Identification of causal risk factors for intracranial hemorrhage using Mendelian Randomization analysis

Sahil Adhawade
September 30, 2020
Word Count: 5,405

***Abstract*—Background: Over 88% of stroke-related deaths are caused by intracranial hemorrhage (ICH). Currently, there is a myriad of observational analyses and evidence that investigates the association between various exposures or risk factors and the outcome, ICH. However, a severe underlying issue with correlational studies is the risk of confounding variables. Confounders lead to invalid results and are often difficult to measure due to the sheer number of them.**

Methods: I implemented a Mendelian Randomization analysis in order to uncover *causal* relationships between ICH and various risk factors by eliminating the worry of confounding variables. Specifically, I utilized the IVW, MR-Egger, weighted median and mode, and simple mode statistical models to find causal relationships. For risk factors displaying high pleiotropy amongst its variants, I implemented a contamination mixture method.

Results: Strong causal relations were observed between hypertension and ICH and body-mass index (BMI) and ICH. Additionally, there was no significant causal association between high-density lipoprotein cholesterol (HDL-c) and ICH. Finally, the results displayed a significant, inverse causal association between low-density lipoprotein cholesterol (LDL-c) and ICH, via the contamination mixture method, implying that low levels of LDL-c caused a greater risk of ICH.

Conclusion: In essence, through the Mendelian Randomization study, not only did I find causal relationships between various risk factors and ICH leading to better detection and treatments, but it also aided in finding evidence to how clinical treatments, such as the prescription of statins, may need to be revised.

***Index Terms*—biostatistics, causal inference, intracranial hemorrhage, Mendelian Randomization**

I. Introduction

Humans subsist on the passage of blood through certain arteries and vessels to deliver vital oxygen and nutrients to different organs and muscles. However, the structural integrity of the walls of arteries may not always be able to resist the constant hydraulic stress due to the pumping of blood. In some cases, this could lead to a ballooning of weaker segments of an artery known as an aneurysm [1]. If these aneurysms exceed a mean diameter of around 7 mm, they are likely to hemorrhage, colloquially known as a rupture, which could potentially lead to hemorrhagic stroke, permanent brain damage, or death [2]. Further, intracranial hemorrhage (ICH) frequently occurs in the caudate head, associated with storing and processing memories, the putamen/globus pallidus, which regulates movement and influences various types of learning, and the thalamus, known for its sensory relay for a variety of physiological systems [3]. Therefore, a hemorrhage in these regions could diminish quality of life or possibly lead to death. Annually, over 88% of stroke-related deaths are caused by ICH with a 30-day mortality rate of around 45% [4]. Currently, however, the identification of risk factors related to ICH occurs through observational evidence, which is often questionable due to their potential confounding effect. Consider this simplified example: if there exists a strong negative correlation between vitamin D levels, the risk factor, and the risk of cancer, the outcome, there is still a possibility that a patient exists with high vitamin D levels and a high risk of cancer. As per the associational evidence, this patient with high vitamin D levels should not receive treatment even though they have a high risk of cancer. The false-negative error, a red-herring caused by a confounding variable other than the risk factor, is fatal because the patient should be receiving treatment. Additionally, an associational study was conducted comparing the effect of high and low-density lipoprotein cholesterol (HDL-c and LDL-c), also known as "good" and "bad" cholesterol respectively, on the risk of hemorrhagic stroke in which they concluded an inverse relationship between LDL-c and the outcome [5]. However, the results of this associational study should be carefully considered due to the existence of potential confounding variables. For instance, in the scenario of LDL-c and ICH, a possible confounding variable could be a patient's diet or the amount of exercise, which may act as an external influence on their LDL-c levels and their risk of ICH. In fact, it is hard to pinpoint exactly which confounder is potentially swaying the results making it very hard to control confounding effects properly. This demonstrates the risk of blindly accepting purely observational evidence because the presence of confounders makes it difficult to definitively determine whether the risk factor or a confounding variable had a direct effect on the outcome. This is especially detrimental when forming the proper treatment for a patient as doctors could possibly prescribe an unnecessary treatment or fail to prescribe a necessary treatment. A more effective method would be to find which factors directly *cause* ICH by statistically removing effects of confounding variables. Therefore, unlike previous observational studies conducted to assess risk factors' effects on ICH, this study utilizes a Mendelian Randomization analysis to identify the putative

causal relationships between ICH and several risk factors: hypertension, body-mass index (BMI), HDL-c, and LDL-c.

II. The Distinction Between Statistical Association and Causality

In order to understand causal inference, I must first draw the distinction between statistical associations and causation. Statistical associations are typified by regressions, hypothesis testing, and probability in which certain parameters, or characteristics of a population, are assessed to infer associations among variables and estimate predictions [6]. Additionally, these tasks, including updating beliefs or probabilities through new data, are only well managed under static experimental conditions [7]. On the contrary, causal inference attempts to estimate predictions and infer beliefs under both static and dynamic experimental conditions, which could be induced through intervention or switching from observational to experimental data [7].

Unequivocally, statistical association and causality are two distinct entities. Consider a joint distribution, or a probability distribution, in which symptoms and a certain disease are two random variables. As per the laws of probability theory, a change in the distribution function's property, such as symptoms, does not necessarily suggest a consequent, definitive change of another property, such as the disease [7]. This highlights the central issue with statistical associations. Their main function is to quantify relationships between different parameters to understand how they vary with one another. However, these associations fail to explain why an intentional change in one variable might affect the outcome variable. This lack of information is especially dangerous in medicine because if a particular risk factor's relationship to the risk of a disease is unknown or skewed by confounding effects, then formulating a treatment could be inaccurate as it would be difficult to pinpoint and suppress the exact risk factor causing the disease [7].

III. Mendelian Randomization

Identification of causal relationships leads to the development of new guidelines of treatment and medical therapeutics [8]. As exemplified in section 2, inferring causal relationships between an exposure and an outcome is imperative to developing proper treatments [7]. With any statistical analysis, there exists the risk of confounding variables or external influences. However, the risk of confounding variables can be mitigated by utilizing a randomized control trial [9]. Assume a randomized controlled trial was conducted to observe the potential causal effect of smoking and the risk of cardiovascular disease (CVD). First, the experimenter would amass a large, random pool of participants and randomly assign half of the population to the first group and the other half to the second group. The first group is forced to smoke while the second group is not allowed to smoke. After waiting a few years, the experimenter then observes which participants have and do not have CVD. This experimental analysis could reveal a causal relationship between smoking and the risk of CVD as confounding influences, such as diet, exercise, or other external factors, are controlled by the experimenters [10]. However, this method is not only lengthy and expensive, making it impractical, but it is also grossly unethical [8].

Mendelian Randomization is an epidemiologic approach to finding causal relationships between exposures and an outcomes while circumventing the issues stemmed from a traditional randomized controlled trials [8], [10]. This approach of analyzing observational data requires an understanding of instrumental variables, which are similar to a random assignment to a treatment group in experimental analysis. The instrumental variables function as natural experiments in which they vary primarily with respect to the risk factor but not with the confounders associated with the exposure-outcome relationship [11]. Furthermore, Mendelian Randomization uses genetic variants as proxy measures for modifiable exposures. These genetic variants, more formally known as single nucleotide polymorphisms (SNPs), are the underlying framework of Mendelian Randomization analysis as they are virtually unaffected by any environmental variables and immune from reverse causation because they are congenital and not subject to change as one ages. However, in order to infer causality between an exposure and an outcome, Mendelian Randomization requires the genetic variants to satisfy certain assumptions necessary for an instrumental variable:

i the SNP must be associated with the risk factor,
ii the SNP is not associated with any confounding variables related to the exposure-outcome association, and
iii the SNP is not associated with the outcome of the exposure-outcome relationship other than through the risk factor [11].

If these assumptions are satisfied, then any association between the exposure and the outcome can be considered as causal. This is because these assumptions allow us to conclude that there is only a single causal pathway from the genetic variant to the outcome, which is through the identified risk factor (Fig. 1).

IV. Etiology of Intracranial Hemorrhage

In order to choose the proper risk factors to test on the outcome, ICH, it is essential to understand the etiology or the originating factors of the disease. ICH mainly occurs due to weak arterial walls caused by the constant hydraulic pressure of the blood passage. One of the primary risk factors that are a culprit for causing immense stress to arteries is high blood pressure, more formally known as hypertension. Hypertension, a condition related to blood pressure, can increase due to old-age, current or past smoking, or high intake of alcohol [12]. In addition to hypertension, I analyzed the causal effect of BMI, on ICH. An increase in BMI, indicating a patient as overweight or obese, leads a greater strain on arteries, thus sacrificing its structural integrity, making it a possible risk factor for the cause of ICH [13]. High-density lipoproteins, or good cholesterol, is associated with the deaccumulation of plaque in arteries, while low-density lipoproteins, or bad cholesterol, has an antithetical effect. These risk factors may

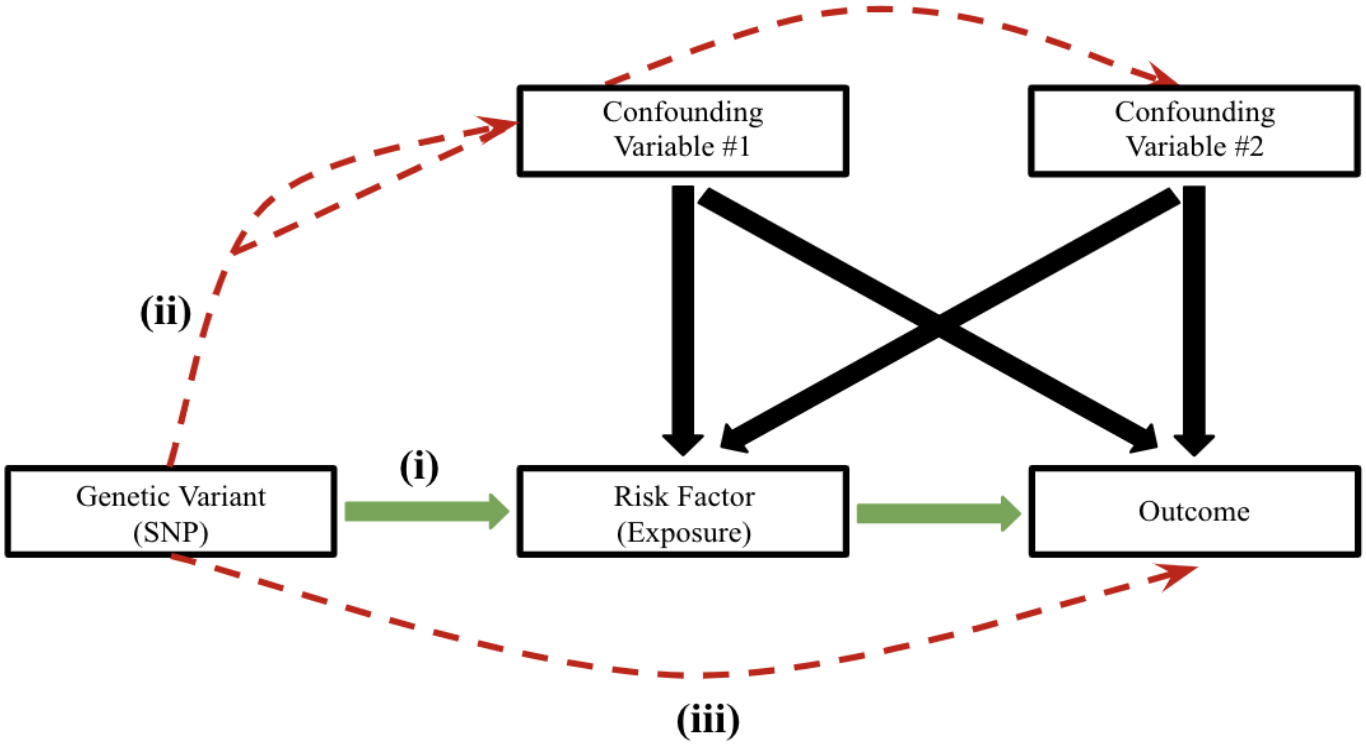

Fig. 1. Assumptions i, ii, and iii are graphically modeled above. The dotted, red arrows represent no association, and the solid, green arrows represent an association. As shown above, if there is no association between the SNP and the confounding variables and the outcome (assumptions ii & iii), then a single causal pathway remains represented by the solid, green arrows.

have a relationship to ICH because the accumulation of plaque in arteries can also lead to obstructed bloodstreams causing the rupture of a hemorrhage [13].

V. Methods

A. The Genome-Wide Association Study Data

For the purposes of this Mendelian Randomization analysis, I used a genome-wide association study (GWAS) data, which is an approach employed in genetics research to associate SNPs with different outcomes or exposures. More specifically, the GWAS data was acquired from the UK Biobank, which is a large repository of genetic data with over 90 million quality controlled and imputed SNPs, which was publicly released in 2017 [14]. The outcome, ICH, data utilized in the Mendelian Randomization analysis consists of SNPs from both sexes and over 8 million SNPs [14]. The exposures—hypertension, BMI, HDL-c, and LDL-c—tested contained genetic variants from over 500,000 samples only from European ancestry between ages 40-69 at the time of recruitment for the trial [15].

B. Languages and Packages

In order to conduct the Mendelian Randomization analysis, I decided to use the R programming software as it contains the necessary packages to perform statistical genetic analysis. Specifically, I utilized the `TwoSampleMR` library that consists of the necessary methods needed to carry out the Mendelian Randomization analysis. Additionally, this library consists of a large GWAS database of `available_outcomes()`, which is a compilation of different summary statistics of outcomes and exposures labeled with a specific id (exposure data collected from Neale Lab are labeled with a `'ukb-b-'`). Although I was able to find all the exposure data from this database, the outcome, ICH, data was not found. Using ICH data from an external UK Biobank source exhibited several issues. For instance, in order to modify this data set to match the correct formatting necessary to be compatible with the exposure data sets and the methods, I had to add, as well as rename several columns. Additionally, as it was missing an SNP column with the correct "rs" IDs (which are used to access specific SNPs), I had to manually merge the ICH data frame with the exposure data frame to ensure that the SNPs in that particular exposure data set match with the coinciding SNPs in the outcome data set. Through this, I ensured that I was only assessing the specific subset of the outcome data with the coinciding SNPs in the exposure data to infer a potential causal association between that particular exposure and ICH. After this step was completed, I used methods, such as `extract_instruments()` (extract data using the exposure id), `clump_data()` (prune data to contain only SNPs with low p-values or high significance), `harmonise_data()` (ensures that effect of SNP on the exposure and effect of SNP on the outcome correlate to the same allele), and `mr()` (conduct the MR analysis using

harmonized data), in order to gather my preliminary results of the causal association between each exposure and the outcome. Finally, I manually coded the contamination mixture analysis for LDL-c, which is a robust method of analysis that generates significant causal estimates when invalid genetic variants are present.

C. Types of Mendelian Randomization Methods

In this study to unveil causal relations between ICH and various risk factors through a Mendelian Randomization approach, I applied several different methods, each with slightly different semantics, in order to ensure consistent and significant results. Specifically, I employed the Inverse Variance Weighted (IVW), Simple Mode, Weighted Median, and MR-Egger methods.

1) **Inverse Variance Weighted (IVW)** [16], [17]: In conjunction with the assumptions outlined in section 3, IVW also typically assumes that the instrumental variables and the outcome are linear allowing for the calculated causal estimate to also be linear. Assume a genetic variant *k*, where the association of *k* with the exposure is $\hat{\beta}_{X_k}$ and standard error $\sigma_{\hat{\beta}_{X_k}}$, and the association of *k* with the outcome is $\hat{\beta}_{Y_k}$ and standard error $\sigma_{\hat{\beta}_{Y_k}}$. I can estimate both the causal effect ($\hat{\theta}_k$) and standard error ($\sigma_{\hat{\theta}_k}$) using:

$$\hat{\theta}_k = \frac{\hat{\beta}_{Y_k}}{\hat{\beta}_{X_k}} \tag{1}$$

$$\sigma_{\hat{\theta}_k} = \frac{\sigma_{\hat{\beta}_{Y_k}}}{\hat{\beta}_{X_k}} \tag{2}$$

These two equations, however, are only applicable to find the causal estimate, along with the respective standard error, of a single genetic variant *k*. However, if I want to estimate the causal effect of multiple instrumental variables such that $k \in$ K where all the instrumental variables (K = $\{k_1, k_2, k_3, k_4, k_5, ..., k_n\}$) are assumed to be valid, I could perform a two-stage least square regression. Through this approach, I first linearly regress the instrumental variables, or genetic variants, with the risk factors and, then, regress the risk factors on the outcome. The same effect can be modeled by calculating the weighted mean of single instrumental variable estimates and standard errors (1), (2). This is known as the inverse variance weighted, or IVW, estimate:

$$\hat{\theta}_{IVW} = \frac{\sum_k \hat{\beta}_{X_k}\hat{\beta}_{Y_k}\sigma^{-2}_{\hat{\beta}_{Y_k}}}{\sum_k \hat{\beta}^2_{X_k}\sigma^{-2}_{\hat{\beta}_{Y_k}}} \tag{3}$$

The inverse-variance weighted method only proves to be a consistent causal estimator if all the candidate SNPs are considered to be valid instrumental variables, assuming there to be zero pleiotropy [1] amongst the genetic variants and, therefore, a zero y-intercept on the scatter plots.

[1]Pleiotropy is when a single genetic variant seems to affects two or more unrelated phenotypic traits. Pleiotropic variants are often a violation of assumption (i) defined in section 3.

2) **Simple Mode** [18]: The simple mode method in Mendelian Randomization analysis generates clusters of SNPs that have a similar causal estimate. These variant-specific causal estimates are determined by the ratio of the SNP-outcome association to the SNP-risk factor association (1). The simple mode method forms clusters using the kernel density function, which is a non-parametric method that models both linear and non-linear associations without needing to specify a rigid function form. Furthermore, this simple mode method returns an unbiased causal estimate of the exposure-outcome relationship if the SNPs within the largest cluster of SNPs are all valid instruments. This is also known as the Zero Modal Pleiotropy Assumption (ZEMPA), which essentially means that the SNPs only exhibit influence on a single phenotypic trait, which is, in this case, the exposure. If this assumption is not satisfied, that genetic variant will be an invalid instrumental variable.
3) **Weighted Median** [18]: Another approach to the Mendelian Randomization analysis is the median-based method, specifically the weighted median. The primary advantage of employing the weighted median method is that only half of the total number of SNPs need to exhibit valid instrumental variables, satisfying all three assumptions described in section 3, in order to produce an unbiased causal estimate. In addition, the weighted median method also requires that 50% of the calculated weights must originate from valid instrumental variables. In terms of calculations, the variant-specific causal effects, denoted by $\hat{\theta}_k$ such that $\hat{\theta}_1 < \hat{\theta}_2 < \hat{\theta}_3 < ... < \hat{\theta}_k$, are first sequentially ordered. Second, I found the inverse-variance weights for each of the causal estimates:

$$w_k = \frac{\left(\sigma_{\hat{\theta}_k}\right)^{-2}}{\sum_i \left(\sigma_{\hat{\theta}_i}\right)^{-2}} = \frac{\frac{\hat{\beta}^2_{X_k}}{\sigma^2_{Y_k}}}{\sum_i \frac{\hat{\beta}^2_{X_i}}{\sigma^2_{Y_i}}} \tag{4}$$

Next, the standardized sum of all the weights is calculated where s_j is less than 0.50, representing weights from invalid instrumental variables, and s_{j+1} is greater than 0.50, representing weights from valid instrumental variables:

$$s_j = \sum_{k=1}^{j} w_k = w_1 + w_2 + w_3 + ... + w_k < 0.50 \tag{5}$$

and

$$s_{j+1} = \sum_{k=1}^{j} w_{k+1} = w_1 + w_2 + w_3 + ... + w_k + w_{k+1} > 0.50 \tag{6}$$

Finally, using the range of values determined by s_j and s_{j+1}, a linear extrapolation is performed to determine the weighted median causal estimate:

$$\hat{\theta}_{WM} = \hat{\theta}_k + \left[\left(\hat{\theta}_{k+1} - \hat{\theta}_k\right)\left(\frac{0.50 - s_j}{s_{j+1} - s_j}\right)\right] \tag{7}$$

Through using the inverse-variance weights, the weighted median method allows for stronger SNPs to exude a greater effect on the weighted median causal estimate, $\hat{\theta}_{WM}$.

4) **MR-Egger** [19]: An alternative method is to use Mendelian Randomization Egger regression, commonly known as MR-Egger. This method is essentially a weighted linear regression of the SNP association with the exposure and outcome, $\hat{\beta}_{X_k}$ and $\hat{\beta}_{Y_k}$ respectively. More importantly, the y-intercept is allowed to be at a value other than zero, therefore, displaying directional (unbalanced) pleiotropy, unlike the IVW method. In order to calculate the MR-Egger causal estimate, a weighted regression model is used:

$$\begin{aligned} \hat{\beta}_{Y_k} =& \hat{\theta}_{0_{MRE}} + \hat{\theta}_{MRE}\hat{\beta}_{X_k} + \epsilon_{MRE_k}; \\ & \epsilon_{MRE_k} \sim N\left(0, \phi^2 \sigma^2_{\hat{\beta}_{Y_k}}\right) \end{aligned} \tag{8}$$

In this equation, I added a residual term, ϵ_{MRE_k} , and an intercept term $\hat{\theta}_{0_{MRE}}$ in order to represent the y-intercept or the directional pleiotropy that is accounted for by the MR-Egger estimation, $\hat{\theta}_{MRE}$. Additionally, the validity of the MR-Egger estimate is determined by the Instrument Strength Independent of Direct Effect (InSIDE) assumption, which, simply put, means that the pleiotropic effect of the SNPs is independent of the strength of the instrumental variables. However, this assumption is weaker than simply assuming no directional pleiotropy, as in the IVW method. Therefore, the MR-Egger method is less statistically powerful than other weighted-estimate approaches, particularly the IVW method.

D. Visual and Sensitivity Analysis of Mendelian Randomization Results

Along with numerically calculating the causal estimate for each exposure-outcome pair, several plots were used to visualize the causal estimate and test my result's sensitivity. Specifically, scatter and funnel plots were utilized to visually represent each causal estimate and understand their degree of pleiotropy. A leave-one-out forest plot was implemented to test the sensitivity of my Mendelian Randomization results.

1) **Scatter Plot** [20]: The scatter plot graphs $\hat{\beta}_X$ against $\hat{\beta}_Y$, which represent SNP associations with the exposure and outcome respectively. The lines represent each different Mendelian Randomization method and its calculated causal estimate.
2) **Funnel Plot** [20]: The funnel plot graphs the variant-specific Wald ratios, $\hat{\beta}_{IV} = \frac{\hat{\beta}_X}{\hat{\beta}_Y}$, against their precision, $\frac{1}{\sigma_{IV}}$. This plot is used to represent pleiotropy in the Mendelian Randomization results through assessing the approximate symmetry of the plot. A symmetric plot means there is an approximately equal number of pleiotropic variants that both increase and decrease bias, therefore, canceling out and making the net effect of pleiotropic variants negligible. On the other hand, an asymmetric plot means that either pleiotropic variants that increase or decrease bias are dominant, leading the overall net pleiotropy in the results.
3) **Leave-one-out Forest Plot** [20]: In the leave-one-out forest plots, the black points on the forest plot represent the causal estimate of the risk factor on ICH by using each of its SNPs separately. The red points show the combined causal effect of hypertension on ICH using the MR-Egger and IVW methods. The lines spanning out from each point represent the 95% confidence interval. More specifically, the black points represent the maximum likelihood Mendelian Randomization method applied to generate the causal estimate between the risk factor and ICH by excluding that particular SNP. This plot is useful in visualizing the sensitivity of my Mendelian Randomization analysis. A fairly uniform curve means that the results are not sensitive to any one specific genetic variant, therefore, indicating the lack of outlier SNPs.

VI. Results

For the Mendelian Randomization analysis, the methods outlined in section 5c were used, and the statistical significance of each causal relation and the causal estimates were assessed through a calculated p-value (*p*) and beta-value (β) statistics, respectively. For the beta-value, if $\beta > 0$, then it indicated that the particular exposure increases the risk of the outcome (direct variation), and, antithetically, if $\beta < 0$, then the particular exposure would reduce the risk of the outcome (inverse variation). Furthermore, a multiple comparisons problem arose, which typically occurs when performing multiple independent statistical tests on a single set of data, which, in this case, is the outcome data set [21]. The study used four different risk factors ($n = 4$), therefore, making the family-wise error rate, $\alpha_{FWER} = 1 - (1 - \alpha)^n$ = 18.55%, where $\alpha = 0.05$. To bring this increased error rate down to ≈ 0.05, a Bonferroni correction was used, in which $\alpha_c = \frac{\alpha}{n}$ = 0.0125. Using this corrected error rate, I noticed that the family wise error rate, $\alpha_{FWER} = 1 - (1 - 0.0125)^4 \approx 0.05$. Hence, the statistical significance of all results is determined at $\alpha = 0.0125$.

A. The Causal Effect of Elevated Hypertension on the Increased Risk of ICH

In order to test the causal effect of hypertension on the risk of ICH, I utilized GWAS summary statistics for hypertension from the MR database. The extracted dataset of genetic variants associated with hypertension consisted of 68 SNPs once clumped and harmonized. Through the Mendelian Randomization analysis (Table 1), I concluded that hypertension has a significant causal estimate via the IVW and Weighted Median method ($p < 0.0125$). To further analyze these results, I examined visual representations of the data. In the scatter plot (Fig. 2a), the five slopes fitted to the data points, or SNPs, represent the causal estimate of each of the five Mendelian Randomization methods. According to Figure 2a, the estimates had nearly the same slope, and, there was little to no deviation in the y-intercept for the MR-Egger method

TABLE I
SUMMARY STATISTICS FOR EACH RISK FACTOR USING MENDELIAN RANDOMIZATION ANALYSIS.

	Hypertension			BMI			HDL-c			LDL-c		
Methods	β	σ	p	β	σ	p	β	σ	p	β	σ	p
MR-Egger	-4.68e-4	0.009	0.959	-9.91e-5	7.70e-4	0.897	-7.81e-4	4.39e-4	0.079	-2.20e-4	3.65e-4	0.549
Weighted Median	0.009	0.003	0.007	7.00e-4	4.98e-4	0.159	-7.99e-4	3.93e-4	0.042	-4.69e-4	3.47e-4	0.177
IVW	0.007	0.002	0.002	7.63e-4	2.84e-4	0.007	-5.71e-4	2.65e-4	0.031	-5.38e-4	2.45e-4	0.028
Simple Mode	0.015	0.008	0.083	0.002	0.001	0.179	1.38e-4	8.15e-4	0.866	-0.001	6.33e-4	0.081
Weighted Mode	0.012	0.007	0.109	6.88e-4	8.34e-4	0.409	-5.87e-4	3.62e-4	0.108	-5.64e-4	3.51e-4	0.117

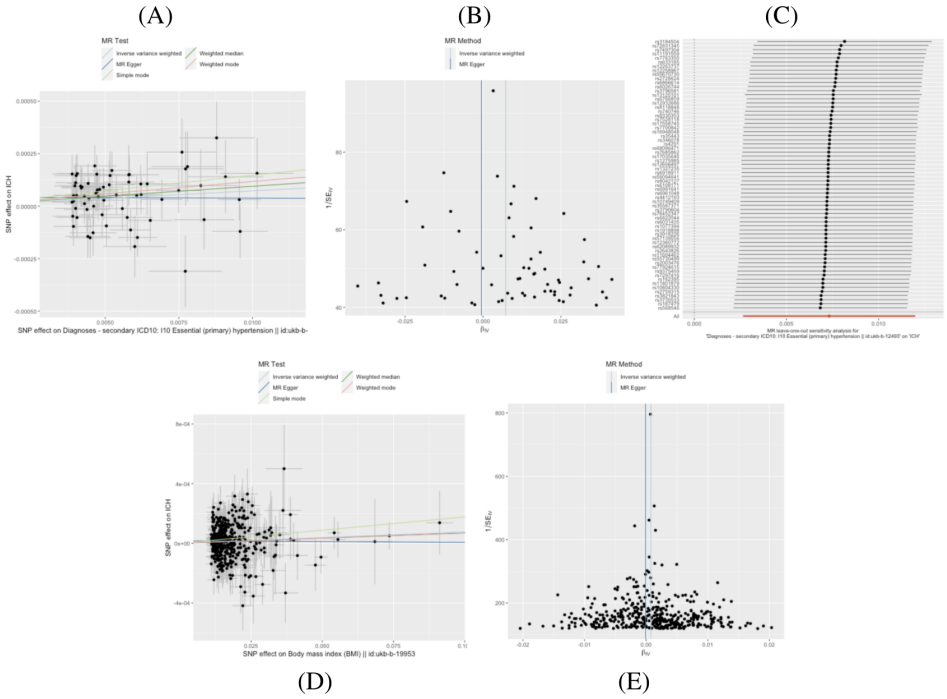

Fig. 2. Hypertension-ICH displayed in scatter (a), funnel (b), and leave-one-out forest plot (c). BMI-ICH displayed in scatter (d) and funnel plot (e).

indicating limited pleiotropy amongst the genetic variants. To confirm this, the funnel plot (Fig. 2b) also displays limited pleiotropy due to a fairly symmetric distribution of data points. In addition to the scatter and funnel plot, I generated a leave-one-out forest plot to examine the sensitivity of my results. As shown in Figure 2c, uniformity in the leave-one-out forest plot suggests that my Mendelian Randomization results are not sensitive to any one SNP, indicating the lack of any outlier genetic variants. Thus, through these robust results, I can infer a statistically significant causal effect of hypertension on ICH.

B. The Causal Effect of Heightened BMI on the Increased Risk of ICH

The GWAS summary statistics for BMI from the MR database were used in order to estimate the causal effect of BMI on ICH. After clumping and harmonizing the extracted instruments, I evaluated a causal estimate using 434 SNPs acquired from the BMI data set. According to Table 1, I noticed a significant causal estimate between BMI and ICH via the IVW method ($p < 0.0125$). In order to confirm this statistical significance, I analyzed several visual plots to view the reliability and sensitivity of my results. Through analyzing the scatter plot (Fig. 2d), I noticed that four Mendelian Randomization methods have a zero y-intercept indicating no horizontal pleiotropy, including the MR-Egger method which displays a near zero y-intercept indicating limited to no pleiotropy present. I further confirmed this low degree of pleiotropy in my results by constructing a funnel plot, which shows symmetric distribution of points indicating almost no pleiotropy (Fig. 2e). Finally, I constructed a leave-one-out forest plot to test the sensitivity of my causal estimate between BMI and ICH. I observed a fairly uniform curve that suggests that the exposure-outcome causal relationship is not sensitive to any one particular SNP indicating no outliers. Therefore, it is reasonable to suggest that there exists a reliable putative causal relationship between BMI and ICH.

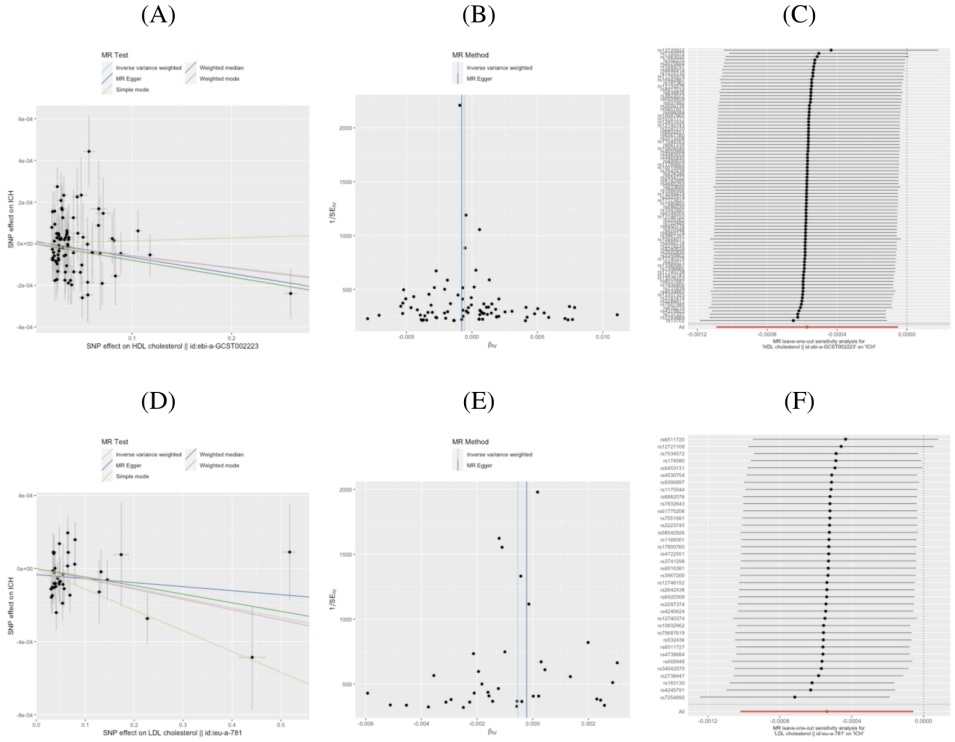

Fig. 3. HDL-c-ICH displayed in scatter (a), funnel (b), and leave-one-out forest plot (c). LDL-c-ICH displayed in scatter (d), funnel (e), and leave-one-out forest plot (f).

C. HDL and LDL Cholesterol Risk Factors

According to Table 1, there was little evidence for a significant causal relationship between HDL and LDL cholesterol and the risk of ICH ($p > 0.0125$) using the traditional Mendelian Randomization methods. To further investigate, I observed the scatter, funnel, and leave-one-out forest plots for both LDL-c and HDL-c.

The causal estimate generated by each method for HDL-c and ICH seems to have an inverse relationship indicated by the negative slopes (Fig. 3a). More importantly, in Table 1, none of the methods produced a significant p-value, implying that HDL-c is likely not a strong causal risk factor for ICH. Additionally, through an analysis of the scatter plot for HDL-c and ICH (Fig. 3a), I found very little pleiotropy in the results, indicated by the small magnitude of the y-intercept for the MR-Egger estimate. I confirmed the small degree of pleiotropy in my results by analyzing the funnel plot for HDL-c and ICH (Fig. 3b), which displayed a fairly symmetrical distribution of points. Finally, ensure accuracy of my results, I tested the sensitivity of my results with a leave-one-out forest plot for HDL-c and ICH (Fig. 3c). The uniform curve indicates that the calculated causal estimates were not swayed by certain SNPs. Due to a lack of evidence of a significant number of invalid genetic variants or inaccuracies, a contamination mixture analysis was not conducted.

For LDL-c and ICH, the scatter plot (Fig. 3d) represents a negative sloping causal estimate across all traditional Mendelian Randomization methods, indicating a key inverse relationship. In addition, I noticed the fairly large magnitude of the y-intercept for the MR-Egger estimate indicating the presence of pleiotropic variants. To confirm this, through an analysis of the funnel plot (Fig. 3e), I noticed that it displays asymmetry, which further affirms the existence of pleiotropy.

In order to mitigate the effects of the invalid, pleiotropic genetic variants, I employed the contamination mixture analysis. Through this approach, I found a significant relationship between a subset of SNPs associated with LDL-c and ICH. For the 37 SNPs associated with LDL cholesterol, my results consisted of a $\beta = -0.001$, $\sigma = 2.6 \times 10^{-4}$, and calculated z-score of z = 3.85. Using the z-score statistic, I calculated the p-value statistic, $p = 1.2 \times 10^{-4}$. This low p-value statistic ($p << 0.0125$) suggests a significant causal estimate between LDL-c and ICH once the invalid genetic variants are filtered out. Additionally, the leave-one-out forest plot (Fig. 3f) did not signify that my results were sensitive to any one SNP in particular, which implies a reliable causal estimate.

VII. Discussion

A. Understanding Hypertension as a Causative Risk Factor

Hypertension, also known as high blood pressure, performed exceptionally well across all of the methods applied, which included the five traditional Mendelian Randomization methods. This was expected because, intuitively, arterial stress caused by the constant hydraulic pressure of the passage of

blood can lead to a weakening of the structural integrity of arterial walls resulting in an aneurysm. Therefore, hypertension is an unequivocal risk factor that can have a causative relationship with ICH, which was proved by the Mendelian Randomization analysis. Furthermore, current literature suggests that the average age range of morbidity from ICH is between 60-80 years old, and the risk of developing an intracranial aneurysm increases with age [12]. Similarly, hypertension also has the tendency to increase with age, thus, further proving that it is a likely factor that causes ICH in mostly older patients [22].

B. Understanding BMI as a Causative Risk Factor

BMI is the ratio derived from the mass (kg) divided by height squared (m^2). The BMI value can indicate the weight of a person in relation to their height. This measure can indicate whether a person is underweight, overweight or obese, or at the proper weight for their height. As discovered through the Mendelian Randomization analysis, I found that BMI had a significant causal estimate with ICH and a positive causal effect on ICH. Therefore, this implies that an increase in BMI increases the risk of ICH. Once again, intuitively, if a patient is overweight or obese, their blood pressure naturally increases, causing immense strain on the arteries. As a result, the arterial walls weaken, making them more susceptible to balloon into an aneurysm. Assessing the demographics of BMI and ICH, on average, 700,000 non-Hispanic White Americans are expected to experience some form of hemorrhagic stroke [23]. In conjunction with this evidence, non-Hispanic Whites also make up the largest population of overweight (BMI of 25.0 - 29.9) and obese (BMI of $\geq$ 30.0) people [23]. My Mendelian Randomization analysis reveals an evident causal relationship between BMI and ICH, confirmed by a similar trend seen through the demographical data.

C. HDL-c, LDL-c, and the Issue with Statins

HDL-c, also known as "good cholesterol", gets rid of excess cholesterol in the bloodstream to prevent the build-up of plaque in arterial walls. On the other hand, LDL-c, or "bad cholesterol", is known to have the opposite effect as it leads to the build-up of plaque in arterial walls, primarily resulting in cardiovascular issues. For the HDL-c versus ICH, the inverse relationship likely has little significance as the analysis showed no statistically significant causal association. Nevertheless, I found a significant causal estimate between a subset of valid LDL-c instrumental variables and ICH via the contamination mixture method. However, I also noticed the same negative or inverse causal relationship between LDL-c and ICH, which means higher levels of "bad cholesterol" actually reduces the risk of ICH. This was a rather unexpected result and required further investigation. LDL-c is a culprit for the vast majority of cardiovascular diseases. In order to prevent an increase in the amount of LDL-c in a patient's bloodstream, physicians usually prescribe statins in order to reduce LDL-c. More specifically, statins are lipid-lowering medications, particularly LDL-c, in order to reduce the risk of cardiovascular disease [24]. British experts were recommending statins to anyone over the age of 50 and even to healthy people. Without understanding the side-effects of statins, prescribing them loosely to everyone over the age of 50 is grossly negligent [24]. However, through my Mendelian Randomization analysis, I gained a better understanding of not only the causative relationship between LDL-c and ICH but also a detrimental side-effect of statins. Statins are meant to lower LDL-c levels; however, this, in turn, causes a greater risk of ICH, as mentioned earlier. While statins do tend to lower the risk of cardiovascular disease, they also have an adverse side-effect of increasing the risk of hemorrhagic stroke, specifically through ICH. Other previous studies have shown some evidence that the use of statins to lower LDL-c levels actually heightens the risk of hemorrhagic stroke [5]. However, these studies have been completely based on observational evidence, which lacks definitive reliability due to the presence of confounding variables that are difficult to account for in a simple observational analysis. My Mendelian Randomization analysis suggests that the evident inverse variation between LDL-c and ICH, such that, with decreases in LDL-c, the risk of ICH increases, not only corroborates known observational evidence but also further suggests a causal relationship between reduced LDL-c and increased ICH risk. These results conclusively highlight an unintentional effect of statins, which are known to decrease LDL-c concentration, in turn, causing a higher risk of ICH rather than being beneficial to one's health. Statins should only be given to patients who have a really high risk of cardiovascular diseases, not everyone above the age of 50 despite the fact that most of them are healthy because I found that the over-prescription of statins leads to a higher risk of ICH. Therefore, understanding this newly discovered causal nature between LDL-c and ICH, physicians can make more informed decisions before prescribing statins to anyone above the age of 50, especially if they are healthy and run a lower risk of cardiovascular disease.

VIII. Conclusion

Thus far, there has been a myriad of associational studies regarding risk factors and ICH. However, I introduced a new approach to identifying risk factors that *cause* ICH without the effect of confounding variables that might produce errors in my results. Through Mendelian randomization, I was successfully able to identify four risk factors—hypertension, BMI, HDL-c, and LDL-c—as causative risk factors for ICH. Additionally, through my analysis of the results, I was able to highlight the central issue of statin therapy as a treatment to lower LDL-c levels as it can definitively lead to a greater risk of ICH in patients.

Despite these significant findings, there are some flaws with Mendelian Randomization that make it difficult to apply these results to real clinical settings. Most of the methods for Mendelian Randomization rely on some sort of assumptions to generate causal estimates. Through this, the results vary slightly from method to method. Thus, it is hard to apply these results to real clinical settings. Therefore, the future directions of this research can lie in doing a clinical study

to verify the truth of these causal results. Additionally, further research can be done to use machine-learning techniques to verify if the causal results can also be used as early diagnostic tools by using symptoms (hypertension, BMI, LDL-c, HDL-c, etc...) to detect ICH. This would be beneficial in detecting and developing treatments for ICH before it becomes too severe. Research regarding Mendelian Randomization studies on ICH can be continued to test other risk factors that might be associated with ICH.

REFERENCES

[1] J. Kaj, "Aneurysms," vol. 247, no. 1, pp. 110–125, 1982.

[2] O. Yashuiro, H. Toru, S. Masao, Y. Tsutomu, and N. Hideaki, "Size of cerebral aneurysms and realted factors in patients with subarachnoid hemorrahge," vol. 61, no. 3, pp. 239–245, 2004.

[3] D. Candice, S. Shoichiro, Z. Shihong, S. Else, Z. Danni, C. Xiaoying, H. Maree, A. Hisatomi, H. Jun, H. Emma, S. Rustam, R. Thompson, D. Leo, L. Pablo, L. Richard, S. Christian, C. John, and A. Craig, "Intracerebral hemorrhage location and outcome among interact2 participants," vol. 88, no. 15, p. 1408–1414, 2017.

[4] F. Margret, G. Alan, C. Yuchiao, H. Elaine, H. Lori, J. Nancy, and S. Daniel, "Death and disability from warfarin-associated intracranial and extracranial hemmorhages," vol. 120, no. 8, pp. 700–705, 2007.

[5] W. Xiang, D. Yan, Q. Xuangqian, H. Chenggaung, and H. Lijun, "Cholesterol levels and risk of hemorrhagic stroke," pp. 1833–1839, 2013.

[6] J. Frost. (2019) Causation versus correlation in statistics. [Online]. Available: https://statisticsbyjim.com/basics/causation/

[7] P. Judea, "Causal inference in statistics: An overview," vol. 3, pp. 96–146, 2009.

[8] D. Neil, H. Michael, and D. George, "Reading mendelian randomisation studies: a guide, glossary, and checklist for clinicians," vol. 362, 2018.

[9] K. Maria, R. Christian, S.-G. Monika, and B. Maria, "Randomized control trials," vol. 663, no. 8, 2011.

[10] K. Csaba and K.-Z. Kamyar, "Observational studies versus randomized control trials: Avenues to causal inference in nephrology," vol. 19, no. 1, 2012.

[11] B. Stephen, B. Jack, F. Tove, I. Erik, and T. Simon, "Sensitivity analyses for robust causal inference from mendelian randomization analyses with multiple genetic variants," vol. 28, no. 1, pp. 30–42, 2017.

[12] C. E., L. M., B. S., L. D., and A. M., "The role of age in intracerebral hemorrhage: An intricate relationship," vol. 1, no. 5, 2014.

[13] N. Sunil, S. P., S. Sharma, S. Ratnakar, and D. Tarun, "Etiology and outcome determinants of intracerebral hemorrhage in a south indian population, a hospital-based study," vol. 15, no. 4, pp. 263–366, 2012.

[14] R. Collins. (2020) About uk biobank. [Online]. Available: 14. https://www.ukbiobank.ac.uk/

[15] B. Clare, F. Colin, P. Desislava, B. Gavin, E. Lloyd, S. Kevin, M. Allan, V. Damjan, D. Olivier, O. Jared, C. Adrian, W. Samantha, M. Gil, L. Stephen, D. Peter, and M. Jonathan, "Genome-wide genetic data on $\sim 500,000$ uk biobank participants," 2017.

[16] B. Stephen, F. Christopher, A. Elias, S. James, and H. Joanna, "A robust and efficient method for mendelian randomization with hundreds of genetic variants," vol. 11, no. 376, 2004.

[17] H. Fernando, S. George, and B. Jack, "Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption," vol. 46, no. 6, pp. 1985–1998, 2017.

[18] H. Liang-Dar, L. Deborah, F. Rachel, E. David, and W. Nicole, "Using a two-sample mendelian randomization design to investigate a possible causal effect of maternal lipid concentrations on offspring birth weight," vol. 48, no. 5, pp. 1457–1467, 2019.

[19] B. Stephen and T. Simon, "Interpreting findings from mendelian randomization using the mr-egger method," vol. 32, no. 5, pp. 377–389, 2017.

[20] G. Hemani and P. Haycock. (2020) Perform mr. [Online]. Available: https://mrcieu.github.io/TwoSampleMR/articles/perform_mr.html

[21] C. Alexander. (2015) 10 things to know about multiple comparisons. [Online]. Available: https://egap.org/resource/10-things-to-know-about-multiple-comparisons

[22] M. Carmel, W. Ian, and A. Albert, "Age, hypertension and arterial function," vol. 34, no. 7, pp. 665–671, 2007.

[23] A. Mohamad, B. Ruchi, W. Christina, and B. Brandi, "Racial disparities in stroke awareness: African americans and caucasians," vol. 33, no. 4, pp. 462–490, 2011.

[24] M. Margaret, "Medicine and the media. statins for all?" vol. 345, 2012.