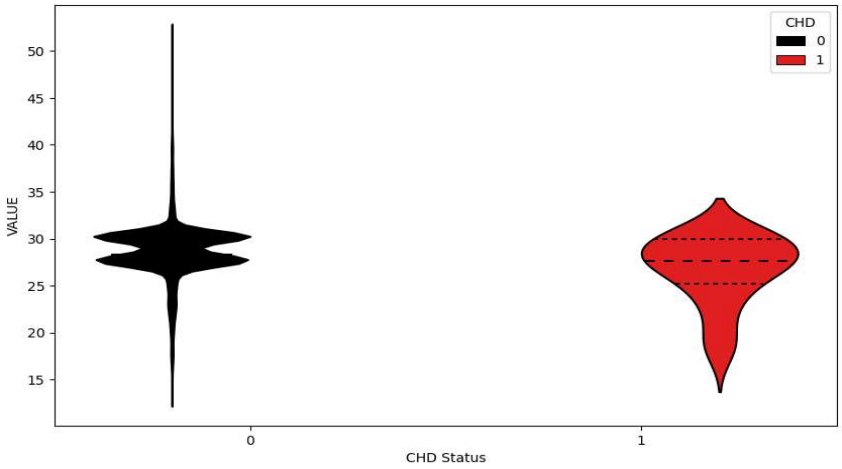

Figure 1

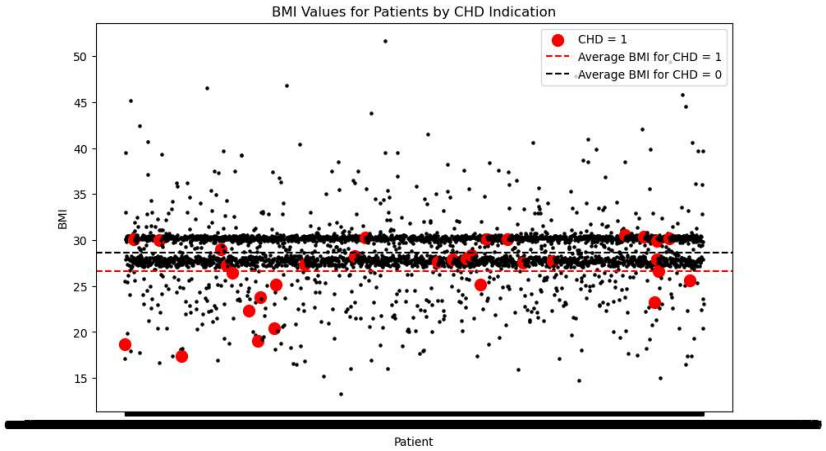

Figure 2

To address this bias, we refined our analysis by incorporating a binary column to indicate CHD status (1 for CHD, 0 for no CHD) and selecting only the highest recorded BMI value per patient (See Figure 1). This adjustment aimed to mitigate the disproportionate influence of patients with numerous encounters on the dataset. Following this refinement, a meaningful difference emerged, with patients with CHD exhibiting a slightly lower average BMI compared to those without CHD. However, the Pearson correlation coefficient for this apparent inverse relationship was -0.066, signifying a very weak or negligible correlation (See Figure 2).

Our investigation into the genetic predispositions associated with cardiometabolic conditions and CHD yielded several noteworthy findings using Chi-Square tests and logistic regression models.

We focused on variations in two specific genes: PON1 and LDLR, located on Chromosome 7 and 19, respectively.

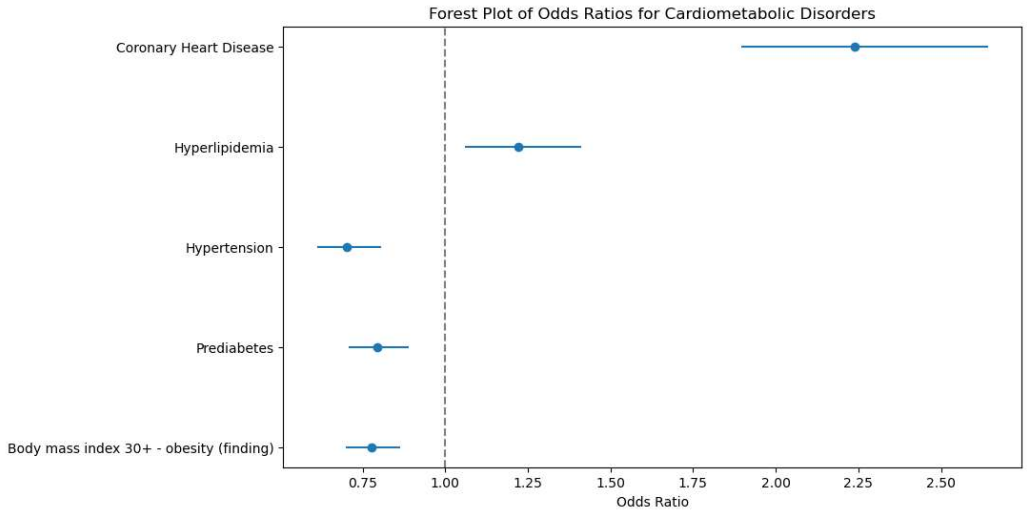

Figure 3

- **Chi-Square Test Results:** The presence of variations in LDLR and PON1 was significantly associated with all the cardiometabolic conditions investigated (p-value < 0.05 for all conditions). This suggests that these genetic variations are indeed linked to these health outcomes.
- **Logistic Regression and Odds Ratios:** Patients with variations in LDLR and PON1 were approximately 22% less likely to have obesity (odds ratio [OR] = 0.78, 95% confidence interval [CI]: 0.70 - 0.86), and the presence of these variations was associated with a 21% reduction in the odds of having prediabetes (OR = 0.79, 95% CI: 0.71 - 0.89). Similarly, these individuals had 30% lower odds of hypertension (OR = 0.70, 95% CI: 0.61 - 0.81). Conversely, the presence of variations was linked to a 22% increase in the odds of hyperlipidemia (OR = 1.22, 95% CI: 1.06 - 1.41). Most strikingly, individuals with variations in LDLR and PON1 were more than twice as likely to have coronary heart disease (CHD) (OR = 2.24, 95% CI: 1.89 - 2.64). P-values for these associations were all significant (less than 0.05) (See Figure 3).
- **Model Performance:** The pseudo-R-squared values for our models ranged from 0.001 to 0.015, indicating that while statistically significant, the genetic variations and demographic factors included in the models only account for a small portion of the variance in these conditions and they do not fully explain the complex etiology of cardiometabolic diseases.
- **Discriminatory Power:** The discriminatory power of the models was assessed using AUC scores. Obesity's AUC was 0.53, indicating minimal discriminatory power. CHD's AUC was 0.60, indicating a better-than-random performance but still relatively weak (See Figure 4).

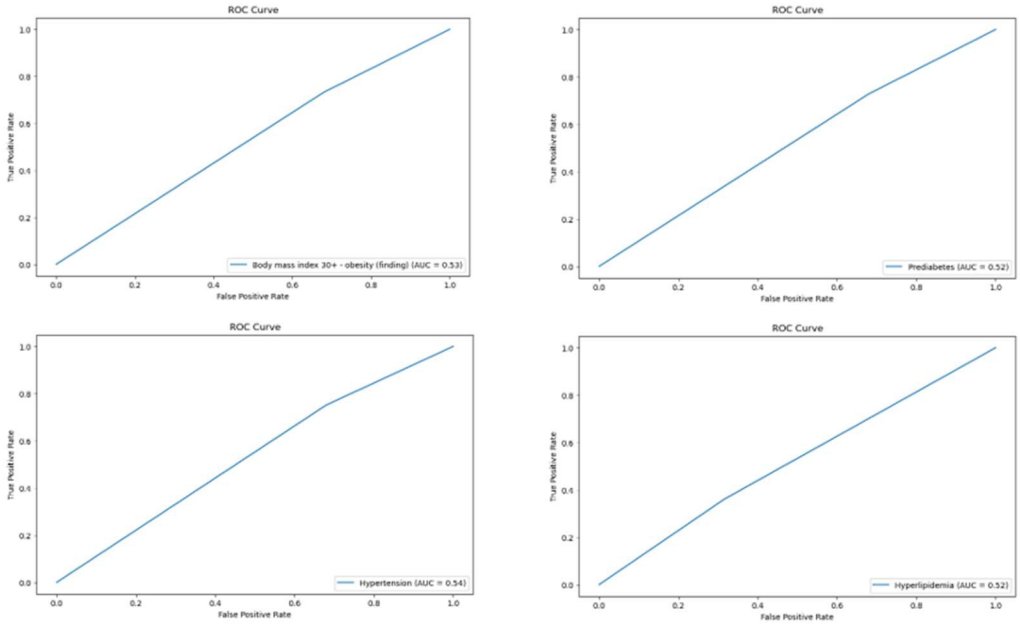

Figure 4

Discussion

Our study sought to clarify the relationship between BMI and CHD, revealing a weak or non-existent association when adjusting for potential biases. This suggests that BMI alone may not be a reliable predictor of CHD risk within this patient population, highlighting the need to consider a broader range of risk factors in clinical practice.

Interestingly, despite the differences, the average BMI for both groups remained in the overweight category by CDC standards, suggesting the complexity of using high BMI as a sole predictor of CHD risk. This finding emphasizes the necessity of considering other contributing factors beyond BMI.

Our analysis of the genetic predispositions associated with cardiometabolic conditions including CHD revealed significant associations between variations in the LDLR and PON1 genes and the presence of these conditions. The presence of these genetic variations was linked to a decreased likelihood of obesity, prediabetes, and hypertension. Conversely, these variations were associated with an increased likelihood of hyperlipidemia and CHD. The particularly strong association observed with CHD (odds ratio [OR] = 2.24, 95% confidence interval [CI]: 1.89 - 2.64), suggests that individuals carrying these variations may benefit from early screening and targeted preventive interventions.

The logistic regression models confirmed the statistical significance of these associations, indicating that the observed relationships are unlikely to be due to random chance. However, the low pseudo-R-squared values (ranging from 0.001 to 0.015) highlight that genetic variations alone do not fully explain the variability in these conditions. This suggests the involvement of other factors, such as environmental and lifestyle influences, in the complex etiology of cardiometabolic diseases. Furthermore, the discriminatory power of the models, as assessed by

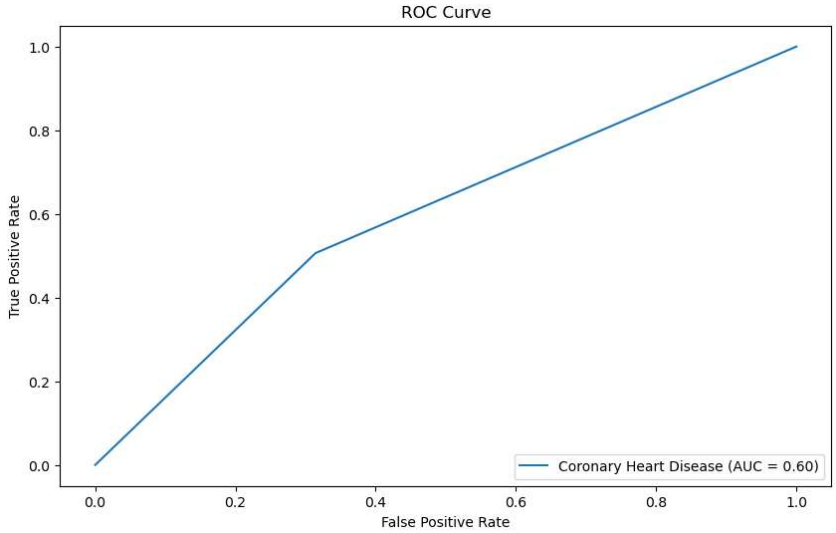

Figure 5

the area under the curve (AUC) metric, was relatively weak. The AUC for obesity was 0.53, suggesting minimal discriminatory power, while the AUC for CHD was 0.60, indicating a better-than-random performance but still with limited predictive capability. These findings emphasize the multifactorial nature of these conditions and the need for a more comprehensive approach to risk assessment and management (See Figure 5).

Limitations

This study has limitations that must be acknowledged. The biggest is that the data utilized in this analysis is synthetic, which means it potentially will not apply to real-world scenarios. Consequently, the findings cannot be confidently generalized to real-world populations. The artificial nature of the data also introduces challenges in evaluating the true impact of the lack of diversity within our patient demographics. This limitation hampers our ability to accurately interpret the results and diminishes the external validity of the study. Future research should aim to incorporate real-world, diverse datasets to provide more robust and generalizable insights into the associations between cardiometabolic factors, genetic predispositions, and CHD.

Another limitation of this study is the age demographics. Firstly, this data's age distribution is skewed towards older people. As shown in tables 1, 26% of the participants of this study are in the 91-100-year age range, with over 50% of patients being over 70 years old. This can be interpreted in multiple ways. One interpretation of this demographic is that they are significantly healthier compared to a typical patient population, as evidenced by a large portion of patients living past the average lifespan in the USA. This is significant because if our patient population is healthier than the average person, the factors affecting them may not be representative of the factors affecting the average person. Another interpretation is that this population may be developing these conditions due to old age. Additionally, the skewed age distribution may limit our ability to generalize the study's findings to younger populations. It is important to consider that older populations may have different lifestyle factors, comorbidities, and genetic predispositions that could affect the study's outcomes.

Finally, another limitation is the race demographics. As stated earlier, the racial composition of the dataset included White (84.2%), African American (8.9%), Asian (6.6%), Native Americans (0.25%), and other races (0.08%). While this is not an uncommon demographic, we may be limiting ourselves by having such a large portion of White individuals. This potentially means that our findings may not be consistent in all race demographics. Ideally, if we had more data for races that are not White, we could find separate results for each race demographic. This would allow for a more comprehensive understanding of how different racial groups are impacted by PONS1 and LDLR. Therefore, future studies should aim to include a more diverse patient population to ensure the findings are more generalizable and applicable for all racial demographics.

Future Direction

Our study reveals the complex relationship between cardiometabolic factors, genetic predisposition, and the risk of CHD. The findings highlight that while BMI may not be a strong independent predictor of CHD, genetic variations in LDLR and PON1 genes play a significant, albeit limited, role in influencing the risk of CHD and other cardiometabolic conditions. The potential for personalized medicine approaches in managing cardiometabolic health is evident, emphasizing the need to integrate genetic data into risk assessment and treatment strategies. However, the limitations of relying solely on genetic information are also apparent, underscoring the multifactorial nature of these conditions and the need to consider a broader range of factors, including environmental and lifestyle influences.

Future research should prioritize a more comprehensive and individualized approach, incorporating additional genetic, environmental, and lifestyle factors to enhance our understanding and prediction of CHD and related cardiometabolic conditions. The insights gained from this research pave the way for the development of more effective healthcare strategies and personalized treatment plans, ultimately contributing to the fight against CHD-related mortality and emphasizing the critical need for continued funding in this vital area of research.

References

1. Coronary Artery Disease (CAD): Symptoms & Treatment (clevelandclinic.org)
2. Brown JC, Gerhardt TE, Kwon E. Risk Factors for Coronary Artery Disease. [Updated 2023 Jan 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2024 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554410/