Mendelian Randomization for Causal Inference of the Relationship between Obesity and 28-day Survival following Septic Shock

by

Dilshani Induruwage

B.Sc., University of Colombo, 2014

Project Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2017

Approval

Name:	**Dilshani Induruwage**
Degree:	**Master of Science (Statistics)**
Title:	***Mendelian Randomization for Causal Inference of the Relationship between Obesity and 28-day Survival following Septic Shock***
Examining Committee:	**Chair:** Dr. Rachel Altman Associate Professor

Dr. Brad McNeney
Senior Supervisor
Associate Professor _____

Dr. Jinko Graham
Supervisor
Professor _____

Dr. Keith R. Walley
External Examiner
Professor
Department of Medicine
University of British Columbia _____

Date Defended: 10 August 2017

Abstract

Septic shock is a leading cause of death in intensive care units. Septic shock occurs when a body-wide infection leads to low blood pressure, and ultimately organ failure. Some recent studies suggest that overweight and obese patients have a better chance of survival following septic shock than normal or underweight patients. In this project we apply Mendelian randomization to assess whether the observed obesity effect on 28-day survival following septic shock is causal or more likely due to unmeasured confounding variables. Mendelian randomization is an instrumental variables approach that uses genetic markers as instruments. Under modelling assumptions, unconfounded estimates of the obesity effect can be obtained by fitting a model for 28-day survival that includes a residual obesity term. Data for the project comes from the Vasopressin and Septic Shock Trial (VASST). Our analysis suggests that the observed obesity effect on survival following septic shock is not causal.

Keywords: Obesity; Septic Shock; Causal Inference; Instrumental Variables; Mendelian Randomization

Acknowledgements

I would like to express my gratitude to my senior supervisor Dr. Brad McNeney for his continuous support and guidance throughout my time at SFU. I would never have been able to finish my project without his tremendous help, and I really appreciate his unlimited patience and kindness to me. I would also like to give special thanks to Dr. Jinko Graham for her valuable suggestions, guidance and resources for improving my project.

At the same time, thank you to Dr. Rachel Altman, Dr. Jinko Graham and Dr. Keith Walley, for agreeing to be part of my committee and for the insightful comments and reviews on this project. Furthermore, I would like to thank Dr. Keith Walley for providing an interesting research idea and data for my master's project.

I also want to thank all the faculty members of the Department of Statistics and Actuarial Science who taught me during my time at SFU, especially Dr. Steve Thompson, Dr. Boxin Tang, Dr. Rachel Altman, Dr. Dave Campbell and Dr. Tom Loughin. My sincere gratitude to Sadika, Kelly, and Charlene for their kind assistance. I extend my gratitude to my graduate student colleagues for their friendship and camaraderie and for the fun times we had together.

Finally and most importantly, I wish to thank my better half, Chamara, for standing by my side all the time and making me laugh every time when I get stressful. Also, I would like to express my special thanks to my parents. I am eternally grateful for all the love and support they have given me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Septic shock has been reported as one of the most common causes of death in the Intensive Care Unit (ICU), with a mortality rate of up to 45% [SCCM]. According to the U.S. National Library of Medicine, septic shock occurs when a body-wide infection leads to dangerously low blood pressure and that can lead to heart failure, organ failure and death. People with septic shock are usually cared for in ICUs. It most commonly affects people with weakened immune systems. On the other hand, obesity is a growing health problem in the world and it is reported that nearly 30% of the world's population are either obese or overweight [WHO]. Although higher BMI is associated with various diseases and reduces overall life expectancy, it has been suggested that higher BMI may improve survival following septic shock [Wacharasint et al., 2013]. This study is of the causal relationship between BMI and 28-day survival following septic shock.

Several studies have been conducted to test the association between BMI and survival following septic shock. Some suggest that patients with higher BMI have lower risk of death from septic shock [Wurzinger et al., 2010, Wacharasint et al., 2013], while others suggest that BMI has no effect on survival [Arabi et al., 2013, Gaulton et al., 2015]. A possible explanation for the conflicting results is unmeasured confounding variables. A confounder is a variable that is associated with both the exposure of interest and the outcome. Failing to account for a confounding variable will lead to biased inference of the exposure effect, and differences between the distribution of a confounder in different study populations will lead to differential bias. It is therefore of interest to obtain an unconfounded, or *causal* estimate of the effect of BMI on survival.

The starting point for this project is Wacharasint et al. [2013]. These authors conducted a retrospective analysis of the Vasopressin and Septic Shock Trial (VASST) to determine whether being overweight or obese altered mortality of septic shock. They found that mortality in obese (BMI of 30 kg/m^2 or more) and overweight (BMI of 25-29.9 kg/m^2)

patients was significantly lower than in low or normal weight patients (BMI of less than 25 kg/m^2). Whereas Wacharasint et al. [2013] analysed the effect of BMI on survival time, in this project we simplify and use a binary indicator of 28-day survival as the outcome. After data cleaning and adjustment for known confounders (Chapter 3), we find that an overweight patient is estimated to have a 1.4-times higher odds of survival than a normal or low weight patient with the same values of the confounding variables, and an obese patient is estimated to have a 1.8-times higher odds of survival than a normal or low weight patient with the same confounders. The objective of this project is to re-analyse the VASST data using methods for causal inference. The question is whether the observed association between BMI on 28-day survival is causal or more likely due to confounding.

Instrumental variables (IVs) are used to control for unmeasured confounding. An IV G is a variable that is (i) predictive of the exposure X, (ii) associated with the outcome Y only through the association with X and (iii) is independent of unobserved confounders U [Smith and Hemani, 2014]. Figure 1.1, provides a graphical representation of instrumental variable (IV) assumptions. The arrows indicate the direction of causal relationships between variables. The absence of any arrow between two variables indicates that the variables are not related. In addition to G, X, U and Y, the Figure includes observed confounders O; these are depicted as independent of G, but such independence is not required by the models we use in Chapter 2. In the Econometrics literature, the variables with no arrows pointing towards them (such as G, U and O) are called *exogenous* variables, and those with arrows pointing towards them (such as X and Y) are called *endogenous* variables. Thus, exogenous variables are not influenced by other variables in the system of causal relationships and endogenous variables are internal, being affected by the other variables.

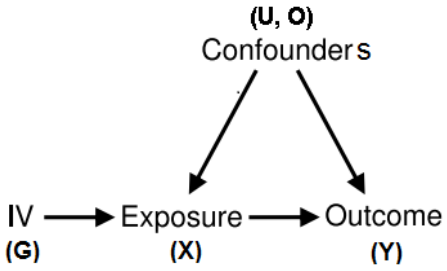

Figure 1.1: Instrumental variable (IV) assumptions

IV methods infer a causal relationship between the endogenous exposure and the outcome by studying the association between the exogenous instrumental variable and the outcome. According to the diagram, the exposure-outcome association is confounded by unmeasured confounders, but the IV-outcome association is not since there is no direct line between IV and unmeasured confounders. Since IVs are associated with the exposure, but

not directly associated with the outcome, any association of the IVs with the outcome must come *via* the IV's association with the exposure, and this is taken as evidence for causality between exposure and the outcome. Causal inferences based on IV methods are only valid if these IV assumptions are satisfied. However, it is difficult to prove that these assumptions hold. It is only possible to justify the validity of these assumptions based on subject matter or background knowledge.

In our study, BMI is the exposure, 28-day survival following septic shock is the outcome and observed confounders are age, gender and APACHE II score. In our data set, vasopressin and norepinephrine drug data were not available. If drug data were available, it would have been a confounder and would affect second-stage model and might lead to better prediction of outcome. We used genetic variants (single-nucleotide polymorphisms, or SNPs) as instruments. Genetic variants are suitable instruments because they are inherited at conception and do not change over one's lifetime. This random inheritance from parents to offspring ensures that association between genetic variants and outcomes are not likely to confounding [Smith and Hemani, 2014]. Instrumental variable analysis with genetic variants as instruments is called Mendelian Randomization (MR).

The project is organized as follows. In Chapter 2 we discuss causal inference methods for the VASST data and in Chapter 3 we apply these methods. Chapter 4 summarizes our findings. Supplementary data summaries and analysis results are given in Appendix A.

Chapter 2

Instrumental Variables: Models and Inference

In this chapter we first describe models for the relationships between the variables in Figure 1.1, and then discuss the IV assumptions.

2.1 Models

As discussed in the Introduction, an association between the IV, G, and the outcome, Y, implies a causal relationship between the exposure, X and Y. Thus, under the IV assumptions, testing for a G-Y association is a test of the causal effect of X on Y. However, to *estimate* causal effects we must specify models for the exposure and outcome. We use the general model of Terza et al. [2008], specialized to the case of a binary outcome and a categorical exposure. In this general model, there is a first stage model for the mean exposure as a function of the IVs and observed confounders, and a second stage model for the mean outcome as a function of the exposure and all confounders. We discuss the second stage model first, as it is the model of primary interest.

2.1.1 Second-stage Model

With a binary outcome Y, the second stage model is a logistic regression of Y on a vector of covariates Z that encodes information on X, O and U. Let $O = (O_1, \ldots, O_p)$ denote a row vector of information on the observed covariates and $X = (X_1, X_2)$ encode BMI status, with $X = (0, 0)$ for low or normal weight, $X = (1, 0)$ for overweight and $X = (0, 1)$ for obese. The precise definition of $U = (U_1, U_2)$ depends on the first stage model, and is described below. Lastly, let $Z = (O, X, U)$. Corresponding to Z is a column vector $\beta = (\beta_O^T, \beta_X^T, \beta_U^T)^T$ of parameters of length $p + 2 + 2$, where ^{T} denotes vector transpose. For convenience we suppose that $O_1 \equiv 1$ so that β_{O1} is an intercept term. The logistic model is for the log-odds

of survival:

$$\log\left(\frac{P(Y=1|Z)}{P(Y=0|Z)}\right) = Z\beta = \sum_{k=1}^{p} O_k\beta_{Ok} + X_1\beta_{X1} + X_2\beta_{X2} + U_1\beta_{U1} + U_2\beta_{U2}. \qquad (2.1)$$

Equivalently,

$$P(Y=1|Z) = \frac{\exp(Z\beta)}{1+\exp(Z\beta)}.$$

The parameters β_{X1} and β_{X2} are the causal effects of BMI on survival and are the object of inference. Model (2.1) cannot be fitted because U is not observed. Without U the model is of $P(Y=1|O,X)$ which depends on the unknown joint distribution of (Y,O,X,U) in the population and need not be of logistic form [Greenland et al., 1999]. To illustrate the dependence of $P(Y=1|O,X)$ on the joint distribution of (Y,O,X,U), suppose U is a discrete random variable taking values $u_1, u_2, \ldots,$. Then

$$P(Y=1|O,X) = \frac{P(Y=1,O,X)}{P(O,X)} = \frac{\sum_{i=1}^{\infty} P(Y=1,O,X,U=u_i)}{\sum_{i=1}^{\infty}\sum_{y=0}^{1} P(Y=y,O,X,U=u_i)}.$$

2.1.2 First-stage Model

The first stage model is a multinomial logistic regression of X on G and O. Whereas logistic regression is used when the outcome has two possible categories, multinomial logistic regression is used when the outcome has more than two categories. A baseline category is chosen (e.g., low or normal weight) and the probability of each remaining category relative to the probability of the baseline category is modeled as log-linear in the covariates. Let G denote the genotype information. This could be a column vector of allele counts at multiple SNPs, or an allele score that is a weighted average of allele counts. Let $W=(O,G)$. Then the multinomial logistic model is

$$\log\left(\frac{P(X=(1,0)|W)}{P(X=(0,0)|W)}\right) = W\alpha_1; \quad \log\left(\frac{P(X=(0,1)|W)}{P(X=(0,0)|W)}\right) = W\alpha_2, \qquad (2.2)$$

where $\alpha_1 = (\alpha_{1O}^T, \alpha_{1G}^T)^T$ is a column vector of regression parameters for overweight *versus* low/normal and $\alpha_2 = (\alpha_{2O}^T, \alpha_{2G}^T)^T$ is a column vector of regression parameters for obese *versus* low/normal. The linear predictors $W\alpha_i$ expand to $\sum_{k=1}^{p} O_k\alpha_{iOk} + G\alpha_{iG}$. It can be shown that an equivalent specification of model (2.2) is

$$P(X=(1,0)|W) = \frac{\exp(W\alpha_1)}{1+\exp(W\alpha_1)+\exp(W\alpha_2)} \equiv p_1(W), \text{ and}$$

$$P(X=(0,1)|W) = \frac{\exp(W\alpha_2)}{1+\exp(W\alpha_1)+\exp(W\alpha_2)} \equiv p_2(W).$$

In Terza et al. [2008] the unobserved confounder variables $U = (U_1, U_2)$ from the second stage model are defined as residuals from the first stage model; that is, $U = X - p(W)$, where $p(W) = (p_1(W), p_2(W))$. Comments on this definition are given in the next section. The utility of defining U to be an error term is that it can be estimated by fitting the first stage model, and these residuals can be used in the second stage regression. That is, we (i) fit model (2.2) to n observations to obtain $\hat{p}(W_i)$, $i = 1, \ldots, n$, (ii) calculate the residuals $\hat{U}_i = X_i - \hat{p}(W_i)$, and then (iii) use $\hat{U}_i$, $i = 1, \ldots, n$ in place of U_i, $i = 1, \ldots, n$ in the second stage model. Terza et al. [2008] show that this two-stage residual inclusion (2SRI) method yields consistent estimates of the causal effects, meaning that, as the sample size grows, the causal effect estimates tend in probability to the true values.

2.2 IV Assumptions

2.2.1 Definition of U

In Figure 1.1 we see that G, O, and U all affect X, but according to the first stage model they do so in different ways. The multinomial logistic regression model $p(W)$ includes G and O, but cannot include the unobserved U. Instead, unmeasured confounders are collected together into an additive error term U in the model $X = p(W) + U$. These U then appear as covariates in the second stage model. It seems reasonable to suppose that the additive errors from the first stage model depend on the unmeasured confounders. However, it seems like a strong assumption to suppose that the second stage model depends on the unmeasured confounders only through these additive errors.

2.2.2 Population Stratification

It is important to take appropriate measures to avoid introducing confounding of the G-Y relationship through population stratification. Population stratification occurs when there exist population subgroups that experience both different disease rates and have different frequencies of alleles of interest [Lawlor et al., 2008]. Figure 2.1 depicts the confounding effect of population stratification.

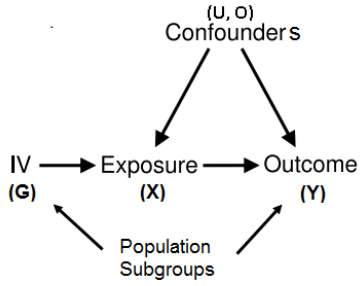

Figure 2.1: Effect of population stratification

To avoid population stratification, one can restrict analyses to ethnically homogeneous groups, and apply correction methods using principal component analysis (PCA). Principal components (PCs) obtained from PCA are orthogonal axes of variation that could represent the population structure of individuals. The top few PCs can be used as covariates in the analysis, in order to adjust for any existing population structure [Price et al., 2006]. An alternative to PCA is multidimensional scaling (MDS; Mardia et al. [1979]).

2.2.3 Many Weak Instruments

An instrument is considered a weak instrument if it explains only a small proportion of variance in the exposure. Weak instruments provide less information about the causal effect. Individual SNPs are weak predictors of BMI, since BMI is a complex trait. To improve the strength of IV, multiple SNPs can be used as IVs [Palmer et al., 2011]. If multiple SNPs cumulatively explain more variability in the exposure, they can jointly serve as better instruments to improve the prediction of the exposure and its causal effect estimate on the outcome. Thus, we use multiple SNPs as instruments to predict BMI.

Multiple instruments can be used as separate explanatory variables or they can be used to construct a single allele score. However, instrumental variable estimates of causal effects could be biased when using many weak instruments as separate explanatory variables [Davies et al., 2015]. Using many instruments as separate explanatory variables will tend to overfit BMI in the first stage and hence predicted BMI will be very similar to observed BMI. Thus we will essentially use observed BMI in the second stage, leading to confounded estimates of the BMI effect. Davies et al. [2015] suggested that constructing a single allele score such as unweighted or weighted Genetic Risk Score (GRS) can eliminate this bias.

Single Allele Scores

Each genetic variant is coded as 0, 1, or 2 depending on the combination of BMI-increasing alleles.

The weighted genetic risk score (GRS) is calculated as the weighted sum of alleles of SNPs associated with the exposure of interest, with weights equal to the published per-allele effects for the exposure. For each individual i, the weighted genetic risk score is calculated using an additive genetic model. The score is the product of individual's allele count for the j^{th} SNP and the weight for the effect of the j^{th} SNP on the exposure, across all J SNPs.

$$WGRS_i = \sum_{j=1}^{J} \hat{\beta}_j * allelecount_{ij}$$

where, $\hat{\beta}_j$ is the estimated weight of the effect of j^{th} SNP on the exposure and $allelecount_{ij}$ is the allele count for the j^{th} SNP of i^{th} individual.

2.3 Bootstrap

The causal effect estimates of 2SRI method introduced by Terza et al. [2008] were shown to be consistent; however, the estimated variances of the second-stage model are incorrect due to inclusion of residuals rather than the actual error terms in the second-stage regression. These estimated residuals are an extra source of variation that is not accounted for in the standard error from the logistic regression. We can utilize the nonparametric bootstrap to obtain approximately correct standard errors and confidence intervals from two-stage models [Guan, 2003].

The bootstrap is a resampling technique. Suppose that we have a random sample of size n from an unknown distribution, and we want to make statistical inferences about parameters. In our case the parameters are $\beta = (\beta_O^T, \beta_X^T, \beta_U^T)^T$ of length $p + 2 + 2$. The method is demonstrated as follows:

1. draw a random sample of size n with replacement from the data set;
2. fit the first-stage regression model to obtain $\hat{p}(W)$;
3. calculate the residuals, $\hat{U} = X - \hat{p}(W)$;
4. fit the second-stage model using $\hat{U}$ as explanatory variables;
5. repeat steps 1-4 for N number of bootstrap replicates, and
6. empirical standard errors are the standard deviation of the empirical distribution of the estimates and a 95% confidence interval is given by the 2.5^{th} and 97.5^{th} percentiles of the empirical distribution [Efron and Tibshirani, 1986].

Chapter 3

Application

In this chapter we apply instrumental variables methods to the VASST data. VASST is a multicenter, randomized, stratified, double-blind trial, evaluating the efficacy of vasopressin versus norepinephrine on mortality in patients with septic shock [Russell et al., 2008]. VASST collected phenotype and genotype data. The phenotype data consists of age, gender, ethnicity, weight, height and APACHE II (Acute Physiology and Chronic Health Evaluation II) measurements of 632 patients with sepsis. APACHE II is a severity-of-disease classification system, which is applied within 24 hours of admission of a patient to an ICU. The score can range from 0 to 71, with higher values corresponding to more severe disease and a higher risk of death [Knaus et al., 1985]. Age of the patients ranged from 17 to 99 years with median age 63. Around 60% of the patients were males and the APACHE II score ranged from 0 to 49.

Body mass index (BMI) for patients was calculated as weight (in kilograms) divided by height squared (in meters). Participants with BMI less than 14 kg/m^2 or greater than 80 kg/m^2 were excluded from the analyses on the basis that values outside this range are unlikely to be physiologically plausible [Shungin et al., 2015]. Among the patients, 589 patients had valid BMI measures. Then patients were grouped into BMI categories established by the World Health Organization (WHO): underweight (BMI < 18.5 kg/m^2), normal weight (BMI 18.5-24.9 kg/m^2), overweight (BMI 25-29.9 kg/m^2) and obese (BMI $\geq$ 30 kg/m^2). Since only 18 patients (4%) were present in the underweight category, normal and underweight patients were grouped into a single category of BMI less than 25 kg/m^2. The outcome measurement was 28-day survival, a binary variable labeled 0 for patients who were admitted to the ICU and died before 28-day follow-up period and 1 for patients who survived 28-day follow-up period, respectively.

The sample was restricted to patients with self-reported ethnicity as Caucasian to limit the influence of any population stratification. We then removed 17 of these patients because

they were outliers, with respect to the first two principal coordinates (PCs) from multidimensional scaling (MDS) [Mardia et al., 1979] based on the identity-by-state (IBS) distance matrix [Purcell et al., 2007]. Further details related to MDS are described in section 3.2.1 and Appendix A. After data cleaning steps all the analyses were carried out with 476 Caucasian patients with complete data on 28-day survival and BMI, of whom 315 (66%) had survived 28-days. In Table 3.1, we show the baseline characteristics of patients and the outcome in each BMI category.

Table 3.1: Baseline characteristics among different BMI categories

Characteristics	BMI<25 kg/m^2 (n=172)	BMI 25-29.9 kg/m^2 (n=134)	BMI≥30 kg/m^2 (n=170)	p-value*
Demographics:				
Age (years)	59.43 (17.37)	61.02 (15.71)	61.59 (14.50)	0.73
Gender (Female)	63 (36.6%)	45 (33.6%)	79 (46.5%)	0.04
Severity of Illness:				
APACHEII score	25.53 (8.03)	26.34 (5.8)	26.70 (8.15)	0.11
D28 Survival:				
Yes	105 (61%)	90 (67.2%)	120 (70.6%)	0.16

Continuous variables are reported as means with standard deviations (SD) and categorical variables as frequencies with percentages.
* p-values for a nonparametric test of any association with BMI categories, as explained in the text

We tested whether any of the covariates had a significant association with BMI categories. The associations of baseline characteristics with BMI were tested using the Kruskal-Wallis test for continuous data and the chi-square test for categorical data. Association tests showed an association of gender with BMI (p=0.04) but no association with age (p=0.73) or APACHE II score (p=0.11). We further investigated the association between BMI and age. Figure A.1. in Appendix A shows BMI versus age, and suggests a quadratic effect of age. We therefore tested for an association between age^2 and the BMI categories and found age^2 was significantly associated with BMI (p=0.005).

3.1 Observational Association

To study the observational association between BMI and 28-day survival, we carried out logistic regression analyses considering BMI less than 25 kg/m^2 as the reference category and two binary variables, indicating overweight and obese, respectively. We evaluated the association of BMI and survival for both unadjusted and adjusted models for known confounders age, gender and APACHE II score (Table 3.2 and Table 3.3).

Table 3.2: Observational association between BMI and 28-day survival unadjusted for known confounders

Coefficients	Odds Ratio	95% Confidence Interval
BMI-overweight	1.3051	(0.8149, 2.1026)
BMI-obese	1.5314	(0.9779, 2.4098)

For a logistic regression model unadjusted for known confounders, an overweight patient is estimated to have a 1.3-times higher odds of survival than a normal or low weight patient (95% CI: 0.8-2.1) and an obese patient is estimated to have a 1.5-times higher odds of survival than a normal or low weight patient (95% CI: 0.98-2.4).

Table 3.3: Observational association between BMI and 28-day survival adjusted for known confounders age, gender and APACHE II

Coefficients	Odds Ratio	95% Confidence Interval
BMI-overweight	1.4372	(0.8815, 2.3591)
BMI-obese	1.8393	(1.1454, 2.9781)

The logistic regression model adjusted for known confounders, showed a significant relationship between BMI and 28-day survival (likelihood ratio test, p=0.03). We find that an overweight patient is estimated to have a 1.4-times higher odds of survival than a normal or low weight patient with the same values of the confounding variables (95% CI: 0.88-2.3), and an obese patient is estimated to have a 1.8-times higher odds of survival than a normal or low weight patient with the same confounders (95% CI: 1.1-3.0). Therefore, based on the observational study, the overweight and obese patients had a higher probability of survival compared to the patients with low or normal weight.

3.2 IV Analysis

Genetic variants (SNPs) for the IV analysis were extracted from the VASST genotype data. Genotype data was available on 662 patients and 1,199,187 SNPs.

3.2.1 Quality Control

We carried out quality control (QC) measures to ensure the quality of the genetic variants extracted for the analysis. For QC we followed the recommendations in Anderson et al. [2010] and all the QC analyses were performed using PLINK [Purcell et al., 2007]. SNPs

with minor allele frequency less than 5% and SNPs with more than 5% missing genotypes were excluded. After these quality control steps, variants that showed a significant deviation ($p<0.001$) from Hardy-Weinberg equilibrium (HWE) were identified and excluded from the study.

MDS was performed to investigate the population structure of the genome data in patients who self-reported as Caucasian. We identified 17 outliers and these patients were removed from the study. Outliers were detected as described in Appendix A Section A.1. After removing outliers, MDS was re-applied to the reduced data and since further principal coordinates did not reveal evidence of population structure, only the first two PCs were used to adjust for any existing population stratification [Price et al., 2006]. We used the top two MDS principal coordinates as covariates in the regression analyses as suggested by visual inspection of the scree plot.

3.2.2 Construction of the single allele score

Genome-Wide Association Studies (GWAS) have identified SNPs related to BMI at different chromosome locations. Our analyses were based on BMI-related SNPs identified by Speliotes et al. [2010] and Locke et al. [2015]. We used these established BMI-related SNPs and their reported effects on BMI in our study. We extracted 47 BMI-related SNPs from the VASST genotype data set. We found alternative SNPs, that can be used as proxies ($R^2 > 0.8$) for the SNPs that were not in our data using linkage disequilibrium (LD) and were able to find 11 "LD proxy" SNPs. Altogether 58 BMI-related SNPs were extracted from the VASST data and used as instruments for the IV analysis (Appendix A Section A.2).

Each SNP was coded 0, 1, or 2 depending on the combination of BMI-increasing alleles each individual had. As we discussed in section 2.2.3, a single allele score was used as an instrument to avoid the many-weak-instruments bias. We constructed the weighted genetic-risk score (GRS) by multiplying the number of risk alleles for the corresponding effect sizes, as reported by Speliotes et al. [2010] and Locke et al. [2015] (Table A.1 and Table A.2).

3.2.3 Estimation

We used the two-stage residual inclusion (2SRI) method explained in Section 2.1 to estimate the causal effect of BMI on 28-day survival.

First-stage Model

For the first-stage, a multinomial logistic regression model for BMI categories on GRS was fitted adjusting for known confounders and principal components. The normal/low weight

category was considered as the baseline category. The coefficients of the first-stage model using GRS as the IV are shown in Table 3.4.

Table 3.4: Regression coefficients of the first-stage model

	BMI-Overweight		BMI-Obese	
Coefficients	Estimate	95% Confidence Interval	Estimate	95% Confidence Interval
Intercept	-1.4448	(-3.2073, 0.3175)	-1.5702	(-3.1906, 0.0502)
GRS	-0.2702	(-1.5587, 1.0182)	0.7219	(-0.5190, 1.9629)
Age	-0.0220	(-1.2894, 1.2454)	0.1366	(-1.0970, 1.3702)
Age^2	-0.6919	(-1.9753, 0.5914)	-1.4602	(-2.7630, -0.1575)
Gender(Fe)	-0.0669	(-0.5493, 0.4154)	0.4591	(0.0150, 0.9032)
APACHEII	0.6291	(-0.8979, 2.1563)	1.0454	(-0.3968, 2.4877)
PC1	0.8206	(-0.4077, 2.0491)	1.1173	(-0.0400, 2.2747)
PC2	1.3111	(-0.1366, 2.7589)	0.3803	(-0.8525, 1.6131)

A comparison of the fitted first-stage model to the null model suggested that the fitted model is better than the null model, with p-value 0.01. However, a likelihood-ratio test to compare the models with GRS and without GRS suggested that GRS is not significant (p=0.29). Therefore, we find that the GRS is a weak instrument. However, since the GRS has been established as an instrument in other Mendelian randomization studies [Jokela et al., 2012, Tyrrell et al., 2016], we proceed to the second-stage analysis with the results of first-stage analysis.

Second-stage Model

Estimated residuals from the first-stage model were used as explanatory variables in the second-stage model accounting for unmeasured confounders. For the second stage, a logistic regression model for 28-day survival on BMI was fitted, adjusting for known confounders and principal coordinates.

The second-stage model uses estimated variables as explanatory variables, therefore, estimated standard errors for the logistic regression model are considered incorrect. Second-stage standard errors needed to be corrected for uncertainty in the estimated residuals. We used the bootstrap, as described in Section 2.3, with 10,000 bootstrap replicates to account for uncertainty in the estimated residuals and to correct standard errors.

Table 3.5: Mendelian randomization analysis of association between BMI and 28-day survival

Coefficients	Odds Ratio	Bootstrap 95% CI
BMI-overweight	1.2451	(0.7591, 2.3033)
BMI-obese	1.2781	(0.7497, 2.4366)
Residuals-overweight	1.4201	(0.8470, 2.2947)
Residuals-Obese	1.8178	(1.2065, 2.9195)

We find that an overweight patient is estimated to have a 1.2-times higher odds of survival than a normal or low weight patient with the same values of the confounding variables (bootstrap 95% CI: 0.76-2.3), and an obese patient is estimated to have a 1.3-times higher odds of survival than a normal or low-weight patient with the same confounders (bootstrap 95% CI: 0.75-2.4) when we used weighted GRS as the instrument. Based on the IV analysis, the overweight and obese patients had a slightly higher odds of survival compared to the patients with low or normal weight, but the odds ratios are not significantly different from one. Interestingly, the odds ratio for residual obesity, which represents the effect of unmeasured confounders on the probability of survival, is very similar to the estimated effect of obesity in the observational analysis (Table 3.3). This similarity suggests that the obesity effect in the observational analysis can be attributed to unmeasured confounders.

A graphical comparison of the estimated effects of BMI on 28-day survival for the unadjusted observational association, adjusted observational association and IV method are shown in Figure 3.1. Confidence intervals for the IV analysis are 95% bootstrap confidence intervals using the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distributions. The obese BMI category was associated with survival in the adjusted observational analysis, but this association weakened and was not statistically significant in the IV analysis, suggesting the absence of a causal effect between obesity and survival.

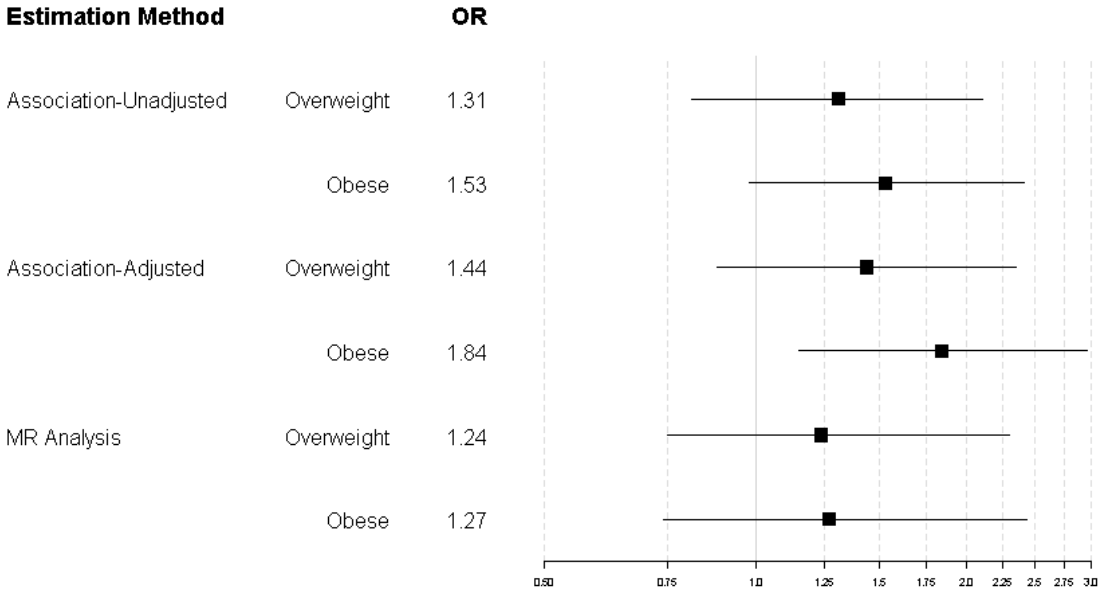

Figure 3.1: Estimated effects of BMI on survival for different estimation methods

Chapter 4

Concluding Remarks

The study conducted by Wacharasint et al. [2013] found a significant association between higher BMI and survival from septic shock in the VASST data. However, it was not certain whether the observed association is causal or more likely due to confounding. We re-analyzed the VASST data using an IV method to obtain an unconfounded estimate of the effect of BMI on survival. In particular, we applied Mendelian randomization with genetic variants as instrumental variables. We used 58 BMI-related SNPs combined into a genetic-risk score (GRS) to produce an instrument for BMI. The goal of introducing the genetic-risk score as an IV was to remove the effects of unmeasured confounders that may confound the relationship between BMI and survival. The GRS was used to predict BMI in the first-stage model. Based on the model introduced by Terza et al. [2008], unmeasured confounders were defined as residuals from the first-stage model. These residuals were used as explanatory variables in the second-stage model to control for confounders.

In this project, we first analyzed the data for the observational association and obtained similar results to Wacharasint et al. [2013] with a significant association between higher BMI and survival. Then, we applied the Mendelian-randomization approach to the same data to assess whether the observed relationship is causal or not. In our study, we could not detect a strong causal association between BMI and 28-day survival following septic shock, though our results suggested that overweight and obese patients had slightly higher odds of survival than normal or low weight patients with the same values of the confounding variables. Taken together, our analyses suggest that the observational finding that obesity increases the probability of 28-day survival is due to unmeasured confounders.

There were several limitations of our analyses. The main limitation is the small number of patients available in our analysis. The first-stage analysis indicated that the GRS used as the instrumental variable was weak. The genetic risk score constructed for this project as an instrumental variable has been established as an instrument in several other studies

[Jokela et al., 2012, Tyrrell et al., 2016]. However, we note that the SNPs used to construct the GRS do not explain very much variation in BMI. For example, the SNPs identified by Speliotes et al. [2010] only accounted for 1.45% of the variation in BMI and SNPs identified by Locke et al. [2015] only accounted for 2.7% of the variation in BMI. This suggests that SNPs alone cannot explain a complex trait such as BMI fully. Furthermore, our sample only included patients with Caucasian ethnicity; however, the genetic-risk score might be differently associated with BMI in different ethnic groups, so our findings may not generalize to other ethnic groups directly. Our data only showed observational associations when we treated BMI as an categorical variable. The first-stage model would have been a simpler least squares regression if we had used BMI, or some normalizing transformation of BMI, in the second-stage analysis. It is possible that the SNPs from the Speliotes et al. [2010] and Locke et al. [2015] studies are more predictive of BMI than they are of overweight and obesity status. However, we did not find an association between BMI and survival in the observational analysis of the VASST data, and so we used BMI categories throughout.

Bibliography

C. A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9), 2010.

Y. M. Arabi, S. I. Dara, H. M. Tamim, A. H. Rishu, A. Bouchama, M. K. Khedr, and A. Kumar. Clinical characteristics, sepsis interventions and outcomes in the obese patients with septic shock: An international multicenter cohort study. *Critical Care (London, England)*, 17(2), 2013.

N. M. Davies, S. V. Hinke Kessler Scholder, H. Farbmacher, S. Burgess, F. Windmeijer, and G. Davey Smith. The many weak instruments problem and mendelian randomization. *Statistics in Medicine*, 34(3), 2015.

B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1), 1986.

T. G. Gaulton, C. Marshall MacNabb, M.E. Mikkelsen, A. Agarwal, S. Cham Sante, C. Shah, and D. Gaieski. A retrospective cohort study examining the association between body mass index and mortality in severe sepsis. *Internal and Emergency Medicine*, 10(4), 2015.

S. Greenland, J. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, 14(1), 1999.

W. Guan. From the help desk: Bootstrapped standard errors. *Stata Journal*, 3(1), 2003.

M. Jokela, M. Elovainio, L. Keltikangas-Järvinen, G. D. Batty, M. Hintsanen, I. Seppälä, M. Kähönen, J. S. Viikari, O. T. Raitakari, T. Lehtimäki, and M. Kivimäki. Body mass index and depressive symptoms: instrumental-variables regression with genetic risk score. *Genes, Brain and Behavior*, 11, 2012.

W. Knaus, E. Draper, D. Wagner, and J. Zimmerman. Apache ii: A severity of disease classification system. *Critical Care Medicine*, 13(10), 1985.

D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. D. Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8), 2008.

Adam E. Locke, Bratati Kahali, Sonja I. Berndt, Anne E. Justice, Tune H. Pers, and Felix R. Day. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518, 2015.

K. V. Mardia, J. T. Kent, and J. M. Bibby. Multivariate analysis. *Probability and mathematical statistics*, 1979.

U.S. National Library of Medicine. Medical encyclopedia - septic shock. URL `https://medlineplus.gov/ency/article/000668.htm`.

T. M. Palmer, J. A. C. Sterne, R. M. Harbord, D. A. Lawlor, N. A. Sheehan, S. Meng, R. Granell, G. D. Smith, and V. Didelez. Instrumental variable estimation of causal risk ratios and causal odds ratios in mendelian randomization analyses. *American Journal of Epidemiology*, 173(12), 2011.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), 2006.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81, 2007.

J. A. Russell, K. R. Walley, J. Singer, A. C. Gordon, P. C. Hebert, D. J. Cooper, C. L. Holmes, S. Mehta, J. T. Granton, M. M. Storms, D. J. Cook, J. J. Presneil, and D. Ayers. Vasst investigators: Vasopressin versus norepinephrine infusion in patients with septic shock. *N Engl J Med*, 358, 2008.

SCCM. Society of critical care medicine, critical care statistics. URL `http://www.sccm.org/Communications/Pages/CriticalCareStats.aspx`.

Dmitry Shungin, Marilyn C. Cornelis, Kimon Divaris, Birte Holtfreter, John R Shaffer, Yau-Hua Yu, and Silvana P Barros. Using genetics to test the causal relationship of total adiposity and periodontitis: Mendelian randomization analyses in the gene-lifestyle interactions and dental endpoints (glide) consortium. *International Journal of Epidemiology*, 1, 2015.

George Davey Smith and Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):89–98, 2014.

E. K. Speliotes, C. J. Willer, S.I. Berndt, K.L. Monda, and G. et al Thorleifsson. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42, 2010.

J. V. Terza, A. Basu, and P. J. Rathouz. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics*, 27(3), 2008.

J. Tyrrell, S. E. Jones, R. Beaumont, C. M. Astley, R. Lovell, H. Yaghootkar, M. Tuke, K. S. Ruth, R. M. Freathy, J. N. Hirschhorn, A. R. Wood, A. Murray, M. N. Weedon, and T. M. Frayling. Height, body mass index, and socioeconomic status: mendelian randomisation study in uk biobank. *BMJ*, 352, 2016.

P. Wacharasint, J. H. Boyd, J. A. Russell, and K. R. Walley. One size does not fit all in severe infection: obesity alters outcome, susceptibility, treatment, and inflammatory response. *Critical Care*, 17(3), 2013.

WHO. World health organization, obesity and overweight. URL `http://www.who.int/mediacentre/factsheets/fs311/en/`.

B. Wurzinger, M. Dünser, C. Wohlmuth, M. Deutinger, H. Ulmer, C. Torgersen, and W. Hasibeder. The association between body-mass index and patient outcome in septic shock: A retrospective cohort study. *Wiener Klinische Wochenschrift*, 122(1), 2010.

Appendix A

Data summaries and analysis results

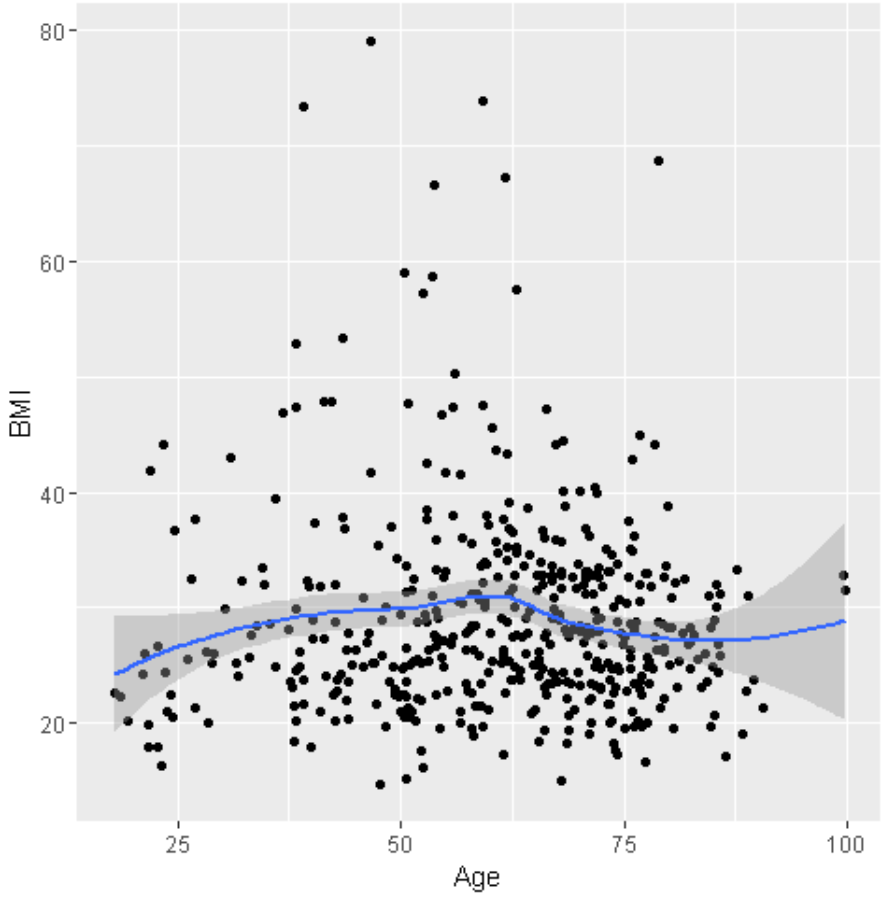

Figure A.1: Plot of BMI vs Age

A.1 Multidimensional scaling (MDS)

Before conducting MDS, SNPs were pruned to remove clusters of highly correlated SNPs. SNPs were pruned calculating the LD between each pair of SNPs and removing one of a pair of SNPs if the LD was greater than 0.5. In this way we were able to obtain a subset of SNPs in which all pairs have low correlations.

MDS was performed on the identity-by-state (IBS) distance matrix for these pruned data. Outliers were detected as follows. We first computed pair-wise distances between observations based on the first 20 MDS PCs. Observations more than six standard deviations from their nearest neighbour were declared outliers. For the covariate adjustment in the first- and second-stage models, we selected the top two PCs based on the scree plot (Figure A.2). Plink [Purcell et al., 2007] was used for prunning and MDS.

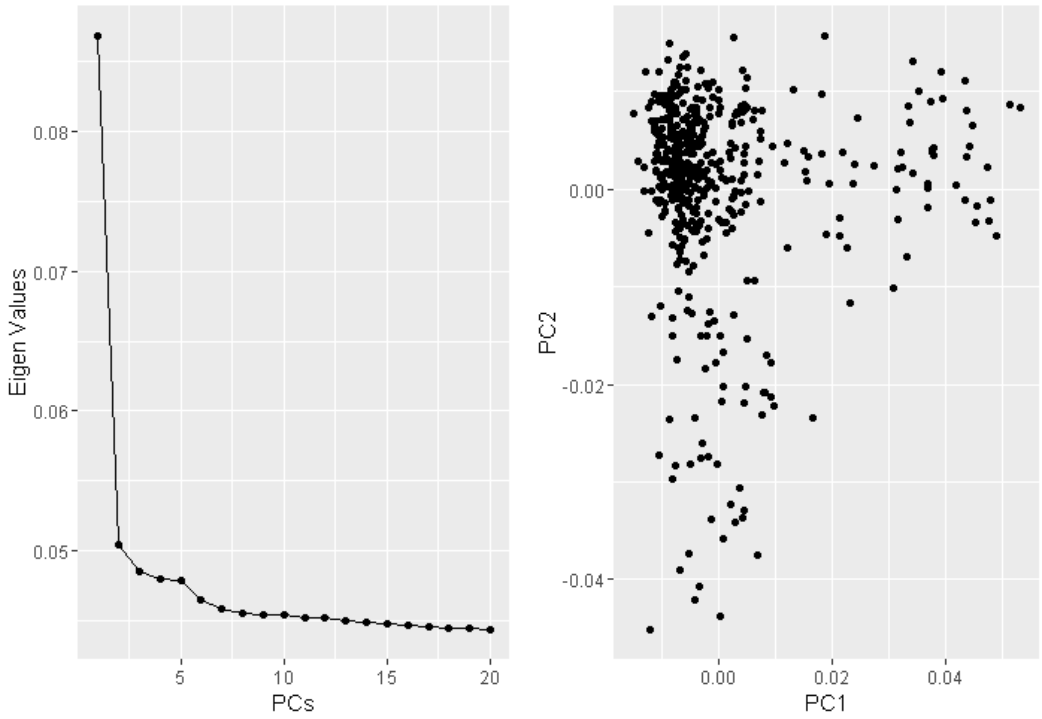

Figure A.2: Scree plot and Plot of the top 2 PCs used in the analysis

A.2 Construction of the single allele score

Speliotes et al. [2010] examined associations between BMI and around 2.8 million SNPs and confirmed 14 known obesity susceptibility SNPs and identified 18 new SNPs associated with BMI. Together, these 32 BMI related SNPs explained 1.45% of the variation in BMI. This study is not adjusted for age and gender, and the 32 BMI related SNPs and the corresponding effect sizes are reported in Table 1 of the original paper. Locke et al. [2015] identified 97 common genetic variants that were associated with BMI at genome-wide significance in the GIANT consortium in studies upto 339,224 people. The 97 BMI related SNPs account for about 2.7% of the variation in BMI and the corresponding effect sizes for the European population based study are reported in Table 7 of the original paper. 16 of the 32 SNPs from Speliotes et al. [2010] and 31 of the 97 SNPs from Locke et al. [2015] were presented in the VASST genotype data. For the rest of the SNPs that were not included in the VASST genotype data, we were able to find 11 LD proxy SNPs that were in linkage disequilibrium (LD) with $R^2 > 0.8$. These 58 BMI related SNPs were used to construct the single allele score as an IV for the analysis.

Some of the SNPs in the VASST genotype data set had switched alleles. For e.g. in the VASST data the SNP is coded as A/G but the risk allele for the particular SNP is reported as C. Then to identify the corresponding risk allele, the alleles needed to be *flipped* as A -> T, T -> A, C -> G and G -> C. Table A.1 and Table A.2 shows the 58 SNPs and the corresponding effect alleles and effect sizes for the BMI.

Table A.1: SNPs associated with BMI and corresponding effect sizes

	BMI related SNPs	Effect Allele	Effect Size
SNPs from Speliotes et. al.	rs2867125	C	0.31
	rs571312	A	0.23
	rs2815752	A	0.13
	rs7359397	T	0.15
	rs3817334	T	0.06
	rs29941	G	0.06
	rs543874	G	0.22
	rs987237	G	0.13
	rs7138803	A	0.12
	rs2241423	G	0.13
	rs2287019	C	0.15
	rs1514175	A	0.07
	rs13107325	T	0.19
	rs10968576	G	0.11
	rs13078807	G	0.10
	rs206936	G	0.06
SNPs from Locke et. al.	rs11583200	C	0.02
	rs3101336	C	0.035
	rs12401738	A	0.022
	rs2820292	C	0.02
	rs10182181	G	0.03
	rs11126666	A	0.015
	rs1016287	T	0.028
	rs11688816	G	0.02
	rs1528435	T	0.02
	rs7599312	G	0.017
	rs6804842	G	0.02
	rs16851483	T	0.056
	rs11727676	T	0.027
	7rs205262	G	0.022
	rs1167827	G	0.023
	rs4740619	T	0.017
	rs6477694	C	0.016
	rs1928295	T	0.021
	rs10733682	A	0.018
	rs11191560	C	0.027
	rs12286929	G	0.017
	rs11057405	G	0.026
	rs12429545	A	0.038
	rs7141420	T	0.029
	rs3736485	A	0.016
	rs758747	T	0.023
	rs1000940	G	0.021
	rs12940622	G	0.015
	rs1808579	C	0.02
	rs7243357	T	0.02
	rs17724992	A	0.02

Table A.2: LD proxy SNPs associated with BMI and corresponding effect sizes

BMI related SNPs	LD Proxy SNPs	R^2	Effect Allele	Effect Size
rs1558902	rs1421085	1	C	0.39
rs10938397	rs12641981	1	T	0.18
rs10767664	rs2030323	1	C	0.19
rs10150332	rs10146997	1	G	0.13
rs713586	rs713587	1	T	0.14
rs12444979	rs11639988	1	A	0.17
rs2112347	rs40060	0.96	T	0.10
rs887912	rs759250	0.98	T	0.10
rs1555543	rs11165643	0.98	C	0.06
rs4771122	rs1475221	0.82	G	0.09
rs4929949	rs9300092	1	C	0.06