The Application of Multiple Imputation to Explore Alcohol Consumption Among Young Adults and Its Impact on Body Mass Index

Maryam Tafreshi

A Thesis

in

The Department

of

Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Science (Mathematics) at

Concordia University

Montreal, Quebec, Canada

June 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: Maryam Tafreshi Motlagh

Entitled: The Application of Multiple Imputation to Explore Alcohol Consumption Among Young Adults and Its Impact on Body Mass Index

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Statistics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Arusharka Sen

_____ Examiner
Dr. Yogendra P. Chaubey

_____ Examiner

_____ Thesis Supervisor(s)
Dr. Lisa Kakinami

_____ Thesis Supervisor(s)

Approved by _____
Dr. Cody Hyndman Chair of Department or Graduate Program Director

_____

Dr. Francesca Scala.

ABSTRACT

"The Application of Multiple Imputation to Explore Alcohol Consumption Among Young Adults and Its Impact on Body Mass Index"

Maryam Tafreshi Motlagh

Obesity is associated with health issues such as high blood pressure, cardiovascular disease, and cancer. Concurrently, excessive alcohol consumption and its negative health effects among young adults are growing global public health concerns. However, the studies on the association between alcohol consumption and weight gain are inconsistent. In particular, analyzing weight gain and alcohol consumption over time is complicated by data missingness. Indeed, one potential reason for the mixed findings could be how those studies addressed missing data. Missing data can occur due to a variety of reasons including loss to follow-up and various non-response cases. As a result, missing data can lead to loss of efficacy in data analysis and is one of the main challenges of longitudinal studies and large surveys. Therefore this thesis attempted to address missing data challenges by utilizing multiple imputation (MI). Generalized linear models were then conducted to assess the relationship between alcohol consumption with BMI, adjusting for age at baseline, sex, race, smoking, depression status, relationship status, employment status, student status and physical activities. Analyses were conducted separately for a complete-case dataset and the imputed dataset.

Data were from the Nicotine in Dependence in Teens (NDIT) cohort, a 20-year prospective cohort initiated in 1999 (n=1294, 52% female). Data were collected every three months when participants were in high school (grade 7 to grade 11; 1999-2005, 20 cycles). After high school, follow-up assessments were conducted approximately every four years (2007-08 and 2011-12 for cycles 21 and 22). The GLM results demonstrated that there was no association between alcohol consumption and BMI at cycle 21, but at cycle 22 a negative association was detected. Many similarities between the complete-case and MI general linear models were observed. However, estimates and standard errors were different between the two models, and were generally smaller in the MI models compared with the complete-case.

Although counterintuitive, the negative association is consistent with the existing literature. Preliminary evidence from the literature further suggests there is an interaction with age on the alcohol and obesity relationship. Indeed, it seems that alcohol consumption does not have an immediate impact on body mass among young adults, or it may have an inverse relationship. However, as alcohol remains one of the main contributing factors to chronic diseases such as high blood pressure, cardiovascular disease, and cancer, it is perhaps a latent effect that does not show any immediate positive relationship with symptoms of these diseases at early stages of young adulthood. Specific details about the alcohol use (such as type of beverage or calorie count) were not available in this study. While this is similar to other cross-sectional study limitations, they are important confounding variables that should be addressed in future studies. Although the statistical methodology is not consistently utilized in the alcohol and obesity risk literature, this thesis demonstrated the use of MI to produce unbiased estimates and smaller standard errors compared to the complete cases analysis. Further research is needed on the same cohort to further track the weight changes and other possible health problems due to alcohol use in the long-term.

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Dr. Lisa Kakinami, for her valuable insights, constant guidelines, continuous motivation and encouragement during this long journey.

I like to extend my appreciation for the learning opportunity provided by the professors of the Department of Mathematics and Statistics of Concordia University. I am also thankful to members of the Mathematics and Statistics Department for their help and cooperation during my study at Concordia University.

Last but not the least, I like to thank my family for believing in me during tough times. Finally, I like to thank my family for their love and support.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIC	Akaike Information Criterion
BMI	Body mass index
CTADS	Canadian Tobacco Alcohol and Drug Survey
FIML	Full Information Maximum-Likelihood
GLM	General Linear Model
MAR	Missing at Random
MCAR	Missing completely at random
MCMC	Markov Chain Monte Carlo
MET	Metabolic Equivalent of Task
MI	Multiple Imputation
MNAR	Missing not at Random
NDIT	Nicotine Dependence in Teens
OR	Odds Ratio
SD	Standard Deviation
SE	Standard Error

1. Introduction

Obesity

Normal body fat varies between 25-30% (Williams, et al., 1992). When body fat exceeds these proportions (adiposity), it is associated with health risks such as diabetes, high blood pressure, heart disease, and cancer (Freedman, et al., 2007). Excessive adiposity is one of the major causes of morbidity and mortality (Singh, et al., 2008). However, measuring adiposity accurately is costly as it requires blood tests, computerized tomography, scanning, or the use of x-rays (Ellis, 2001). Therefore, a person's body mass index (BMI, calculated based on the person's height and weight) is oftentimes used as a proxy for adiposity.

Overweight and obesity are defined by the World Health Organization as BMI 25 and above, and 30 and above, respectively (Statistics Canada, 2015). It has been noted that the proportions of young people with obesity and overweight in Canada have increased significantly over the past four decades (Shields, et al., 2011). Most adults who gain weight in middle-age had a history of gaining weight in young adulthood (Lewis, et al., 2000). Thus, as obesity during adolescence and young adulthood tracks into adulthood, monitoring obesity during this period is essential (Public Health Agency of Canada, 2011).

Obesity Determinants

Different factors, including biological, psychological, sociocultural, and genetics can increase the risk of obesity (Suter & Tremblay, 2008). For instance, eating patterns and physical activity are among the behavioral factors that can lead to weight changes (Statistics Canada, 2015). In particular, alcohol contains 7.1 kcal per gram (Yeomans, 2010) and is thus one of the risk factors for gaining weight because of the high-calorie content (Suter, 2005). It also causes fat accumulation by inhibiting fat oxidation (Suter & Tremblay, 2008). Approximately 50-75% of Canadian young adults drink alcohol (Albanese & Bryson, 2015). However, only a few studies have looked at alcohol consumption and risk for overweight or obesity among teenagers and young adults (Oesterle, et al., 2004; Pajari, et al., 2010).

Knowledge Gaps

Alcohol consumption remains one of the most important preventable causes of death and disability in most countries (De Castelnuovo, et al., 2006). It is now well-established that alcohol consumption has negative effects on health issues, such as obesity, blood pressure, cardiovascular disease, and cancer in adults (Collins, 2016). However, the influence of alcohol consumption on BMI among young adults remains unclear. To be specific, the research to date has tended to focus on alcohol consumption in middle-aged or older adults rather than young adults. Many young adults try drinking alcohol at different stages of their education and continue to drink in their adulthood (Hingson, et al., 2006). Several adult studies have concluded that the quantity of drinking alcohol "occasionally" is associated with BMI, while more frequent drinking is associated negatively (Bendsen, et al., 2013; Sayon-OreaC, et al., 2011; Yeomans, 2010). For instance, some studies showed that the risk of obesity for those who drank alcohol 1-2 drinks per week (Odds Ratio [OR] =1.8) and those who consumed 3-5 drinks per week (OR=1.6) was higher compared to daily drinkers (Dumesnil, et al., 2013). On the other hand, some results showed a negative impact on gaining weight for daily alcohol consumers compared to non-daily drinkers on abdominal obesity (Dorn, et al., 2003). Previously published studies on the effect of alcohol consumption on BMI are not consistent in different age groups, such as young adults. Thus this thesis attempted to examine the relationship between alcohol intake and BMI among a prospective cohort of young adults.

Furthermore, one of the main longitudinal study challenges is missing data. Missing data can occur because of non-response or loss to follow-up. In addition, applying imputation methods for missing information in these prospective investigations are not utilized very often. In several studies, all the analyses were conducted on all complete and observed information (Molenberghs & Ibrahim, 2009).

As a result, this thesis attempts to address several knowledge gaps regarding alcohol consumption and BMI by: (1) focusing on young adults by analyzing existing data from the longitudinal Nicotine Dependence in Teens (NDIT) study, and (2) utilizing multiple imputation methods to address non-response and missing data issues. By applying imputation methods, this study

explores missing data in alcohol consumption among young adults and its effects on BMI. Results between imputed and complete case analyses will be compared.

This thesis will be organized in the following order: (I) introduction and literature review for alcohol consumption and its impact on gaining weight (II) method section containing the overview of the methods of handling missing data, as well as the description of the study cohort and variable selection for statistical analysis; (III) results and (IV) discussion and conclusion.

2. Methods

2.1 NDIT Cohort

Data were from the NDIT (Nicotine Dependence in Teens) cohort, a 20-year prospective cohort initiated in 1999 (n=1294). The study's original objectives were to examine the natural course and determinants of cigarette smoking and nicotine dependence. Other factors including anthropometric measures, diet, alcohol, physical activity, and mental health were subsequently added to the data collection over time (www.nditstudy.ca). A school-based sampling strategy was used for identifying and selecting the NDIT participants. A total of 13 public high schools located in urban, suburban, and rural area of Montreal (comprising a mix of French- and English-language schools) were initially selected. Out of 13 schools, three of them were excluded because of the low return of parental consent forms and no guarantee for further follow-ups after the first year.

Data were collected every three months when participants were in high school (grade 7 to grade 11; 1999-2005, 20 cycles). After high school, follow-up assessments were conducted approximately every four years (2007-08 and 2011-12 for cycles 21 and 22, respectively). For this thesis, the exposure and outcome of interest were taken from cycle 21 and cycle 22. However, some of the demographic characteristics and covariates were taken from baseline or subsequent cycles.

2.2 Introduction to Missing Data

One of the most common issues in longitudinal studies is non-response occurrence, also known as missing data (Fiona, et al., 2006). Missing data can occur because of the long-term follow-up, as well as item non-response (Newman, 2003). For instance, some participants may not feel comfortable responding to some sensitive questions such as drug use or alcohol consumption.

Missing data can bring theoretical issues into the data analysis. For example, less available observations lead to the loss of efficiency. On the other hand, when the observed values do not represent the original full dataset correctly, a biased estimate occurs (Carpenter & Carpenter, 2012). Thus, ignoring missing data issues may lead to biased parameter estimates and loss of

efficiency (Carpenter & Carpenter, 2012). Therefore, it is important to identify and address missing data patterns for perform the valid statistical analysis.

2.2.1 Missing Data Pattern

The missing data pattern refers to the arrangement between observed and missing values in the dataset. For instance, for the dataset with two variables X and Y, there are four potential missing patterns $(\text{numbers of variables})^2$. Missing patterns are classified as univariate, unit non-response, monotone, arbitrary, general, planned, and latent variable patterns (Enders, 2010):

- Univariate: Contains a single variable with a missing value in the whole dataset. This pattern most commonly happens in experimental studies (Enders, 2010).
- Non-response: Occurs when a subset of subjects do not answer a portion of the questionnaire. In survey research, this happens when some subjects refuse to respond to some of the questions (Enders, 2010).
- Monotone: Occurs in many longitudinal studies when the subject drops out permanently from the study. This pattern reduces the mathematical complexity by eliminating the iterative estimation algorithms for both maximum likelihood and multiple imputation, which will be further explained in a later section (Enders, 2010).
- Arbitrary: Contains either the monotone or non-monotone missing patterns. It includes the subjects who quit the study permanently and those who do not answer some questions randomly.
- General: A "haphazard" pattern is seen between the variables. This can be seen as random missingness as well (Enders, 2010). The general pattern contains the missing values in any location in the dataset; subject discontinuation does not happen in this pattern.
- Planned: Based on the pre-defined design of the questionnaire. For instance, out of four sets of questions (Y1, Y2, Y3 and Y4); the researcher designs four different forms: form 1 includes Y1, but is missing one of the Y2, Y3 or Y4, and so on. In this case, a large number of questionnaires are collected.
- Latent variable: There are two sets of variables; observed ("manifested") and unobserved ("latent"). Latent variables can represent complex or abstract concepts that cannot be

directly measured easily (such as beliefs and emotions). Thus in theory, latent variables are entirely missing for all the samples. Statistical procedures are necessary (such as confirmatory factor analysis) in order to define the latent constructs among a set of variables (Y1, Y2 and Y3).

2.2.2 Missing Data Mechanisms

Whereas missing data *patterns* indicate the missing values' location, missing data *mechanisms* define the relationship between observed and missing variables. In addition, missing data mechanisms indicate the probability of the missingness and other variables in the dataset. Survey variables and observations can be shown in a matrix with the available cases as elements of the matrix. If there is missing data in the matrix, the relationship of the missing cell can be random, or it can be dependent on the other variables' value. Defining the source of the missingness is one of the basic steps of missing procedure determination. (Newman, 2003).

Rubin (1976) classified the missing data problems based on this theory. According to Rubin's missing data theory, a probability distribution exists for missing variables. In order to clarify the concept, Rubin defined **R** as a binary indicator variable. For this purpose, he assigned 1 to observed values and 0 to missing ones. **R** becomes a matrix with the same total numbers of rows and columns of raw multivariate data.

Let $\mathbf{Y} = (\mathbf{y_{ij}})$ denote an n × k dataset (n cases with k variables)

$\mathbf{y_{ij}}$= value of variable j for case i

$\mathbf{Y_{0bs}}$ = observed components of **Y**

$\mathbf{Y_{mis}}$ = missing components of **Y**

R is a **Y** dimension matrix with elements of 1 (If **Y** is observed) and 0 (If **Y** is missing) (Schafer, 1997).

$$\mathbf{R} = \begin{vmatrix} y_{11} & y_{12} & y_{13} \cdots \\ y_{21} & y_{22} & y_{23} \cdots \\ y_{31} & y_{32} & y_{33} \cdots \\ \cdots & \cdots & \cdots \cdots \end{vmatrix}$$

The $p(R|Y,\xi)$ is the distribution of **R** given **Y,** while ξ is unknown parameters. Indeed, the notation $p()$ represents the probability density function (Rubin, 1978).

Rubin's classification system includes missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Missing at Random (MAR)

In missing at random (MAR), the probability of missing data on a variable Y is not related to Y's value. Instead, it is related to another measured variable in the analytic model. Therefore, for MAR, missing values are related to other measured or observed variables Y_{obs}, but not to unobserved or missing values Y_{mis}, after controlling for Y_{obs}.

$$p(R|Y,\xi) = p(R|Y_{obs}, Y_{miss}, \xi) = p\,(R|Y_{obs}, \xi) \qquad (2.1)$$

In this case, $\boldsymbol{\xi}$ can be ignored because it is independent of the model's parameters ($\boldsymbol{\theta}$) and the missing data mechanism MAR (Rubin, 1978). Therefore, there is no association between the probability of missing values and variables that are not complete when observed values are controlled in the dataset (Jacobs, 2015). However, one of this mechanism's difficulties is that there is no way to formally examine whether the missing data probability is entirely a function of the other measured variables.

Missing Completely at Random (MCAR)

The MAR's special case is MCAR. In this mechanism, missing data are neither related to the observed data Y_{0bs}, nor Y_{mis}. Thus, the variables have an equal probability of being missing for all cases. This mechanism's formal definition is that **Y**'s missing data is not related to other measured variables nor values of Y itself (Davey & Dai, 2020; Schafer, 1997)

$$P_r\,(R|Y,\xi) = P_r\,(R|Y_{0bs}, Y_{miss}, \xi) = P\,(R|\xi) \qquad (2.2)$$

Missing not at Random (MNAR)

Missing not at random occurs when the probability of Y_{mis} depends on itself. For instance, participants with high income are less likely to answer questions about income (Enders, 2010). In this case, missingness of the income variable is related to income (Enders C. K., 2006).

$$P_r\,(R|Y,\xi) = P_r\,(R|Y_{0bs}, Y_{miss}, \xi) \qquad (2.3)$$

2.3 Methods for Handling Missing Data

Traditionally, missing data were handled by deleting incomplete cases or replacing the missing values. In the following section, an overview of these traditional missing data methods is provided.

2.3.1 Traditional Missing-Data Techniques

Traditional methods of handling missing data include: listwise deletion, pairwise deletion, regression imputation, and single imputation.

- Listwise deletion technique, also known as complete-case analysis, performs the analysis based on the fully observed cases. In this technique, all the missing values are excluded from the dataset before any analysis (Enders, 2010).
- Pairwise deletion technique, also known as available-case analysis, attempts to include the observed values as much as possible. In the pairwise deletion technique, cases are eliminated if only they are missing data on the variables in the analysis. For example, two variables, **X** and **Y**, are part of the dataset with missing values. Only one of the variables (**X)** should be included in the statistical analysis. In order to keep only the observed cases, all missing values should be deleted from **X**. In contrast, no changes to **Y** are made because it is not in any analysis. Indeed, in the pairwise deletion technique**,** deletion is based on variable selection, not the entire dataset.
- In regression imputation, missing values are replaced by the estimate of the regression model with observed values. Stochastic regression imputation solves the regression imputation problem. Customarily a random sample of the residual term from a normal distribution is added to the missing value.

- Single imputation technique replaces every single value with either the mean of the observed values or the last observation carried forward before performing any analysis (Enders, 2010).

However, these traditional methods can lead to biased estimates. Thus, other methods such as full information maximum-likelihood (FIML) and multiple imputation (MI) were developed to find other options for imputation in comparison to these traditional methods (Enders, 2010), as further outlined in the following sections.

2.3.2 Full Information Maximum-Likelihood vs. Multiple Imputation

Both full information maximum-likelihood (FIML) and multiple imputation (MI) are similar techniques because they require multivariate normality and MAR data. The differences between FIML and MI estimates depend on whether the same set of input variables were used in both imputation methods and analysis or not (e.g., auxiliary variables which is described in a later section). If FIML and MI use similar input variables, their estimates may be similar, but MI standard errors may be slightly larger (Enders, 2010). Nevertheless, the two procedures' results tend to be similar due to the assumption of large sample size (Demirtas & Schafer, 2003). In order to elaborate on the concepts, FIML techniques and multiple imputation techniques will be described in detail in the following sections.

2.3.3 Full Information Maximum-Likelihood

One of the fundamental goals of FIML is identifying the value of the population parameters. Parameters' estimates from maximum-likelihood estimation include all available information containing all cases with missing data (Hartley & Hocking, 1971). FIML is also referred to as "direct maximum likelihood", "raw maximum likelihood", or simply "maximum likelihood".

As a matter of fact, missing values are not imputed in this method. All parameter estimates are calculated based on the dataset, including missing information. Maximum likelihood calculates the estimates of the available information that contains missingness. This method does not 'impute' any missing values in the dataset. Instead, the goal of FIML is to define the parameter estimates that maximize the sample log-likelihood. The available data for each case is used to compute the log-likelihood. For instance, some variables have 400 cases but some have only 390 available cases, the model fit information is based on the 400 cases (Enders, 2010).

Log-likelihood is used for fitting the set of the values of the parameter's estimation. Additionally, the maximum-likelihood depends on the probability density function under the assumption of multivariate normal distribution (Enders, 2010). As a case in point, multivariate normal distribution for the population as an assumption is needed for the individual log-likelihood function.

The complete data log-likelihood for a single case *i* is calculated as:

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu) \qquad (2.4)$$

In this formula, Y_i is the vector of raw data for observed values for case *i*. The k_i is defined based on the total number of observed variables for case *i*, μ and Σ are the vector of mean and covariance matrix population, respectively. Estimates of μ and Σ are calculated with this algorithm to maximize the log-likelihood (Enders, 2010). Replacing the score vector and values of density function parameters gives the likelihood value. For avoiding rounding errors, the natural logarithm values can be used instead (Enders, 2010).

The main part of the formula $(Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)$ is called Mahalanobis distance value. This value is calculated from each observed value's squared z-score and the value of the center of the multivariate normal distribution (Enders, 2010). In fact, this value is very important because it specifies the value of log-likelihood (Enders, 2010).

The simple term is shown in the second equation with a specific mean vector and covariance matrix. If there are missing values in the variables, log-likelihood handles this situation by removing the correspondence parameter related to the missing value. The equation below follows the same rule for missing data.

With missing data, the log-likelihood for case i (log is the natural logarithm with base):

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}\left(Y_i - \mu_i\right)^T \Sigma_i^{-1} (Y_i - \mu_i) \qquad (2.5)$$

k_i = The number of observed variables for case i.

Y is a vector of observed variables and the Σ_i and μ_i can be different for each case by deleting the rows and columns of missing variables.

Another point to consider is that the (2.5) formula is calculated based on each individual case. This formula is identical to each individual case because the log-likelihood depends on the case i related to each variable. As a result, the sum of n log-likelihood of each case is presented as the total sample log-likelihood (Enders, 2010).

2.3.4 Multiple Imputation

Broadly, MI includes three phases (Rubin, 1978):

1. To create *m* complete or "imputed" datasets.
2. To analyze each complete dataset individually.
3. To combine all results from the analysis phase to obtain the overall parameters estimates.

These phases are described in the following sections with more details.

2.3.4.1 Imputation Phase

By generating *m* set of acceptable values for each missing data point, the uncertainty of values is reduced in *m* complete datasets from this imputation (Dong & Peng, 2013)

Based on Rubin's (1978) recommendation, the Bayesian approach should be used in creating imputation with two steps: I-step, and as well as a posterior step, which is known as P-step. This approach includes determining a model with parameters for complete data, generating a prior distribution for models with parameters, and drawing *m* times from $p(Y_{mis}| Y_{obs})$ (Schafer, 1999).

Based on the Bayesian perspective, this phase includes two steps:

I-step: Random draw of one of the parameters from the set of θ from the conditional distribution given observed data (Y_{obs}).

P-step: Random draw of missing values Y_{mis} from $p(Y_{mis}|Y_{obs}, \theta)$.

Evaluating the true expression of the posterior distribution is difficult in some studies when the missing data is multivariate, and the missing pattern is arbitrary. In these situations, because of the arbitrary missing pattern, the Markov Chain Monte Carlo method is recommended as a simulator algorithm to calculate the posterior $p(\theta| Y_{obs})$. Thus, the parameter θ is unknown, and it is based

on the observed data. It can be pointed out that random samples are drawn from the Y_{mis} from $p(Y_{mis}|Y_{obs}, \theta)$ and θ from $p(\theta|Y_{obs}, Y_{mis})$ separately. In this way, it is easier to draws samples in two separate steps in the MCMC method (Dong & Peng, 2013).

In practice, data augmentation is used for performing the MI (Tanner & Wong, 1987). In particular, the stationary distribution $p(Y_{mis}| Y_{obs})$ can produce the imputed missing values. For instance, suppose $Y = (Y_1, \ldots, Y_k)$ is a vector with k random variable, following the $P_r(Y|\theta)$.

In fact, the data augmentation algorithm depends on MCMC (Enders, 2010). Simulating random draws from posterior distribution is the main goal of the MCMC algorithm (Enders, 2010). The two equations of MCMC are listed in the subsequent section.

Specifically, in equation 1, random sample $Y_{mis}^{(j+1)}$ is drawn from the predictive distribution of missing values, conditional on the observed values $p(Y_{mis}|Y_{obs}, \theta^{(j)})$, and the parameter values iteration *j* (Dong & Peng, 2013).

$$Y_{mis}^{(j+1)} \sim p(Y_{mis}|Y_{obs}, \theta^{(j)}) \qquad (2.6)$$

Equation 2.7 is Monte Carlo simulation. In this step, new parameter $\theta^{(j+1)}$ is generated based on the conditional posterior distribution, which contains the Y_{obs} as observed data and $Y_{mis}^{(j+1)}$ imputed values from I-step. Theoretically, simulated parameters at P-step *t* are based on the preceding I-step imputed values. The imputation at I-step *j+1* is related to parameter simulation from P-step, and so on. The data augmentation algorithm behaves randomly from one cycle to the next, repeating between I-step and P-step to produce a data augmentation chain. Thus, the mutual association of I-step and P-step makes a relationship among simulated parameters from consecutive P-steps (Enders, 2010).

$$\theta^{(j+1)} \sim p(\theta|Y_{obs}, Y_{mis}^{(j+1)}) \qquad (2.7)$$

Convergence to Markov Chain is achieved after J times repeats or both equations and steps in MCMC continues until the stationary distribution coverages:

$$(Y_{mis}^{(1)}, \theta^{(1)}) , (Y_{mis}^{(2)}, \theta^{(2)}), Y_{mis}^{(3)}, \theta^{(3)}), \dots , (Y_{mis}^{(j)}, \theta^{(j)})$$

$Y_{miss}^{(j)}$: Imputed values at I-step *j*, drawn from a distribution with posterior distribution's average.

$\theta^{(j)}$: Simulated parameter values at P-step after iteration *j*

As a result, the MCMC method is recommended as a simulator to calculate the parameter estimates (Enders, 2010). The MCMC convergence is stochastic; the distribution of the imputed values converges to $p(Y_{mis} \mid Y_{obs})$ and parameter estimates to $p(\Theta|Y_{obs})$ (Enders, 2010).

2.3.4.2 The Analysis Phase

The *m* complete datasets with imputed missing values created from the imputation phase are then used in the analysis phase.

In this phase, *m* imputed datasets generated from the imputation phase are analyzed separately. At this point, analyzing the filled-in datasets is the goal of this phase. For instance, when we need to perform a regression analysis in this phase for completed data, the regression should be repeated *m* times, once for each dataset (Enders, 2010). The output of this step is *m* sets of parameter estimation and standard errors.

2.3.4.3 The Pooling Phase

Finally, the *m* estimates are pooled together to provide the single estimation of the parameters and their standard errors (Dong & Peng, 2013).

Multiple imputation minimizes the bias of the single imputation standard errors of the parameters. In this phase, a single-point estimate is calculated from the m parameter. Rubin (1987) defined the multiple imputation point estimate as:

$$\bar{\theta} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}_t \quad (2.8)$$

$\hat{\theta}_t$: Estimation of the parameter from dataset t

$\bar{\theta}$: Pooled estimate

This equation is the usual formula of the sample mean where the parameters present as a data point. Rubin (1987) developed multiple imputation in the Bayesian scheme; the pooled point estimate is of the parameter θ that is obtained from data with no missingness. Whereas $\bar{\theta}$ is defined as the mean of observed-data posterior distribution by the Bayesian paradigm, the fixed population parameter point estimate is $\bar{\theta}$ in a frequentist standpoint (Little & Rubin, 2002; Rubin, 1978).

The variance of $\bar{\theta}$ contains two-parts, which includes between and within imputation variance. Additionally, this variance is calculated based on the formula presented below (Rubin, 1978):

$$\bar{u} = \frac{1}{m} \sum_{i=1}^{m} \hat{u}_i \quad (2.9)$$

$$B = \frac{1}{m-1} \sum_{i=1}^{m} \left(\hat{\theta}_i - \bar{\theta}\right)^2 \quad (2.10)$$

$$T = \bar{u} + \left(1 + \frac{1}{m}\right) B = \text{the variance of } \bar{\theta} \quad (2.11)$$

$\bar{u}$: Within imputation variance

B: Between imputation variances

m: A finite number of imputations

To test the null hypothesis as: $\bar{\theta} = \theta_0$, the following statistics can be used: $(\theta - \bar{\theta}\ )\ /\ (\sqrt{T})$, which is following t distribution with degree of freedom of (ν_m or ν^*_m) (Barnard J, 1999). F-distribution can be used with 1 and ν degree of freedom as the numerator and denominator, respectively.

$$F_{1,\nu} = \frac{(\bar{\theta} - \theta_0)^2}{T} \quad (2.12)$$

$$r = \frac{\left(1 + \frac{1}{m}\right) B}{\bar{u}} \quad (2.13)$$

$$\nu_m = (m-1)[1 + \tfrac{1}{r}]^2 \quad (2.14)$$

$$\nu^*_m = [1/\nu m + 1/(((1 - \text{gamma})\nu 0\ (\nu 0 + 1))/(\nu 0 + 3))]^{-1}$$

In this formula, the r is the relative increase in variance because of missing data where:
gamma: $(1+\frac{1}{m})\ B/T$

$\nu 0$: the degree of the of freedom for complete data
ν^*_m: correction of ν_m if $\nu 0$ is small with moderate missing rate.

The fraction of B in T is shown with $\hat{\lambda}$. The $\hat{\lambda}$ represents the severity of missing data. respectively.

$$\hat{\lambda} = \frac{\left(1+\frac{1}{m}\right)B + \frac{2}{\nu_m + 3}}{T} \quad (2.15)$$

Following the formula, $\hat{\lambda}$ depends on the correlation between the variables and the missing data rate. When there is an association, this value is less than the missingness rate (Enders, 2010). It is assumed that $\hat{\lambda}$ is the same for all the variables (Rubin, 1978).

Thus, the two sources of uncertainty (between: due to missing data; and within: due to the estimations in the imputation) are combined into the parameter estimation's pooled standard error (Dong & Peng, 2013). Therefore, the pooled standard error is bigger than the derived standard error of each individual imputed dataset.

2.3.4.4 Number of Imputations

$\hat{\lambda}$ The fraction of missing information

$$\mathrm{RE} = (1 + \frac{\hat{\lambda}}{\mathrm{m}})^{-1} \tag{2.16}$$

For instance, if $\hat{\lambda}$= 0.2 (e.g., 20% missing data rate) with m=5 the efficiency rate is $\sqrt{1 + {}^{0.2}/_{5}}$ =1.02. This value shows that with five imputed datasets, the standard error is 1.02 times larger than the standard error of an estimate with infinity many imputations. Some simulation studies showed that at least 25 imputations are recommended to decrease random effect sampling from multiple imputation (Enders, 2010). It is also recommended that the *m* should be at least equal to the percentage of incomplete ceases (Enders, 2010).

2.3.5 Additional Technical Issues

A more extensive set of variables can be used in the imputation phase than is required for the substantive analysis model (Schafer, 1999). It is recommended to include a large set of variables in the imputation phase, but the total variables should not be more than the total number of observations (Rubin, 1978). In general, any variable that predicts the probability of incomplete variables should be included in the imputation model.

Particularly, three kinds of variables are recommended to be included in the imputation model (Schafer, 1997):

1. Variables of analytical interest.
2. Variables that are related to a missing mechanism based on the MAR assumptions.
3. Variables that are correlated to the variables containing missing values.

The second and third categories of variables are known as auxiliary variables (Collins, et al., 2001). Auxiliary variables are highly correlated with incomplete variables (r > |0.40|) and are thus recommended to be included in the MI model (Enders, 2010).

2.3.5.1 Auxiliary Variables

Due to the correlation with observed or missing variables, auxiliary variables are included in the MI model. It is not necessary to include them in the final analysis. These variables could be part of the imputation step. For instance, age is an important auxiliary variable because it influences missingness in health and social science research (Enders, 2010). In the MAR mechanism, adding auxiliary variables helps to reduce bias in the analysis. Consequently, it presents more precise parameter estimates. Thus, including auxiliary variables in missing procedures is beneficial (Collins, et al., 2001).

In general, auxiliary variables do not include any missing values (Enders, 2010). Based on the Monte Carlo simulation by Enders (2008), when auxiliary variable includes missing values, bias contraction is identical for both MAR and MCAR (Enders, 2008). Nevertheless, including auxiliary variables in the model may reduce the parameters' bias. The reduction in the bias is related to the correlation between auxiliary variables and the probability of missingness; variables that are more strongly correlated with missingness will have a larger impact on this reduction than weakly correlated auxiliary variables. As a result, this improves the quality of the imputations (Graham & Hofer, 2000).

2.3.6 Rounding Binary Variables

In comparison to the original variables' possible observed values, imputing discrete variables (such as binary or categorical) may produce implausible values. For instance, in this dataset, "sex" is defined as 0 and 1. After performing the MI, the value 0.657 is replaced for a missing data point, which is neither 0 nor 1. In this case, rounding is helpful to categorize the imputed values to match with observed values. Rounding can be an option to keep the consistency between the imputed values and original values. There are many different methods and strategies to follow for rounding the values after imputation, as described below.

2.3.6.1 Simple Rounding

One of the common rounding methods is called simple rounding or naïve rounding. It works based on fixed 0.5 thresholds. If the value is greater than 0.5, it is rounded to one. On the other hand, if the value is less than 0.5, it is rounded to zero (Enders, 2010). However, this method is flawed because a biased parameter estimate is produced and is not generally recommended.

2.3.6.2 Adaptive Rounding

Adaptive rounding was introduced by Barnard (1999). This method follows the normal approximation to a binomial distribution to calculate the threshold for imputing each missing data point. In contrast to the simple rounding method and fixed threshold, adaptive rounding is applied to the rounding threshold (Jacobs, 2015).

If $\hat{\mu}_{UR}$ is the mean of unrounded imputed binary variables, then the rounding threshold is given by c:

$$c = \hat{\mu}_{UR} - \phi^{-1}(\hat{\mu}_{UR})\sqrt{\hat{\mu}_{UR}(1-\hat{\mu}_{UR})} \qquad (2.17)$$

ϕ^{-1} is the inverse of standard normal cumulative distribution.

The rounding threshold is unique, and it is calculated for each imputed dataset individually. If the imputed values are greater than the calculated threshold, they are rounded to one, and the rest of the values are rounded to zero. For instance, if an unrounded imputed binary variable was $\hat{\mu}_{UR}$= 0.67, then $\phi^{-1}(0.67)$ is equal to 0.44. By replacing the calculated values in the formula, the threshold becomes 0.463. Imputed values that exceed this calculated threshold are rounded up to one, and values below this value are rounded down to zero.

2.3.6.3 Calibration

The calibration rounding method (Bernaards, et al., 2006) determines the rounding threshold using a subset of the imputed values reproducing the frequency of zeros and ones from raw data. The first step of calibration is to create one copy of the raw datasets. It is necessary to delete the

observed values from the variables that contain the missing values. In other words, no observed values for binary variables remain in the second dataset. The second step is to append the original data and new dataset in one stacked file. After imputing the entire stacked dataset, the rounding threshold can be identified ((Enders, 2010). By finding the threshold, simple rounding is applied for producing binary values (Enders, 2010). Some research indicates that even when data are missing at random, biased estimates can occur in the calibration method. However, this has not been found to be the case in simulation studies of adaptive rounding (Enders, 2010).

In one study, three rounding methods for binary missing values were compared to one another (Bernaards, et al., 2007): (1) simple rounding, (2) randomly replacing the imputed values by Bernoulli trial as zero and one, and (3) adaptive rounding. Their results demonstrated that the random replacement performed the worst. In contrast, the adaptive rounding performed the best. There is no specific research comparing the advantages of calibration and adaptive rounding (Bernaards, et al., 2007). Nevertheless, based on the recommendations in the literature, adaptive rounding was applied for rounding the binary variables in this thesis.

2.4 Exposure

Alcohol consumption was based on self-reported measures in the questionnaire. Alcohol intake in the previous 12 months was measured in both cycles 21 and 22. The questionnaire had five response options regarding alcohol consumption: never, less than once a month, 1-3 times per month, 1-6 times per week, and every day. For the purpose of this thesis, different categories of alcohol consumption were compared by nested model comparison to identify this variable's optimal categorization for the statistical analysis.

2.5 Outcome

Weight and height measurements were self-reported in cycles 21 and 22. Body Mass Index (BMI) was calculated as weight in kilogram divided by height in meters ($\mathrm{kg/m^2}$).

$$\text{BMI}=\frac{\text{weight (kg)}}{(\text{height(m)})^2} \tag{2.18}$$

BMI was used both as a continuous variable and a categorical variable (weight status) in the statistical analysis. More specifically, weight status was defined as normal (BMI<25.0 $\mathrm{kg/m^2}$) and overweight with a BMI greater than or equal to 25.0 $\mathrm{kg/m^2}$.

2.6 Covariates

Potential covariates were selected based on their associations with alcohol consumption or BMI with $\mathcal{X}^2$, t-test, or Wilcoxon rank-sum tests. Smoking status, whether the respondent had depression, relationship status, current employment status, whether the respondent was currently a student, physical activity, and demographic characteristics (baseline age, sex, and race) were considered as covariates in the analysis as outlined further in the following section.

2.6.1 Smoking

Smoking was measured in cycles 21 and 22. The questionnaire's original variable consisted of five response options, including never smoked, once or a couple of times in the past 12 months, once

or a couple of times each month, once or a couple of times each week, and every day. For the purpose of this thesis, this variable was categorized as a binary variable (smoker vs. non-smoker) for the linear model and as a categorical variable (never, former, occasional, and daily) in the logistic model.

2.6.2 Depression

The presence of depressive symptoms was assessed using the Major Depression Inventory. The scores can range from 0 to 50. In accordance with the literature, the presence of depression was classified as a binary variable: no depression for scores between 0 to 20, and possible depression for scores greater than 20 (Timmerby, et al., 2015).

2.6.3 Relationship Status

Relationship status was defined as: single, married, living as married (common-law), divorced, separated, and other. For this thesis, relationship status was re-classified into whether the person was in a relationship (married or common-law), compared to being single, divorced, or separated.

2.6.4 Employment Status

Employment status was defined as working either full-time or not currently employed. Employment status was a binary variable as being either a current worker or not currently working.

2.6.5 Student Status

Student status was defined as whether the participant was a full-time student, part-time student, or not currently a student. Three categories were re-classified in two categories as a current student or not a current student.

2.6.6 Physical Activity

Physical activity was self-reported in each cycle using a validated questionnaire (Sallis, et al., 1993). Participants indicated whether they had performed an activity (such as basketball, ice hockey, volleyball from a list) for 5 minutes or more in the past week. Each of the activities was

further classified as light, moderate or vigorous physical activity using the 2018 Youth compendium based on its Metabolic Equivalent of Task (MET). For instance, 1.5 to less than 3 METs is considered light, 3.0 to less than 6 METs is considered as moderate, and more than 6 METs is considered as vigorous physical activity (Wellman, et al., 2020). These values were then converted to the total number of light, moderate, and vigorous physical activity based on the data collection from cycles 1 to 20.

Thus, each cycle contained three physical activity variables: number of activities as light, moderate, and vigorous physical activity. A new variable was created based on the mean number of each of these physical activity variables. For instance, the new light physical activity variable was calculated as the mean of all available light variables (light physical activity) from cycles 1 to 20. The two other categories (moderate and vigorous) were calculated the same way. As a result, three new variables containing the mean number of light, moderate, and vigorous physical activities across cycles 1 to 20, were included in the MI model.

Lastly, the physical activity variable was defined as a new variable: the sum of the two new calculated moderate and vigorous variables (Mean of cycle 1 to 20). This variable was included as a continuous covariate in statistical models for both cycles 21 and 22.

2.6.7 Demographic Characteristics

Demographic characteristics included age at baseline, sex, and race. The questionnaire's original variable for race consisted of nine levels (Arabic, Black, Chinese, Latin American, Southeast Asian, West Asian, White, and Other). The race was re-classified as "white" versus "non-white

2.7 Statistical Analysis

The main predictor of interest (alcohol consumption) was categorized in the questionnaire as never, less than once a month, 1-3 times per month, 1-6 times per week, and every day. Prior to conducting the primary statistical tests, we investigated whether there was an optimal classification of alcohol consumption with non-nested model comparisons.

SAS (version 9.4) was used for all analyses. Generalized linear models were conducted to assess the relationship between alcohol consumption with BMI and alcohol consumption with weight status. Analyses were conducted separately for a complete case dataset and the imputed dataset. Complete case analysis involved all participants with observed values for the variables of interest in the model. For the MI models, following the arbitrary missing pattern, the MCMC method was performed as a first step. In the next step, MI analyzed the 50 sets of data using statistical procedures such as logistic regression and GLM separately. At the end of this step, 50 sets of parameter estimates were obtained. In the last step, MI pooled the 50 estimates. For the imputed datasets, auxiliary variables included age at baseline, sex, race, and physical activities. In addition, all analytical variables that were used in statistical analysis were included in the MI procedure. The MI method was implemented with PROC MI and MCMC. Both PROC GLM and PROC LOGISTIC were used for the regression models; estimates were pooled with PROC MIANALIZE. All the statistical analyses (including GLM and logistic regression) were performed separately for the imputed datasets and the complete dataset.

All regression models included age at baseline, race, sex, smoking status, depression, relationship status, employee status, student status, and physical activity as covariates. Additionally, based on the statistical associations between sex and student status with exposure and outcome, interactions were also incorporated into some regression models.

2.7.1 Non-nested Model Comparisons

Alcohol consumption was measured in the questionnaire with five levels as never, less than once a month, 1-3 times per month, 1-6 times per week, and every day. This variable was re-classified in three different ways and all those three different categorizations were analyzed with non-nested model comparisons. Model comparisons were assessed with Akaike Information Criterion (Hens & Aerts, 2006). Smaller values indicate better model fit.

The following table shows the alcohol consumption categories for each model:

Models	Never	Less than once a month	At least once a month	Every day
Model 1	X		X	X
Model 2	X	X	X	
Model 3	X		X	

2.7.2 General Linear Model (GLM)

Many response modeling approaches commonly utilize generalized linear models (Wright, 2001). Generalized linear models include General Linear Model (GLM), and logistic regression (among others) and are used according to the nature of the response variable.

X is the vector of the parameters, and x_{ij} is the jth covariates' value for observation i (Anderson, et al., 2005).

X is a matrix of the independent variables

- Each column is a variable
- Each row is an observation

β is a vector of parameter coefficients

ε is a vector of residuals

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i \quad (2.19)$$

$$y_i \sim N(\mu_{i,}\varepsilon)$$

$$E(Y_i) = \mu_i = \sum_{i=0}^{n} 1x_{ij}\beta_j \qquad i = 1, \ldots, n$$

2.7.3 Logistic Regression

When the dependent variable is binary, it should be modeled with logistic regression. In this thesis, logistic regression was used to define the categorical probability of having overweight or obesity. Another point to consider is that the odds ratio in logistic regression shows the constant effect of predictor variables. Odds are the ratio between the probabilities; it contains the probability of having overweight/obesity over the probability of not having overweight/obesity; this value is between zero and infinity. Due to the ratio, the chance of the outcome builds on each characteristic. Subsequently, the model is based on the logarithm of:

$$\log\left[\pi/_{(a-\pi)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \quad (2.20)$$

π	Probability of participants' having overweight or obesity
x_j	Exploratory variables
β_0	The reference group
β_i	Regression coefficients associated with the reference group and the x_j explanatory variables

2.7.4 GLM

The GLM was fitted (separately for cycle 21 and cycle 22) to assess the association between each category of alcohol consumption and the continuous variable of BMI. Covariates included age,

gender, race, employee status, relationship status, student status, depression, physical activities, the interaction of gender and alcohol consumption, and the interaction of student status and alcohol consumption. Interactions were added to individual models separately.

y_i	BMI
β_0	Intercept
x_j	Independent variables as alcohol consumption, age at baseline, gender, smoking, relationship status, physical activities, depression, student status, and employee status
β_i	Coefficient of the linear combination
ε_i	Errors independent and identically distributed with $\varepsilon_i \sim N(0, \sigma^2)$

- Model without interaction:

$$\begin{aligned}\mathrm{BMI_i} = {} & \beta_0 + \beta_1 \mathrm{Age} + \beta_2 \mathrm{Alcohol} + \beta_3 \mathrm{Sex} + \beta_4 \mathrm{Race} + \beta_5 \mathrm{Employment\ Status} \\ & + \beta_6 \mathrm{Relationship\ Status} + \beta_7 \mathrm{Student\ Status} + \beta_8 \mathrm{Depression} + \beta_9 \mathrm{Smoking} \\ & + \beta_{10} \mathrm{Physical\ Activity} + \varepsilon_i \\ & \mathrm{i} = (\mathrm{cycle}\ 21\ , 22)\end{aligned}$$

- Model with interaction:

$$\begin{aligned}\mathrm{BMI_i} = {} & \beta_0 + \beta_1 \mathrm{Age} + \beta_2 \mathrm{Alcohol} + \beta_3 \mathrm{Sex} + \beta_4 \mathrm{Race} + \beta_5 \mathrm{Employment\ Status} \\ & + \beta_6 \mathrm{Relationship\ Status} + \beta_7 \mathrm{Student\ Status} + \beta_8 \mathrm{Depression} + \beta_{10} \mathrm{Smoking} \\ & + \beta_{11} \mathrm{Physical\ Activity} + \beta_{12} \mathrm{Student\ Staus} * \mathrm{Alcohol} + \varepsilon_i \\ & \mathrm{i} = (\mathrm{cycle}\ 21\ , 22)\end{aligned}$$

$$\begin{aligned}\mathrm{BMI_i} = {} & \beta_0 + \beta_1 \mathrm{Age} + \beta_2 \mathrm{Alcohol} + \beta_3 \mathrm{Sex} + \beta_4 \mathrm{Race} + \beta_5 \mathrm{Employment\ Status} \\ & + \beta_6 \mathrm{Relationship\ Status} + \beta_7 \mathrm{Student\ Status} + \beta_8 \mathrm{Depression} + \beta_{10} \mathrm{Smoking} \\ & + \beta_{11} \mathrm{Physical\ Activity} + \beta_{12} \mathrm{Sex} * \mathrm{Alcohol} + \varepsilon_i \\ & \mathrm{i} = (\mathrm{cycle}\ 21\ , 22)\end{aligned}$$

2.7.5 Logistic Regression Application

Logistic regression was performed as:

$$\text{Overweight} = \begin{cases} 0, & \text{BMI} \leq 25 \\ 1, & \text{BMI} > 25 \end{cases}$$

$$\log\left[\pi/(a-\pi)\right] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Alcohol} + \beta_3 \text{Sex} + \beta_4 \text{Race} + \beta_5 \text{Employment Status} + \beta_6 \text{Relationship Status} + \beta_7 \text{Student Status} + \beta_8 \text{Depression} + \beta_9 \text{Smoking} + \beta_{10} \text{Physical Activity}$$

3. Results

This chapter begins with the descriptive statistics of the total participants and explores the potential interactions with sex. In the subsequent section, the optimal category of alcohol consumption is selected based on non-nested model comparisons. At the end of this chapter, the association between alcohol consumption and BMI is analyzed with GLM and logistic regression.

3.1 Sample Characteristics

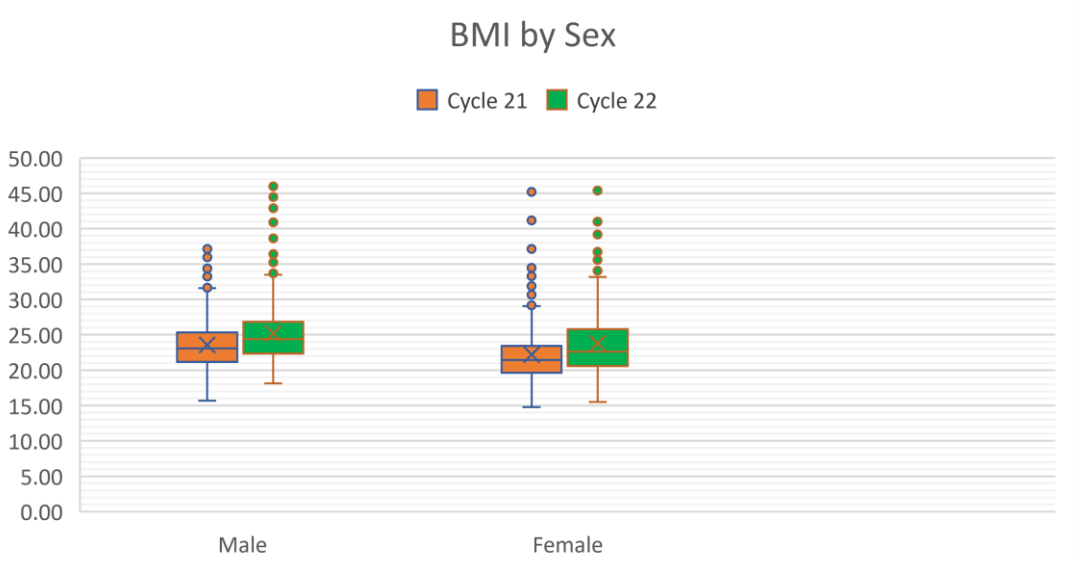
There were 1293 participants included in this study. The general baseline characteristics of the participants are presented in Table 1. The data included 670 females (51.82%) and 623 males (48.18%).

From table 1, sex differences can be observed. Men and women significantly differ in age at baseline (12.8 vs 12.7, p= 0.043), BMI at cycle 21 (23.57 vs 22.21, p= 0.001), and cycle 22 (25.18 vs 23.78, p= 0.001). Figures 1 and 2 show this difference visually. No sex differences in language or race were detected. Approximately 92% were born in Canada (n=1191), and 389 (30.09%) were from French-speaking high schools. Over half of the participants were white. Figure 1 shows the BMI in cycle 21 and 22 from observed values, for males and females.

Table 1 Demographic Characteristics by Sex at Baseline

	Male (n=623)	**Female (n=670)**	***p***
Age, Mean (SD)	12.8(0.56)	12.7 (0.55)	0.043
French Speaking, n (%)	176(28.3)	213(31.8)	0.165
Caucasian (White), n (%)	320(79.6)	374(78.7)	0.753

Figure 1 Comparison of BMI by Sex



3.2 Baseline Characteristics of Participants (Complete vs. Not-Complete Cases)

As described earlier in the method section, associations between baseline demographic characteristics and non-response were assessed (Table 2). In cycles 21 and 22, statistically significant differences in the proportion of Canadian born, and race between complete and incomplete cases were detected. Missing data patterns including all main variables listed in statistical models, are shown in Figure 2. The pattern suggests arbitrary missingness.

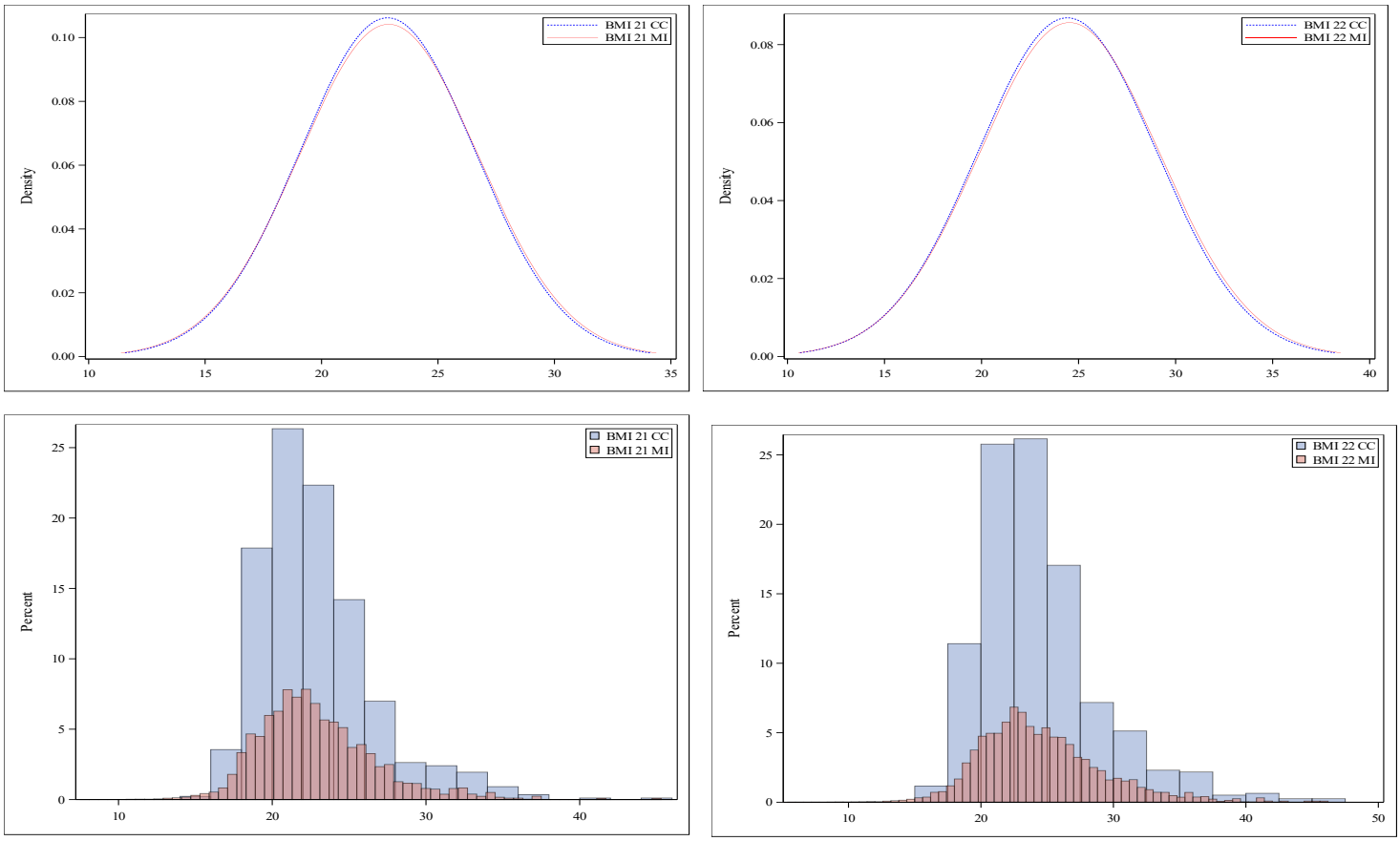
Table 2 Baseline Characteristics for Cycle 21 and 22 for Completed Cases vs. not Completed Cases

	Cycle 21			Cycle 22		
	Complete (n=594)	**Not Complete (n=700)**	***p***	**Complete (n=769)**	**Not Complete (n=525)**	***p***
Age, Mean (SD)	12.7 (0.53)	12.8 (0.58)	0.054	12.7 (0.47)	12.8 (0.65)	**< 0.001**
Born in Canada, n (%)	562 (94.6%)	629 (89.9%)	**0.002**	717 (93.2%)	474 (90.3%)	0.069
French Speakers, n (%)	210 (35.4%)	179 (25.6%)	**< 0.001**	237 (30.8%)	152 (29.0%)	0.486
Caucasian (White), n (%)	486 (81.8%)	208 (29.7%)	**0.003**	562 (73.1%)	132 (25.1%)	0.636

Figure 2 Missing data Pattern

After 50 imputations, the distribution of the imputed BMI variable was plotted and compared with the observed and completed data in the following Figure (blue: before MI, red: after MI; left panels: cycle 21, right panels: cycle 22). No significant difference of density between imputed and observed BMI were detected. Figure shows that the imputed values in both cycle 21 and 22 follows the same density after imputation.

Figure 3 Comparison of BMI (Observed vs. Imputed)



The association between covariates and sex is shown in table 3. There were no sex differences in alcohol consumption between females and males in cycle 21. However, there was a significant association between alcohol consumption and sex in cycle 22 (p=0.016). Alcohol consumption 'never or less than once a month' was 18.1% and 11.4%, and 'at least once a month' was 50.2% and 51.2% in females and males, respectively. No association was detected between sex and student status or physical activity in cycle 21, while a significant association between sex and student status was detected in cycle 22. Depression, being in a relationship, and smoking were significantly associated with sex in both cycles.

Table 3 Exposure and Covariates by Sex

	Cycle 21			Cycle 22		
	Male (n=623)	**Female (n=670)**	***p***	**Male (n=623)**	**Female (n=670)**	***p***
BMI, Mean (SD)	23.57(3.57)	22.21(3.80)	0.001	25.18(4.45)	23.78(4.6)	0.001
Alcohol Consumption						
Never/<1 a month	96 (15.4%)	117 (17.5%)	0.428	71 (11.4%)	121 (18.1%)	0.016
At least once a month	307 (49.3%)	358 (53.4%)		313 (50.2%)	343 (51.2%)	
Depressed	382 (61.3%)	411 (61.3%)	<0.001	363 (58.3%)	414 (61.8%)	0.002
In a relationship	34 (5.5%)	71 (10.6%)	0.003	57 (9.1%)	99 (14.8%)	0.017
Smoking status						
Never	152 (24.4%)	125 (18.7%)		111 (17.8%)	101 (15.1%)	
Former	76 (12.2%)	111 (16.6%)	0.002	95 (15.2%)	152 (22.7%)	0.03
Current	90 (14.4%)	135 (20.1%)		98 (15.7%)	117 (17.5%)	
Every day	85 (13.6%)	105 (15.7%)		81 (13.0%)	98 (14.6%)	
Student	260 (41.7%)	336 (50.1%)	0.055	149 (23.9%)	218 (32.5%)	0.018
Employment	302 (48.5%)	378 (56.4%)	0.114	306 (49.1%)	373 (55.7%)	0.82

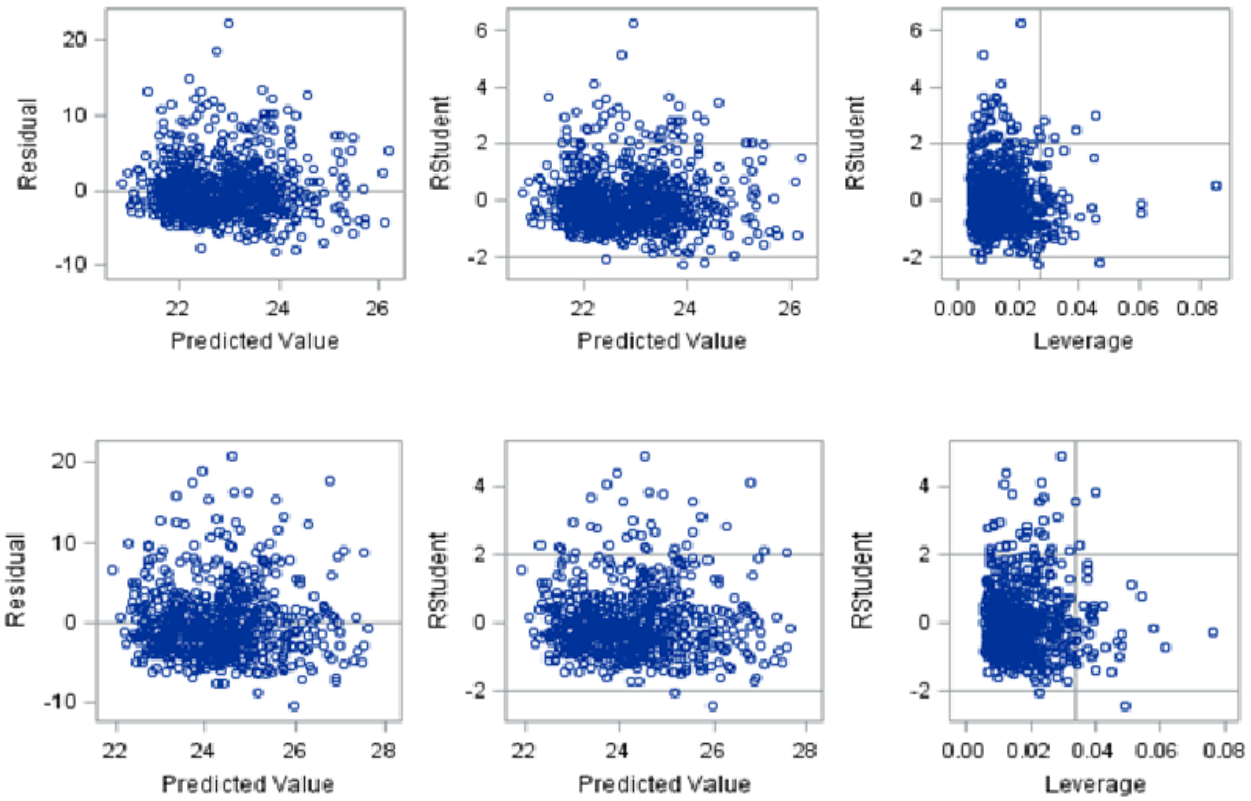
All values except BMI are reported as "n (%)"

Missing percentage and numbers are not presented in the table and all the percentages are based on the total number of each column. Available numbers for alcohol consumption: n=878 (Cycle 21), n=848 (Cycle 22); Depression: n=793 (Cycle 21), n=777 (Cycle 22); In a relationship: n=105 (Cycle 21), n=156 (Cycle 22); Smoking Status: n=879 (Cycle 21), n=853 (Cycle 22); Student Status: n=596 (Cycle 21), n=522 (Cycle 22); Employment Status: n=680 (Cycle 21), n=679 (Cycle 22).

3.3 Assessing Model Assumption

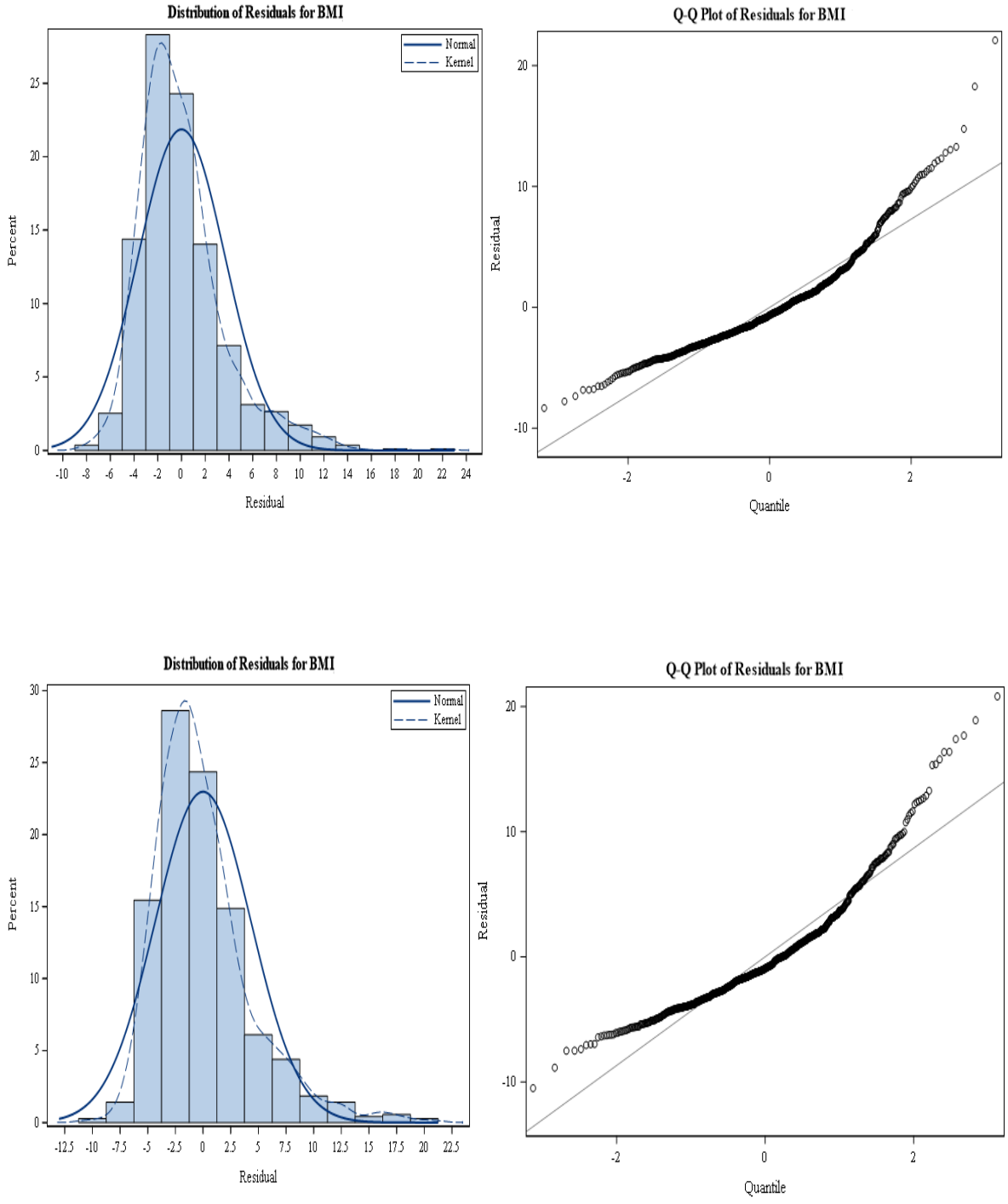
The model assumptions (residual pattern, homogeneity of variance and linearity) were checked for both cycle 21 and cycle 22 (Figure 4). The six panels below show no specific patterns at cycle 21 (top row) and cycle 22 (bottom row), suggesting that errors were independent and random.

Figure 4 Checking Model Assumption: Residual Diagnostics



The assumption of residuals normality (Figure 5) appeared to be violated for cycle 21 (top row) and cycle 22 (bottom row). The assumption was met with a log transformation on the outcome (data not shown). Then the GLM was performed before and after the transformation. No significant differences were seen in the GLM results. Thus the non-transformed data are exclusively presented for ease of interpretations of parameter estimates.

Figure 5 Checking Model Assumption: Normality of Residuals



3.4 Non-Nested Model Comparisons

Model comparisons were performed for three non-nested models with three different alcohol consumption categories. AIC was assessed for these comparisons, with smaller values indicating better model fit. Table 4 shows that the smaller AIC for cycle 21 was model 2, and the smaller AIC for cycle 22 was model 3. These models were carried forward in the statistical analysis accordingly.

Table 4 Non-nested Model Comparisons

Model	Alcohol categories	Criterion	Model Fit Statistics	
			Intercept and Covariates Cycle 21	**Intercept and Covariates Cycle 22**
Model 3	At least once a month vs. Never/Less than once a month	AIC	891.987	**885.408**
		SC	953.963	944.683
		-2 Log L	865.987	859.408
Model 2	Less than once a month vs. at least once a month vs. Never	AIC	**891.708**	885.577
		SC	958.451	949.412
		-2 Log L	863.708	857.577
Model 1	Less than once a month/Never vs. at least once a month vs. Every day	AIC	893.981	887.242
		SC	960.723	951.076
		-2 Log L	865.981	859.242

3.5 General Linear Models

GLM was applied to investigate the association between alcohol consumption and BMI for both complete and imputed datasets in cycles 21 and 22 separately. These models initially included interaction terms between being a student with alcohol consumption or sex with alcohol consumption. As the interaction terms (sex*alcohol consumption; student*alcohol consumption) were not statistically significant in the models (Supplementary table I and II), GLM was conducted after removing all interaction terms (Table 5).

The result in table 5 shows that sex was a significant predictor for BMI. In addition, a significant difference between the participants who drank alcohol at least once a month for model 2_{CC} and model 2_{MI} compared to those who did not drink alcohol in the past year was detected. Those who consumed alcohol at least once a month had significantly lower BMI (model 2_{CC}: B=-1.6, p=<.0001 and model 2_{MI}: B=-1.26, p=0.0006).

The GLM models using complete-case analysis and MI methods are compared in table 5. For both GLM models, in cycles 21 and 22, the MI results do not largely differ from the complete case models. However, estimates in the MI model are smaller for all variables except alcohol consumption and student status for cycle 21; and student status and physical activity in cycle 22. In addition, the standard errors in the MI models are smaller than complete-case model for cycle 21 except alcohol consumption and the depression. Likewise, standard errors are smaller than the complete-case model for all variables in cycle 22, except physical activity.

Table 5 GLM results: Betas (SE) on the Association between Alcohol Consumption with BMI

	Cycle 21				Cycle 22			
Covariates	**Model 1_{cc}**	***p***	**Model 1_{MI}**	***p***	**Model 2_{cc}**	***p***	**Model 2_{MI}**	***p***
Age at Baseline	0.41(0.26)	0.11	0.22(0.22)	0.32	0.26(0.35)	0.45	0.21(0.27)	0.44
Alcohol[a] (less than once a month)	-0.35(0.45)	0.43	0.11(0.52)	0.84				
Alcohol (at least once a month)	0.23(0.53)	0.67	-0.32(0.44)	0.46	-1.6(0.41)	**<.0001**	-1.26(0.37)	**0.0006**
Sex (male vs female)	-1.15(0.26)	**<.0001**	-0.88(0.26)	**0.0007**	-1.39(0.35)	**<.0001**	-0.99(0.32)	**0.0022**
Race (white vs non-white)	0.22(0.32)	0.48	0.04(0.26)	0.87	0.3(0.42)	0.48	0.05(0.31)	0.86
Smoking (smokers vs. non-smokers)	-0.19(0.12)	0.11	-0.17(0.12)	0.17	-0.2(0.16)	0.2	-0.11(0.14)	0.44
Employment (working vs. not working)	0.08(0.3)	0.8	0.07(0.3)	0.82	0.33(0.42)	0.43	0.18(0.41)	0.66
Student Status (student vs non-student)	-0.89(0.29)	**0.002**	-0.93(0.29)	**0.002**	-0.48(0.35)	0.17	-0.51(0.34)	0.14
In relationship (yes vs. no)	0.43(0.39)	0.27	0.39(0.39)	0.32	-0.81(0.44)	0.07	-0.74(0.42)	0.08
Depression (depressed vs. not depressed)	0.26(0.43)	0.55	0.1(0.44)	0.82	-0.2(0.62)	0.75	0.01(0.55)	0.98
Physical Activity	0.02(0.01)	0.1	0.02(0.01)	0.06	0.01(0.02)	0.44	0.02(0.02)	0.16

[a] Never drinkers' category is selected as the reference group for cycle 21 and 22, respectively.

3.6 Logistic Regression

A logistic regression was used to investigate the relationship between BMI and alcohol consumption. The logit linearity assumption was not violated. Model $\mathbf{3}_{\boldsymbol{CC}}$ and $\mathbf{4}_{\boldsymbol{CC}}$ were conducted in complete cases; Model $\mathbf{3}_{\boldsymbol{MI}}$ and $\mathbf{4}_{\boldsymbol{MI}}$ were MI.

Model		Complete case analysis	Imputed dataset
Cycle 21	Model 3_{CC}	X	
	Model 3_{MI}		X
Cycle 22	Model 4_{CC}	X	
	Model 4_{MI}		X

From table 6, sex and being overweight is statistically associated in both models 3_{CC} and 3_{MI}. The odds ratio (OR) is 0.52 and 1.60, respectively, meaning that holding all other variables in the model constant, the odds of males having overweight or obesity is 48% less than females in model 3_{CC} , but inversely the odds of overweight is 60% higher in model 3_{MI}. Alcohol consumption showed no impact on having overweight or obesity in cycle 21 for both 3_{CC} and 3_{MI}. In model 3_{MI}, risk of having overweight or obesity was significantly associated with student status and physical activities. Similarly, participants who were students during cycle 21 have a 60% higher risk of having overweight or obesity compared to non-students.

Table 6 shows that gender was associated with weight status in cycle 22 in both 4_{CC} and 4_{MI} models, odds of being overweight or obese in males were 60% more than females in model 4_{CC} and 51% less in model 4_{MI}. Alcohol consumption and overweight were statistically significant in cycle 22 for both 4_{CC} and 4_{MI} models, the participants who drank at least once a month compared to those who drank less than once a month were less than half as likely to have overweight or obesity at cycle 22 in model 4_{CC} and 36% less for the same group in model 4_{MI}. Odds of having overweight or obesity in occasional smokers was 53% more than never smokers and the participants in a relationship were 44% less likely to have overweight or obesity compared to those who were not in a relationship in model 4_{CC}.

Table 6 Logistic Regression results: Odds Ratios on the Association between Alcohol Consumption and the Risk of Living with Overweight or Obesity

Parameters	Model 3_{CC} OR	*p*	Model 3_{MI} OR	*p*	Model 4_{CC} OR	*p*	Model 4_{MI} OR	*p*
Age at Baseline	0.99	0.94	0.99	0.97	1.2	0.3	1.13	0.3
Alcohol[a] (less than once a month)	0.97	0.94	0.98	0.96	0.54	**0.002**	0.64	**<.0001**
Alcohol (at least once a month)	0.72	0.25	0.81	0.43				
Sex (male vs female)	0.52	**<0.0001**	1.6	**<0.0001**	0.49	<.0001	1.6	<.0001
Race (white vs non-white)	0.91	0.66	1.41	**0.03**	1.06	0.79	1.27	0.79
Smoking[b] (daily vs. non-smokers)	0.55	**0.03**	0.65	0.1	0.93	0.76	0.75	0.76
Smoking (former vs. non-smokers)	0.8	0.36	0.83	0.38	0.89	0.61	0.74	0.61
Smoking (Smoker vs. non-smokers)	0.97	0.88	1	0.99	1.53	0.05	0.82	0.054
Employment (working vs. not working)	1.06	0.77	1.01	0.95	1.37	0.14	1.17	0.14
Student Status (student vs non-student)	1.65	**0.01**	1.65	**0.004**	1.1	0.6	1.19	0.6
In relationship (yes vs. no)	1.08	0.79	1.1	0.72	0.66	0.05	0.71	0.055
Depression (depressed vs. not depressed)	1.23	0.49	1.05	0.85	1.01	0.98	1	0.98
Physical Activity	1.02	**0.03**	1.01	**0.04**	1	0.75	1.01	0.75

[a] Never drinkers' category is selected as reference group in cycle 21 and never and less than once a month drinkers' category as reference group in cycle 22.
[b] Never smokers are selected as reference group.

The logistic models generated by using complete-case analysis and MI methods can also be compared in Table 6. The complete-case model used only 46% of the data in cycle 21 and 61% in cycle 22. When we compared the complete-case and MI models, it was noted that the odds ratio for the two models were similar for most of the variables. However, larger odds ratio for sex and race in MI models were seen except employment, depression, and physical activity in cycle 21. In addition, age, smoking, employment, and depression were smaller in the cycle 22 MI model.

4. Discussion and Conclusion

It has been widely acknowledged that obesity is associated with health issues such as high blood pressure, cardiovascular disease, and cancer (Freedman, et al., 2007). In addition, excessive alcohol consumption and its negative health effects among young adults are growing concerns of global public health (Peltzer, et al., 2014).

However, the studies on the association between alcohol consumption and weight gain are inconsistent (Suter & Tremblay, et al., 2008). Whereas some researchers found positive association between alcohol consumption and obesity (Park, et al., 2017), others found a negative between alcohol intake and weight gain (Haffner, et al.; et al.,1990) or no association (Keenan, et al.,1992; Wakibayashi, I;, 2012). Due to various factors that may have an impact on this relationship, it is difficult to explain the inconsistency in the studies. For this reason, this thesis attempted to focus on a prospective cohort of young adults to monitor the impact of alcohol in gaining weight in this transition period of life. However, tracking weight gain and alcohol consumption over the time with follow-up studies may cause some missingness.

Indeed one potential reason for the mixed longitudinal findings could be how those studies addressed missing data. Missing data can occur due to a variety of reasons including loss to follow-up, and non-response. As a result, missing data can lead to loss of efficiency in data analysis and is one of the main challenges of longitudinal studies and large surveys. For instance, loss to follow-up in most longitudinal studies ranges from 5% to 53% (Forster, et al., 2008; McNairy, et al., 2017; Kaplan, et al., 2017). In this thesis, the number of participants in the follow-up cycles was 596 in cycle 21 and 796 in cycle 22 in comparison to the participants at baseline (n=1293). Therefore, this thesis attempted to address missing data challenges by utilizing multiple imputation. All the statistical results were compared between complete cases and imputed cases. One of the advantages of multiple imputation model is that it can decrease standard errors by providing both within-imputation and between-imputation variabilities. The other advantage of utilizing a multiple imputation model is that it can provide unbiased estimates and may therefore be more valid. Nevertheless, in this thesis many similar findings were observed between the complete-case and MI general linear models in cycles 21. However, estimates and standard errors were generally

smaller for the MI models compared to the complete case models. In addition, when we compared the complete-case and MI logistic models, we noted that the odds ratio were consistent for most of the variables, with the exception of sex and race.

In this study, different alcohol consumption categories were defined to assess the relationship with BMI for cycles 21 and 22. For cycle 21, no relationship was seen between alcohol consumption and body weight or the risk of having overweight/obesity. However, in the subsequent cycle, compared to those who never had a drink in the past year, those who consumed alcohol at least once a month had lower BMI and lower odds of having overweight or obesity in both the complete-case and the MI models. These findings are aligned with a prospective cohort study in Japan in which an inverse association between age, alcohol consumption, and obesity among Japanese men was detected (Wakibayashi, 2012).

Preliminary evidence from the literature further suggests there is an interaction with age on the alcohol and obesity relationship (Wakibayashi, 2012). Indeed it seems that alcohol consumption does not have an immediate impact on body mass among young adults, or it may have an inverse relationship. However, as alcohol remains one of the main contributing factors to chronic diseases such as high blood pressure, cardiovascular disease, and cancer (Freedman, et al., 2007), it is perhaps a latent effect that does not show any immediate positive relationship between symptoms of these diseases at early stages of young adulthood. Nevertheless, the exact reason for an inverse relationship remains unclear. In order to get a better understanding follow-up studies investigating the long-term impacts of alcohol consumption in large cohorts of young adults are needed. Perhaps frequency and consistency of alcohol consumption can increase the likelihood of developing symptoms in other stages of life (Hvidtfeldt, et al., 2010).

The generalizability of this study is subject to certain limitations. For instance, the participants were young adults in Montreal. Therefore, they may not be representative of the general Canadian young adult population. Although data were from a cohort study, many variables of interest were more cross-sectional in nature and no causality can be inferred. Another point to consider is that many measures of interest were based on self-reported data, with their known limitations in bias and error (Carpenter & Carpenter, 2012). In addition, measurement of alcohol use such as type of

beverage or calorie count were not available in this study. While this is similar to other cross-sectional study limitations, they are important confounding variables that should be addressed in future studies. Lastly, further research should include other participants from different regions other than Montreal to compare the result with this current study.

As overweight and obesity is related to many different conditions, it is difficult to assess the independent impact of alcohol consumption on the risk of obesity. Although the current study was limited by the short-term follow up period with certain difficulties under the free-living condition to control for lifestyle habits, this study focused on young adults to complement those of earlier studies and addresses a critical gap in the literature. In addition, this study attempted to rely on two existing follow-ups among young adults to explore association between alcohol intake and weight gain which is a major public health concerns. However, in order to get a better understanding of the longitudinal effect of alcohol on body mass, additional follow-up studies are needed. Considerable more work and follow-up studies will need to be done to determine the exact impact of alcohol consumption on BMI in the long term.

Nevertheless, the study of cycle 21 and cycle 22 in this prospective cohort of young adults highlight the importance of the passage of time in providing better estimates of the impact of alcohol consumption on obesity. Although cycle 21 shows no association between alcohol consumption and BMI, cycle 22 indicates a negative association between them. This unanticipated result can suggest the significance of calorie density. For instance, the calorie content of a 5 oz. glass of wine is the same as the calorie content of 12 oz. glass of beer and investigating types of alcohol and their related calorie density may be useful for more accurate results. In addition, the frequency of alcohol consumption and the amount of alcohol intake must be taken into consideration.

In this study, an inverse relationship between alcohol consumption and BMI in cycle 22 was detected. The results must be interpreted with caution because it cannot be generalized for all ages, nor all populations. Although alcohol intake was inversely associated with risk of overweight in cycle 22, further research is needed on the same cohort to further track the weight changes and other possible health problems for alcohol consumers in the long-term. Indeed, longitudinal studies

suggest that increasing alcohol intake may cause weight gain over time (Traversy & Chaput, Alcohol Consumption and Obesity: An Update., 2015).

Therefore, it seems that the impact of alcohol on BMI is a long-term process rather than an immediate effect among young adults. In other words, the impact of alcohol on obesity is not among short term effects of alcohol, but it might be counted as long term consequences of alcohol consumption.

Taking all into consideration, it seems that age, passage of time, type of alcohol, the amount of alcohol intake, frequency of alcohol consumption, consistency of use, may play important roles on BMI in the prospective cohort of young adults. Perhaps the cause-and-effect association between alcohol consumption and weight is a slow process and the impact of alcohol on obesity is not among short term effects of alcohol, but it might be counted as long term consequences of alcohol consumption. Therefore, more follow-up studies may be needed to track weight gain. Taking all into consideration, it seems that age, passage of time, type of alcohol, the amount of alcohol intake, frequency of alcohol consumption, consistency of use, may play important roles on BMI in prospective cohorts of young adults and should be further investigated. However, because missing data issues are a common concern for all longitudinal studies, this thesis demonstrated the application of multiple imputation. Although the statistical methodology is not consistently utilized in this literature, in this thesis its use demonstrated unbiased estimates and smaller standard errors compared to the complete cases analysis. Therefore incorporating the aforementioned unmeasured covariates alongside an MI model would address a notable gap in the literature and should be further explored.

Appendix

Model 1a, 2a, 3a, 4a were conducted in complete cases.

Cycle		Interaction	Model
21	**Complete-Case**	**Sex * Alcohol**	**Model 1a**
	Imputed Data		**Model 1b**
22	**Complete-Case**		**Model 3a**
	Imputed Data		**Model 3b**
21	**Complete-Case**	**Student * Alcohol**	**Model 2a**
	Imputed Data		**Model 2b**
22	**Complete-Case**		**Model 4a**
	Imputed Data		**Model 4b**

The result in Table I shows that sex in models 1a, 2a and 2b was a significant predictor for BMI (model 1a: B =-2.71, p=0.02, model 2a: B =-1.38, p=0.001, model 2b: B =-0.97, p=0.002). The interaction terms between sex and student status with alcohol consumption were not significant in any models.

Similar to cycle 21, Table II shows that sex is a significant predictor for BMI. The association between mean BMI and student status is also significant. Alcohol consumption and BMI are significantly negatively associated (model 4a: B=-2.05, p=0.0011, model 4b: B=-1.61, p= 0.0003). Alcohol drinking (at least once a month) is negatively associated with BMI, for both the complete case and imputed model (4a and 4b). For a person who consumed alcohol at least once a month, the predicted BMI was 2.05 units lower than a person who drank less than once a month.

Table I GLM results with Interaction terms: Betas (SE) on the Association between Alcohol Consumption with BMI (Cycle 21)

Covariates	Model 1a (Complete)	*p*	Model 1b (MI)	*p*	Model 2a (Complete)	*p*	Model 2b (MI)	*p*
Age at Baseline	0.28(0.35)	0.43	0.21(0.27)	0.43	0.28(0.35)	0.43	0.21(0.27)	0.45
Alcohol[a] (less than once a month)	-3.57(1.97)	0.07	-2.86(2.34)	0.22	-1.97(0.9)	0.03	-0.73(0.92)	0.43
Alcohol (at least once a month)	-2.54(2.5)	0.31	-3.07(1.91)	0.11	0.12(1.03)	0.91	-2.11(0.79)	0.01
Sex (male vs female)	-2.71(1.17)	0.02	-2.02(1.13)	0.07	-1.38(0.35)	<.0001	-0.97(0.32)	0.002
Race (white vs non-white)	0.27(0.42)	0.52	0.03(0.31)	0.91	0.29(0.42)	0.49	0.04(0.31)	0.89
Smoking (smokers vs. non-smokers)	-0.2(0.16)	0.2	-0.11(0.14)	0.45	-0.2(0.16)	0.21	-0.11(0.14)	0.45
Employment Status (working vs. not working)	0.29(0.42)	0.49	0.18(0.41)	0.67	0.32(0.42)	0.44	0.17(0.41)	0.67
Student Status (student vs non-student)	-0.46(0.35)	0.2	-0.5(0.34)	0.14	-1.24(1.18)	0.3	-1.69(1.12)	0.13
In a relationship (yes vs. no)	-0.78(0.44)	0.08	-0.74(0.42)	0.08	-0.79(0.44)	0.08	-0.74(0.42)	0.08
Depression (depressed vs. not depressed)	-0.18(0.62)	0.77	0.02(0.55)	0.96	-0.2(0.62)	0.75	0.04(0.55)	0.95
Physical Activity	0.01(0.02)	0.42	0.02(0.02)	0.16	0.01(0.02)	0.44	0.02(0.02)	0.16
Sex* Alcohol (less than a month)	1.39(1.23)	0.26	1.61(1.4)	0.25				
Sex* Alcohol (at least once a month)	1.75(1.5)	0.24	1.03(1.18)	0.38				
Student * Alcohol (less than a month)					0.98(1.23)	0.43	0.79(1.4)	0.57
Student * Alcohol (at least once a month)					0.05(1.46)	0.97	1.36(1.14)	0.23

[a] Never drinkers' category is selected as reference group and Physical activity is calculated based on mean of moderate and vigorous levels of physical activity variable defined in method section.

Table II GLM results with Interaction terms: Betas (SE) on the Association between Alcohol Consumption with BMI (Cycle 22)

Covariates	Model 3a (Complete)	*p*	Model 3b (MI)	*p*	Model 4a (Complete)	*p*	Model 4b (MI)	*p*
Age at Baseline	0.27(0.35)	0.44	0.21(0.27)	0.44	0.28(0.35)	0.43	0.21(0.27)	0.44
Alcohol[a] (At least once a month)	-2.05(1.38)	0.14	-1.26(1.21)	0.3	-2.05(0.56)	<.0001	-1.61(0.49)	<.0001
Sex (male vs female)	-1.61(0.72)	0.03	-0.99(0.65)	0.13	-1.38(0.35)	<.0001	-0.98(0.32)	<.0001
Race (white vs non-white)	0.29(0.42)	0.48	0.05(0.31)	0.86	0.3(0.42)	0.48	0.05(0.31)	0.87
Smoking (smokers vs. non-smokers)	-0.2(0.16)	0.21	-0.11(0.14)	0.44	-0.2(0.16)	0.21	-0.11(0.14)	0.44
Employment Status (working vs. not working)	0.33(0.42)	0.43	0.18(0.41)	0.66	0.33(0.42)	0.43	0.17(0.41)	0.68
Student Status (student vs non-student)	-0.48(0.35)	0.18	-0.51(0.34)	0.14	-1.21(0.7)	0.08	-1.15(0.67)	0.08
In relationship (yes vs. no)	-0.81(0.44)	0.07	-0.74(0.42)	0.08	-0.79(0.44)	0.07	-0.74(0.42)	0.08
Depression (depressed vs. not depressed)	-0.2(0.62)	0.75	0.01(0.55)	0.98	-0.2(0.62)	0.75	0.01(0.55)	0.98
Physical Activity	0.01(0.02)	0.44	0.02(0.02)	0.16	0.01(0.02)	0.43	0.02(0.02)	0.16
Sex* Alcohol (At least once a month)	0.28(0.81)	0.73	0(0.74)	1				
Student * Alcohol (At least once a month)					0.96(0.78)	0.22	0.83(0.73)	0.26

[a] Never or less than once a month drinkers' category is selected as reference group and Physical activity is calculated based on mean of moderate and vigorous levels of physical activity variable defined in method section.

Bibliography

Adam Davey, T. D. (2020). A Systematic Approach to Identify and Evaluate Missing Data Patterns and Mechanisms in Multivariate Educational, Social, and Behavioral Research.

Albanese, S., & Bryson, J. (2015). Report on Alcohol Use, Harms & Potential Actions in Thunder Bay District. 55: Thunder Bay District Health Unit. Retrieved from https://www.tbdhu.com

Anderson, D., Feldblum, S., Modlin, C., Schirmacher, D., Schirmacher, E., & Thandi, N. (2005). A Practitioner's Guide to Generalized Linear Models. University of Connecticut, 122.

Barnard J, R. D. (1999). Small-Sample Degrees of Freedom with Multiple Imputation. Biometrika, 86(4), 984-955.

Bendsen, N., Christensen, R., & Bartels, E. (2013). Is Beer Consumption Related to Measures of Abdominal and General Obesity? Nutrition Reviews, 71(2), 67-87.

Bernaards, C., Belin, T., & Schafer, J. (2006). Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data. Statistics in Medicine, 26(6), 1368-1382.

Breslow, R., & Smothers, B. (2005, February 15). Drinking Patterns and Body Mass Index in Never Smokers. American Journal of Epidemiology, 161(4), 368-376.

Canada, P. H. (2011). Public Health Agency of Canada. Actions taken and future directions. Goverment of Canada. Retrieved from https://www.canada.ca/en/public-health/services/health-promotion

Carpenter, M., & Carpenter, J. (2012). Multiple Imputation and its Application. John Wiley & Sons, Ltd, 345.

Collins, L., Schafer, J., & Kam, C. (2001). A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. Psychological Methods, 6(4), 330-351.

Collins, S. (2016). Outcomes, Associations Between Socioeconomic Factors and Alcohol Outcomes. Alcohol Research, 38(1), 83-94.

De Castelnuovo, A., Costanzo, S., Bagnardi, V., Donati, B., Iacoviello, L., & de Gaetano, G. (2006). Alcohol Dosing and Total Mortality in Men and Women: An Updated Meta-Analysis Of 34 Prospective Studie. Arch Intern Med.,166(22), 2437-2445.

Demirtas, H., & Schafer, J. (2003). On the Performance of Random-Coeffi Cient Pattern-Mixture Models for Non-Ignorable Drop-Out.. Statistics in Medicine, 22(16), 2553-2275.

Dong, Y., & Peng, J. (2013). Principled Missing Data Methods for Researchers. SpringerPlus.

Dorn, J. M., Hovey, K., Muti, P., & Freudenheim, J. (2003). Alcohol Drinking Patterns Differentially Affect Central Adiposity as Measured by Abdominal Height in Women and Men. The Journal of Nutrition, 133(8), 2655–2662.

Dumesnil, C., Dauchet, L., Ruidavets, J. B., Bingham, A., Arveiler, D., & Ferrieres, J. (2013). Alcohol Consumption Patterns and Body Weight. Ann Nutr Metab, 62(2), 91-97.

Ellis, K. (2001). Review Selected Body Composition Methods Can Be Used in Field Studies. Journal of Nutrition, 131(5), 1589S–1595S.

Enders, C. K. (2006). A Primer on the Use of Modern Missing-Data Methods in Psychosomatic Medicine Research. Psychosomatic Medicine, 68(3), 427-436.

Enders, C. K. (2010). Applied Missing Data Analysis. (S. E. Little, Ed.) New York: The Guilford Press.

Forster, M., Bailey, C., Brinkhof, M., Graber, C., Boulle, A., & Spohr, M. (2008). Electronic Medical Record Systems, Data Quality and Loss to Follow-Up: Survey of Antiretroviral Therapy Programmes in Resource-Limited Settings. Bull World Health Organ., 939-947.

Freedman, D., Mei, Z., Srinivasan, S. R., Berenson, G. S., & Dietz, W. H. (2007). Cardiovascular Risk Factors and Excess Adiposity Among Overweight Children and Adolescents. Original Article, 150(1), 12-17.

French, M. T., Norton, E. C., Fang, H., & Maclean, J. C. (2010). Health Econ. 19(7), 814-832.

Graham, J. W., & Hofer, S. M. (2000). Multiple Imputation in Multivariate Research. Psychology Press, 16.

Guerri, C., & Pascual, M. (2010). Mechanisms Involved in The Neurotoxic, Cognitive, and Neurobehavioral Effects of Alcohol Consumption During Adolescence. Elsevier, 44(1), 15-26.

Haffner, S., Stern, M., Hazuda, H., Pugh, J., Patterson, J., & Malina, R. (1986). Upper Body and Centralized Adiposity in Mexican Americans and Non-Hispanic Whites: Relationship to Body Mass Index and Other Behavioral and Demographic Variables. Int J Obes, 10(6), 493-502.

Hartley, H., & Hocking, R. (1971). The Analysis of Incomplete Data. Biometrics, 27(4),783-823.

Hens, N., & Aerts, M. (2006). Model Selection for Incomplete and Design-Based Samples. Stat Med, 25(14), 2502-2520.

Hingson, R. W., Heeren, T., & Winter, M. R. (2006). Age at Drinking Onset and Alcohol Dependence. Arch Pediatr Adolesc Med, 160(7), 739-746.

Hvidtfeldt, U., Tolstrup, J., Jakobsen, M., Heitmann, B., Grønbæk, M., O'Reilly, E., Ascherio, A. (2010). Alcohol Intake and Risk of Coronary Heart Disease in Younger, Middle-aged and Older Adults. Circulation, 121(14), 1589-1597.

Jacobs, M. (2015). Improved Rounding Methods for Binary and Ordinal Variables Under Multivariate Normal Imputation. Thesis, The University of Western Australia.157.

Kaplan, S., Oosthuizen, C., Stinson, K., Little, F., Euvrard, J., & Schomaker, M. (2017). Contemporary Disengagement from Antiretroviral Therapy in Khayelitsha South Africa: A Cohort Study. PLOS Medicine, 14(11): e1002407.

Kaye, S., Folsom, A., Prineas, R., Potter, J., & Gapstur, S. (1990). The Association of Body Fat Distribution with Lifestyle and Reproductive Factors in a Population Study of Postmenopausal Women. International Journal in Obesity, 14(7), 583-591.

Keenan, N., Strogatz, D., James, S., Ammerman, A., & Rice, B. (1992). Distribution and Correlates of Waist-To-Hip Ratio in Black Adults : The Pitt County Study. American Journal of Epidemiology, 135(6), 678–684.

Lewis, C., Jacobs, D., McCreath, H., Schreiner, P., Smith, D., & Williams, O. (2000). Weight Gain Continues in the 1990s: 10-year Trends in Weight and Overweight from the CARDIA Study. American Journal of Epidemiology, 151(12), 1172–1181.

Lieber, C. S. (2000). Its Metabolism and Interaction With Nutrients. Annual Review of Nutrition, 20(1), 395-430.

McNairy, M., Joseph, P., Unterbrink, M., Galbaud, S., & Mathon, J.-E. (2017). Outcomes after Antiretroviral Therapy During the Expansion of HIV Services in Haiti. PLOS ONE, 12(4): e0175521.

Molenberghs, G., & Ibrahim, J. (2009). Missing Data Methods in Longitudinal Studies: A Review. Test (Madr), 18(1), 1-43.

Newman, D. A. (2003). Longitudinal Modeling with Randomly and Systematically Missing Data: A Simulation of Ad Hoc, Maximum Likelihood, and Multiple Imputation Techniques. Organizational Research Methods, 6(3), 328-362.

Oesterle, S., Hill, K., Hawkins, J., Guo, J., Catalano, R., & Abbott, R. (2004). Adolescent Heavy Episodic Drinking Trajectories and Health in Young Adulthood. Journal of Studies on Alcohol, 65(2), 204-212.

Pajari, M., Pietilainen, K., Kaprio, J., & Saarn, S. (2010). The Effect of Alcohol Consumption on Later Obesity in Early Adulthood -- A Population-Based Longitudinal Study. Alcohol Alcohol, 45(2), 173–179.

Park, K., Park, H., & Hwang, H. (2017). Relationship Between Abdominal Obesity and Alcohol Drinking Pattern in Normal-Weight, Middle-Aged Adults: the Korea National Health and Nutrition Examination Survey 2008-2013. Public Health Nutrition, 20(12), 2192-2200.

Peltzer, K., Pengpid, S., Samuels, T., Özcan, N., Mantilla, C., Rahamefy, O., Gasparishvili, A. (2014). Prevalence of Overweight/Obesity and Its Associated Factors Among University. Int. J. Environ. Res. Public Health, 11(7), 7425-7441.

Raben, A., Agerholm-Larsen, L., Flint, A., Holst, J. J., & Astrup, A. (2003). Meals with Similar Energy Densities but Rich in Protein, Fat, Carbohydrate, or Alcohol Have Different Effects on Energy Expenditure and Substrate Metabolism but not on Appetite and Energy Intake. The American Journal of Clinical Nutrition, 77(1), 91-100.

Roberts, K., Shields, M., Degroh, M., Aziz, A., & Gilbert, J. (2011). Overweight and Obesity in Children and Adolescents: Results from the 2009 to 2011 Canadian Health Measures Survey. Health Report, 23(3), 37-41.

Rubin, D. (1978). Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse. Applied Mathematics, 5(21), 20-28.

Rubin, D. B., & Little, R. J. (2002). Statistical Analysis with Missing Data (2nd ed.). Hoboken, NJ: Wiley.

Sallis, J., Condon, S., Goggin, K., Roby, J., Kolody, B., & Alcaraz, J. (1993). The Development of Self-Administered Physical Activity Surveys for 4th Grade Students. Research Quarterly for Exercise and Sport, 64(1), 25-31.

Sayon-OreaC, Martinez-Gonzalez, M., & Bes-Rastrollo , M. (2011). Alcohol Consumption and Body Weight: a Systematic Review. Nutrition Reviews, 69(8), 419–431.

Schafer, J. (1997). Analysis of Incomplete Multivariate Data. Statistics in Medicine, (1st Edition ed.). New York, 444.

Schafer, J. (1999). Multiple Imputation a Primer. SAG Journals, 8(1), 3-15.

Shields, M., Groh, M., Aziz, A., Gilbert, J., & Roberts, K. (2011). Overweight and Obesity in Children and Adolescents: Results from The 2009 to 2011 Canadian Health Measures Survey. Health Reports, 23(3), 37-41.

Shrive, F., Stuart, H., Quan, H., & Ghali, W. (2006). Dealing with Missing Data in a Multi-question Depression Scale: a Comparison of Imputation Methods. BMC Med Res Methodol 6, 57.

Singh, A., Mulder, C., Twisk, J., Mechelen, W., & Chinapaw, M. (2008). Tracking of Childhood Overweight into Adulthood: A Systematic Review of the Literature. Journal of the International Association for the Study of Obesity, 9(5), 474-488.

Smothers, B. A., & Breslow, R. A. (2005). Drinking Patterns and Body Mass Index in Never Smokers. American Journal of Epidemiology, 161(4), 368–376.

Statistics Canada. (2015). Table 13-10-0456-01, Measured Adult Body Mass Index (BMI), World Health Organization classification, by age Group and Sex,. Retrieved from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310045601

Suter, P. (2005). Is Alcohol Consumption A Risk Factor For Weight Gain and Obesity? Crit Rev Clin Lab Sci, 42(3), 197-227.

Tanner, M., & Wong, W. (1987). The Calculation of Posterior Distributions by Data Augmentation (with discussion). Journal of the American Statistical Association, 82(398), 528-540.

Timmerby, N., Martiny, K., Lunde, M., & Soendergaad, S. (2015). Psychometric Evaluation of the Major Depression Inventory (MDI) as Depression Severity Scale in Chinese Patients With Coronary Artery Disease. Findings From the MEDEA FAR-EAST Study. BMC Psychiatry, 15(190).

Tolstrup JS, Heitmann BL, Tjønneland AM, Overvad K, Sørensen A, Grønbaek M,. (2005). The Relation Between Drinking Pattern and Body Mass Index and Waist and Hip Circumference. Int J Obes (Lond).,29(5), 490-497.

Traversy, G., & Chaput, J. (2015). Alcohol Consumption and Obesity: An Update. Current Obesity Report, 4(1), 122-130.

Wakibayashi, I. (2012). Age-Dependent Inverse Association Between Alcohol Consumption and Obesity in Japanese Men. Obesity a Research Journal, 19(9), 1881-1886.

Wannamethee, S., Shaper, A., & Whincup, P. (2005). Alcohol and Adiposity: Effects of Quantity and Type of Drink and Time Relation with Meals. Int J Obes (Lond)., 29(12), 1436–1444.

Wellman, R., Sylvestre, M., Abi Nader, P., Chiolero, A., Mesidor, M., Dugas, E., O'Loughlin, J. (2020). Intensity and Frequency of Physical Activity and High Blood Pressure in Adolescents: A Longitudinal Study. J Clin Hypertens (Greenwich), 22(2), 283-290.

Williams, D., Going, S., Lohman, T., Harsha, D., Srinivasan, S., Webber, L., & Berenson, G. (1992). Body Fatness and Risk for Elevated Blood Pressure, Total Cholesterol, and Serum Lipoprotein Ratios in Children and Adolescents. American Journal of Health, 82(3), 358-363.

Wright, J. (2001). International Encyclopedia of the Social & Behavioral Sciences. Elsevier, 23185.

Yeomans, M. (2010). Alcohol, Appetite and Energy Balance: Is Alcohol Intake a Risk Factor for Obesity? Physiology & Behavior, 100(1), 82-90.