Body Mass Index and Missing Data:

Examining the Levels, Patterns, and Impacts of Missing Data in a Large Cohort Study of Canadian Youth

by

Amanda Doggett

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Public Health Sciences

Waterloo, Ontario, Canada, 2022

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Leah Lipsky Staff Scientist, Eunice Kennedy Shriver National Institute of Child Health and Human Development National Institutes of Health
Supervisor:	Scott T. Leatherdale Professor, School of Public Health Sciences University of Waterloo
Internal Members:	Ashok Chaurasia Associate Professor, School of Public Health Sciences University of Waterloo
	Jean-Philippe Chaput Professor, Department of Pediatrics University of Ottawa Cross-Appointment, School of Public Health Sciences University of Waterloo
Internal-External Member:	Marina Mourtzakis Professor, Kinesiology and Health Sciences University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis contains three manuscripts which have been published or submitted for publication. Exceptions to sole authorship are:

Chapter 4: Doggett A, Chaurasia A, Chaput JP, Leatherdale ST. Learning from missing data: examining nonreporting patterns of height, weight, and BMI among Canadian youth. *International Journal of Obesity.* 2022 9;46(9):1598–1607. https://doi.org/10.1038/s41366-022-01154-8

Chapter 5: Doggett A, Chaurasia A, Chaput JP, Leatherdale ST. Using classification and regression trees to model missingness in youth BMI, height, and body mass. Under review at *Health Promotion and Chronic Disease Prevention.*

Chapter 6: Doggett A, Chaurasia A, Chaput JP, Leatherdale ST. Assessing the impact of missing data in youth overweight and obesity research: complete case analysis versus multiple imputation. Submitted to *International Journal of Social Research Methodology.*

As lead author for Chapters 4, 5, and 6, I was responsible for the development of the research questions, conducting the literature review, planning and implementation of statistical analyses, interpretation of findings, and drafting all sections of each manuscript. My co-authors, Drs Chaurasia, Chaput, and Leatherdale provided guidance throughout each step, gave feedback on manuscripts drafts, and approved each final manuscript for submission to academic journals. Under the supervision of Dr. Leatherdale, I also prepared the remaining chapters of this thesis, which were not written for publication.

Abstract

Missing data are generally unavoidable in survey-based research. Small amounts of random missingness may not pose significant problems; however, issues arise when data are missing in large proportions or when missingness follows a systematic pattern. Survey items that are tied to social desirability can be markedly impacted by non-response. Youth are a major target for survey-based research, with many cohort studies using surveys to monitor youth health and their health behaviours. Two common health-related measures that are often collected using self-reported surveys are height and weight, used to calculate body mass index (BMI). BMI is used as a proxy for body adiposity at the population level to identify individuals with overweight or obesity (OWOB). BMI is an important measure for research and population surveillance as it is a well-established predictor of future chronic disease. Among youth, OWOB trajectories tend to track into adulthood, and there is substantial literature exploring youth OWOB and associated factors. However, few existing studies have examined youth nonreporting of height and weight. Those studies which have examined nonreporting suggest that for youth, BMI tends to be missing in high proportions (exceeding missingness for other measures) and often follows a systematic pattern of missingness.

There are several methods through which researchers can manage missing data. The most common approach is complete case analysis (CCA), whereby missing cases are deleted, and analyses are performed using only complete data. Due to the loss of information, CCA can introduce inefficiencies and bias into statistical results. Hence, in situations where data are missing systematically and in high proportions, CCA is not recommended. More sophisticated techniques, such as multiple imputation (MI), can yield unbiased and efficient estimates in these situations; however, they are not commonly leveraged in epidemiological studies. In fact, systematic reviews have suggested that information on missing data is typically not presented.

This dissertation aimed to explore levels, patterns, and impacts of missing data among youth, specifically focusing on nonreporting of height, weight, and subsequently calculated BMI using data from a large youth-focused survey. This research leveraged data from 74,501 youth who participated in the 2018/19 wave of the COMPASS study. The COMPASS study is a survey-based cohort study among youth aged 12-19 years in Canada examining a variety of different aspects of health and health behaviours. Study 1 examined variables associated with missingness in BMI, height, and weight using model selection in three separate logistic regressions. Study 2 examined patterns, hierarchies, and subgroups of missing BMI, height, and weight data using classification and regression tree (CART) models. Finally, Study 3 compared the differences in findings between CCA and MI missing data approaches in the context of factors associated with youth BMI through linear mixed models.

Study 1 found that nearly 1 in 3 youth in this sample were missing BMI data. Among those with missing BMI, 32% did not report their weight, 20% did not report their height, 36% reported neither weight nor height, and 12% were reduced to missing due to unrealistic values. A greater proportion of females were missing weight only, whereas a greater proportion of males were missing height only. Of all the youth-reported measures, BMI,

height, and weight showed the highest degree of missingness. For both males and females, perceiving oneself as overweight was associated with a greater likelihood of BMI being missing. Indicators of poor diet and physical inactivity were also significantly associated with missing BMI. Taken together, results of Study 1 suggest that social desirability played a significant role in nonreporting patterns, and it is likely that those who have a higher BMI are less likely to report their height or weight.

Study 2 identified that certain subgroups of youth (characterized by various health behaviours and indicators) were more likely to be missing BMI. Confirming findings from Study 1, patterns of systematic missingness in BMI were identified using CART models. Examining the identified subgroups highlighted that a combination of weight perception, low physical activity, poor academic performance, and poor mental health almost certainly lead to nonreporting. Study 2 also identified a hierarchy of importance for the variables related to missingness in BMI, height, and weight, providing more context to the associations observed in Study 1 and highlighting the utility of a CART approach to examine missing data.

Studies 1 and 2 illustrated that in this sample, BMI missingness was highly prevalent and non-random. Using the findings from these two studies, Study 3 illustrated the bias that can occur when missing data are not managed appropriately. MI and CCA approaches produced contrasting results across sex-stratified models examining factors associated with youth BMI. These results illustrated how bias from deleting cases may impact findings and lead to considerably different research conclusions, highlighting the importance of thorough examination and appropriate handling of missing data.

This dissertation fills an important gap in the research examining patterns and impacts of missingness in youth BMI, height, and weight. In this dissertation, missingness in youth BMI was found to be highly prevalent and followed a systematic pattern. Identified patterns indicated that nonreporting was likely influenced at least in part by social desirability, and that those with a higher BMI were less likely to report their height and/or weight. Subgroups of youth who had poorer outcomes for physical activity, school grades, and mental health were nearly guaranteed to be missing BMI. When carried forward into an analysis examining factors related to youth BMI, deleting the missing cases introduced bias into findings. This research highlights a great need for improved missing data reporting and handling within youth OWOB research. Similar cohort studies that collect youth height and weight through self-report measures should perform thorough examinations of missing data and choose appropriate methodologies to manage missingness. This research also suggests that researchers should exert caution when interpreting and utilizing results from studies where missing data are not well-reported.

Acknowledgements

Thank you to all three of my amazing committee members. Scott, in your time as my supervisor you have showed great trust in me by allowing me the freedom to choose research topics, letting me run with my ideas, and connecting me to countless opportunities within your network. I am incredibly appreciative of the graduate experience that you have facilitated for me and the support you have provided over the years. Ashok, thank you for always making time to chat stats with me. I value the substantial expertise you have shared throughout this dissertation as well as the many other projects we have worked on together. JP, you have been a wonderful committee member to work with, and I have greatly appreciated the insights and perspectives that you have added to this research. Your reliability and dependability throughout this process is something that I greatly valued.

Thank you to all members of the COMPASS team, both past and present, who have provided substantial support and expertise over the years; it was invaluable in helping me navigate graduate school. Thank you Mahmood for always making time to chat and patiently work through complex stats problems with me! Also thank you Kelly Skinner, for teaching me all I know about program evaluation and connecting me with many opportunities that I would not have otherwise had. Thank you to my colleagues at the Center for Teaching Excellence; you were incredible people to work with and I really enjoyed learning from all of you.

To my family, Derhyk, Kait, Elizabeth, Andrea, and Allen, I have greatly appreciated your support, questions, encouragement, and jokes throughout this process. To Sharon and Bruce (better known as Mom and Dad), thank you for your unconditional love and support throughout my entire education. I am pretty sure that I'm finally done school (maybe)! To my partner, Carlos, thank you for your unwavering love and support throughout this process. You were a soundboard, a voice of reason, a challenger, and a cheerleader. This PhD would not have been possible without you.

To, Alle, Claire, and Isabella, you could probably each have your own page(s) here, but I'll keep it short. You'll recall that I introduced Alle to my parents as "the reason I am still in graduate school," which is perhaps the most accurate introduction I have ever given, since I can't imagine my graduate school experience without you three. You have all been there for every acceptance, every rejection, and every "hidden curriculum" situation that I have faced. There is something unquestionably special about going through graduate school with other incredible women who love, support, and improve each other. My greatest wish for others is that they can find phenomenal friends like you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIC Akaike Information Criterion

AmED Alcohol Mixed with Energy Drinks

BIC Bayesian Information Criterion

BMI Body Mass Index

CART Classification and Regression Trees

CCA Complete Case Analysis

CESD-R-10 Center for Epidemiologic Studies Depression Scale Revised

CI Confidence Interval

Cq COMPASS Student Questionnaire

DERS Difficulties in Emotional Regulation Scale

EDA Exploratory Data Analysis

EM Expectation-Maximization

FCS Fully Conditional Specification

FIML Full Information Maximum Likelihood

GAD-7 Generalized Anxiety Disorder 7-item

GLMM Generalized Linear Mixed Model

LMM Linear Mixed Model

LOCF Last Observation Carried Forward

MAR Missing At Random

MCAR Missing Completely At Random

MCMC Markov Chain Monte Carlo

MI Multiple Imputation

MICE Multiple Imputation by Chained Equations

ML Maximum Likelihood

MVPA Moderate to Vigorous Physical Activity

NMAR Missing Not At Random

OWOB Overweight or Obesity

PMM Predictive Mean Matching

STSB Screen Time Sedentary Behaviour

Chapter 1

Introduction

1.1 Missing Data

Missing data is a problem in most research, but particularly for epidemiological studies that utilize surveys or questionnaires as data collection instruments. Due to the voluntary nature of research, missing data is largely unavoidable. The optimal way to deal with missing data problems is to prevent missing data in the first place. There are a variety of strategies that can be used to ensure that study design will be as robust as possible to avoid missing data. Designs that incorporate potential for follow-up to gather missing information from initial collection, or those that offer multiple methods of participation, may by nature be more robust against missing data [1]. However, the ability of quantitative survey-based research to accomplish this may be limited, particularly for large epidemiological studies where follow-up or multi-modal data collection is not feasible. Certainly, propensity for missingness should still be limited as much as possible in the survey design; for example, by ensuring that questions are written in such a way that they are simple and avoid confusing terminology or structure [1]. However, there are inevitably aspects which cannot be controlled for with survey design. Questions inherently tied to social desirability (e.g., asking about substance use, finances, etc.), are more likely to suffer from non-response [2, 3]. Research that focuses on areas linked to social desirability may find that they have greater rates of missing data in comparison to research from other domains.

As missing data is largely unavoidable, researchers must develop and use appropriate methods of handling missingness post-data collection through statistical techniques. While many novel statistical approaches exist to address missing data, they are often underutilized. There are a variety of barriers to adoption of robust statistical methods for missing data, including researcher time and skill, software or hardware limitations, and what methods are considered the status-quo for particular fields. The following sections give an overview of missing data assumptions and methods, and identify the most common approaches while highlighting some of the barriers to uptake of more advanced methods.

1.1.1 Identifying Missingness

Data can be missing in a variety of ways. It is essential to understand the different types and patterns of missing data in order to understand potential causes of missingness, as well as make decisions about methodological approaches. There are two main types of missing data: unit non-response and item non-response. Distinguishing between these is important; each introduce unique issues into quantitative research with different strategies for dealing with them.

Unit Non-Response

Unit non-response refers to the complete lack of data at a particular time point. For example, a subject who refuses to participate in a study would represent an instance of unit non-response (in this case, the unit being a person). In human-based research, unit non-response can be considered interchangeable with term subject non-response. A subject dropping out of a longitudinal study (i.e., attrition) would also represent unit non-response at that particular time point. Unit non-response reduces sample size, although a strong research design should consider non-response at the outset and adjust sampling procedures accordingly to obtain sufficient statistical power [1]. Those who do not participate in data collection may be systematically different than those who do, in which case unit non-response may lead to biases. Research has suggested that for this reason, prevalence of many health risk behaviours are typically underestimations of true population rates, as those who are more likely to engage in risky behaviours are also those who are more likely to be selective non-responders [4]. Unit non-response can be handled through weighting techniques, although the feasibility of adjusting for any potential biases will depend on the availability of auxiliary information about non-responders [5]. Depending on the research design and sampling technique, there may be little to no available information on non-responders, and therefore ability to mitigate bias is limited. Where research design does not allow for information to be obtained on non-responders, it is important to mitigate unit non-response as much as possible through other aspects of design. For example, the consent procedure of a study; passive consent procedures drastically limit non-response bias compared to active consent procedures [6, 7]. Given the limitations with respect to handling unit non-response, reference to "missing data" throughout this dissertation refers to item non-response.

Item Non-Response

Item non-response refers to missingness on a particular item, rather than complete missingness for that subject. For example, if a subject filled out a questionnaire, but did not answer one or more of the items, this is considered item non-response. Item non-response is more feasibly dealt with compared to unit non-response, since there is at least some data available on subjects with missing data, which can then be used to inform statistical techniques. Item non-response may lead to bias similar to unit non-response, since those who choose not to answer a particular question may be systematically different than those who do.

1.1.2 Missing Data Patterns

A pattern of missingness refers to the structure of data for a particular subject, or within a dataset. The most important broad distinction is between **monotone** and **non-monotone** patterns of missingness. Missing data patterns are best represented in diagram format (see Figure 1.1), and are always discussed in the context of a dataset being organized from the variable with the least missingness to that with most missingness.

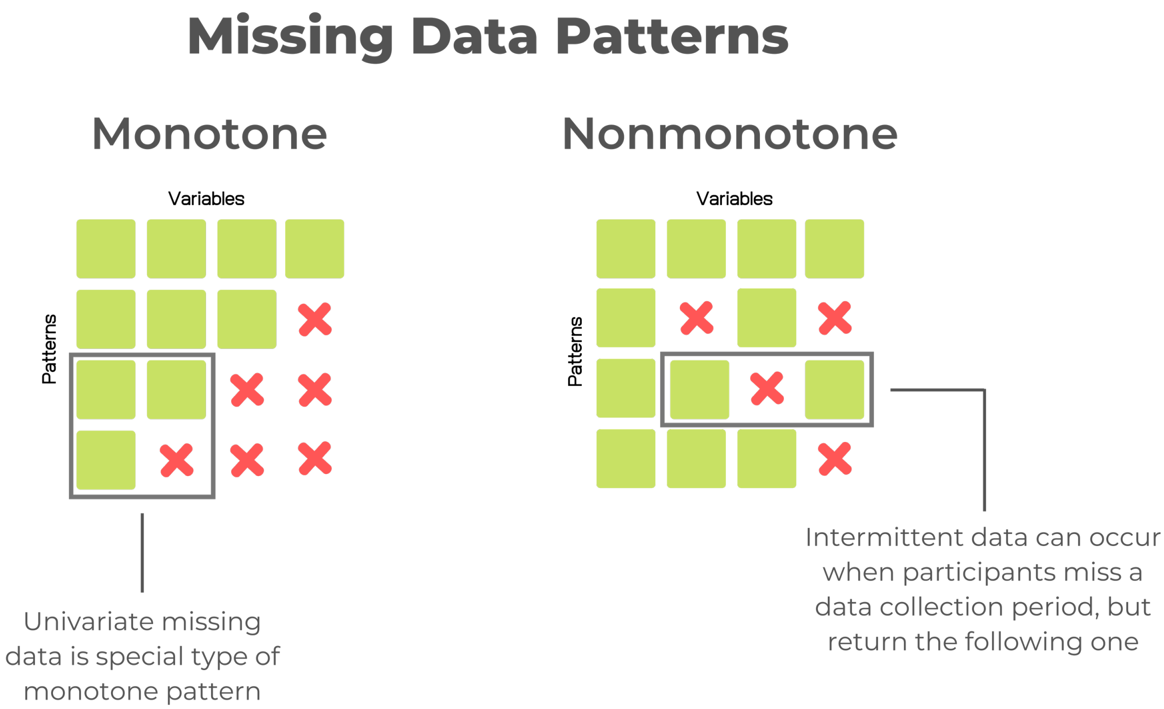


Figure 1.1: Missing Data Patterns. Figure adapted from van Buuren [8]. Green squares represent present data, whereas red x's represent missing data. Variables in this figure are intended to be shown as if data were in wide-format, and as such can refer to distinct variables, or the same variables measured at different time points.

Monotone missing data are such that each variable with missingness will only be followed by a variable that also has missingness [8]. Monotone missingness is unlikely outside of when the dataset has been modified to adhere to such a pattern (which may be appropriate if only a few observations need be removed), or in longitudinal studies due to dropout [8]. If data are only missing on one variable, this is special case of monotone missing data, and can also be referred to as a univariate missing data pattern [8]. A monotone pattern of missingness can simplify some of the methodological approaches discussed later in Section 1.2.

Non-monotone broadly refers to anything that does not meet the requirements to be monotone. This can also be referred to as "arbitrary" missing data, which reflects that the missing data do not appear to follow any sort of pattern [9]. Another term sometimes used to describe non-monotone missing data is intermittent, which specifically refers to situations where data are present on a variable after a missing value has occurred; for example, if a participant were to miss one data collection period but attend a following one [10]. It is technically feasible to modify a non-monotone pattern of missingness to be monotone through deletion, however this is usually not recommended as it is likely to result in many deleted cases and can be a somewhat arbitrary approach [11].

1.1.3 Missing Data Mechanisms

The different classifications of missing data mechanisms were operationalized by Rubin in 1976 [12] and are foundational to the understanding and use of methods for handling missing data. A missing data mechanism simply refers to the equation which expresses the probability of missingness on a variable, as a function of other variables [13]. For example, a missing data mechanism can be a logistic regression equation, where the outcome is whether or not the variable of interest is missing, and predictors are the variables that are associated with missingness on the outcome.

There are three main assumptions that can be made regarding a missing data mechanism: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (NMAR) [13]. These are referred to as *assumptions* because researchers cannot prove with certainty that data follow a specific missingness mechanism. The most stringent assumption, MCAR, can be tested to some degree [14], but there is no way to prove or disprove MAR or NMAR with certainty in non-simulated data [15]. Definitions of these mechanisms along with contextual examples are provided below.

Missing Completely at Random (MCAR)

Missing data are assumed MCAR if the probability that data are missing for a certain variable does not depend on any other variable(s), nor does it depend on the value of itself. In effect, this assumes the data set with missingness is a simple random sample of the hypothetical "complete" sample.

Missing Data Notation: MCAR

Suppose D_{inc} represents a data set containing rows and columns for which the data are incomplete. $D_{inc} = (D_{obs}, D_{miss})$, where D_{obs} is the observed part of the data and D_{miss} is the missing part of the data. Referencing back to Figure 1.1, the green squares would represent D_{obs}, and the red x's would represent D_{miss}.

Suppose R represents the missingness variable, which is of the same dimension as D_{inc}, such that where data are missing $R = 1$ and where data are not missing, $R = 0$. As such, if data are completely observed, each column and row in R is equal to 0.

The missingness mechanism defines the relationship between R and D_{inc}; for MCAR this is:

$$f(R|D_{inc}) = f(R) \tag{1.1}$$

Here the mechanism makes the assumption that the distribution of R is **independent** of D_{inc}, which is the most strict of the three possible assumptions.

Example 1: MCAR

Consider a simple linear regression where the dependant variable *weight* has some missing data, and there are two independent variables which are fully observed: *sex* and *age*. Under the MCAR assumption, the probability that *weight* is missing does not depend on the value of *weight* itself. This means that people who weigh more are no more or less likely to report their *weight* than someone who weighs less. The MCAR assumption also assumes that *sex* and *age* are unrelated to the probability that *weight* is missing. Consider this example written as an equation, where Y is *weight* and X is a vector of *sex* and *age*:

$$Pr(Y\ is\ missing|Y, X) = Pr(Y\ is\ missing) \tag{1.2}$$

Missing at Random (MAR)

Missing data are assumed MAR if the probability that data are missing for a certain variable does not depend on that variable itself, but can depend on other variables [13]. Notably, MCAR is a special type of MAR; if a method is said to work under the MAR assumption, it should also work under the MCAR assumption. MAR is sometimes referred to as ignorability, which refers to the idea that the missing data mechanism can be ignored and need not be modeled [13].

Missing Data Notation: MAR

Suppose the same notation as previous. MAR relaxes some of the assumptions made in MCAR, such that:

$$f(R|D_{inc}) = f(R|D_{obs}) \tag{1.3}$$

Here the mechanism assumes that the distribution of R is dependent on the observed data, but not the missing data.

Example 2: MAR

Consider the same example as used previously in Example 1. Under the MAR assumption, the probability that *weight* is missing does not depend on *weight* itself, but on other observed variable(s) (e.g., *sex*, *age*, or both):

$$Pr(Y\ is\ missing|Y, X) = Pr(Y\ is\ missing|X) \tag{1.4}$$

In other words, someone who weighs more is neither more nor less likely to report their *weight* than someone who weighs less; but, perhaps females are less likely to report their *weight* than males. The probability that data for *weight* will be missing dependant on *sex* and/or *age* can be tested, but since it is not possible to prove if missingness in *weight* depends on the value of *weight* itself, MAR remains an assumption rather than a provable phenomenon.

Not Missing at Random (NMAR)

If missing data are assumed NMAR (sometimes also abbreviated as MNAR), then the missing data mechanism is not considered ignorable. In this case, any method that is used will require that the missing data mechanism be estimated alongside the estimation of the model of scientific interest [13]. In order to obtain appropriate estimates based on the NMAR assumption, one must to have solid prior knowledge of the missing data mechanism [13]. While existing literature could give some information about a reasonable missing data mechanism, if the research is sparse or contradicting, this may be difficult.

Missing Data Notation: NMAR

Suppose the same notation as previous. Unlike MCAR or MAR, the NMAR mechanism cannot be simplified, and as such is just:

$$f(R|D_{inc}) = f(R|D_{inc}) \tag{1.5}$$

Here the mechanism assumes that distribution of R is dependent on the observed data, and the missing data.

Example 3: NMAR

Continuing with the same example from Examples 1 and 2, now suppose missing data are assumed NMAR because there is prior knowledge to indicate that people with higher *weight* tend to choose not to report their *weight*:

$$Pr(Y\,is\,missing|Y,X) = Pr(Y\,is\,missing|Y) \tag{1.6}$$

It will also be important to consider whether or not this differs based on *sex* or *age*:

$$Pr(Y\,is\,missing|Y,X) = Pr(Y\,is\,missing|Y,X) \tag{1.7}$$

Assuming missing data are NMAR may greatly complicate analyses and should only occur where there is sufficient information to rationalize the decision. While it might seem logical based on what is known about social desirability to assume that individuals with a higher *weight* would be less likely to report their *weight*, there is little literature on this topic. Of course, this is difficult to prove in practice, as one would need to collect objective and self-report data for weight on the same participants, and somehow ensure absence of response bias. Without this information, we likely do not have enough to inform the missing data mechanism to avoid biased estimates.

It would be beneficial in this scenario to "convert" NMAR to MAR, through the inclusion of other variables that are known to be correlated with *weight* (e.g. eating habits, socioeconomic status) [15]. Adjusting for such variables can help decrease any residual correlation between *weight* and the probability of *weight* being missing (further discussed in Section 1.2.2) [15].

1.2 Missing Data Methods

Analyses can be done without much consideration of missing data, as modern statistical software typically deletes missingness by default. The problem with ignoring missing data is the potential introduction of bias, which may impact research results and conclusions. As such, there are three main goals of any missing data technique [16]:

1. Minimize bias (which can occur if data that are missing are different than the data that are not missing)
2. Maximize the use of available information
3. Reduce variability in estimates for standard errors, test statistics, confidence intervals, and p-values

Different methods of handling missing data, and how they meet the above goals under differing assumptions, are discussed in the following sections.

1.2.1 Traditional Methods

Complete Case Analysis

Complete Case Analysis (CCA), also known as listwise deletion, is the most common technique for handling missing data in applied research due to its simplicity and wide applicability [8]. CCA simply involves deleting the cases with missing data, such that only the complete cases (i.e., those with no missing data) are included in analyses [13]. CCA is the default for many analysis functions across many statistical programming languages (SAS, R, SPSS, etc.).

If the cases with missing data are statistically different than those without missing data, CCA will introduce bias [8]. If data are MCAR, then CCA is a valid method that will produce unbiased parameter estimates, since the complete cases are a simple random sample of the total sample [13]. However, CCA does not maximize the use of all available information; by deleting cases, statistical power is decreased and the standard errors produced will be larger than methods which utilize more of the available information [15]. Researchers should consider missing data assumptions as well as the fraction of missing data present when deciding whether or not to use a more sophisticated method than CCA.

Example 4: CCA where data are MAR

CCA may produce bias results if data are MAR. Consider the same example used previously throughout Section 1.1.3 with *weight*, *sex* and *age*, and suppose there is a moderate amount of missing data on the outcome *weight*. Let's say that missingness on *weight* does not depend on *weight* itself but does depend on *sex*, such that females are less likely to report their *weight* (i.e., MAR):

$$Pr(Y\ is\ missing|Y, X) = Pr(Y\ is\ missing|X) \tag{1.8}$$

In this situation, CCA will remove more females than males, and thus parameter estimates will be biased since females will be underrepresented in the sample. For example, since women tend to weigh less than men, in a scientific model that estimates mean *weight*, the parameter estimate would be biased upwards from men being over-represented.

Example 5: CCA where predictors are NMAR

Interestingly, CCA is one of the most robust techniques for avoiding bias when missing data on predictors is NMAR, but only if missingness does not depend on any other variables [13, 15]. For example, assume that within the predictor *age*, people who are younger are less likely to report how old they are. As long as the probability of missingness in *age* **does not** depend on the outcome variable *weight*, parameter estimates will not be biased with CCA. Of course, if we believe that *age* and *weight* are highly correlated (a reasonable assumption), then this may not hold true.

$$Pr(X\ is\ missing|X, Y) = Pr(X\ is\ missing|X) \tag{1.9}$$

Other Methods

There are a variety of other methods that can be used to address missing data, such as pairwise deletion, single imputation, last observation carried forward, etc. However, both applied research and simulation studies suggest that these methods do not match the utility of the methods discussed in Section 1.2.2 below, in particular as proportion of missingness increases [17–19]. For example, Last Observation Carried Forward (LOCF) was (and to some degree, still is) a commonly used method across several disciplines that have repeated measures, whereby the observation from a previous measurement is used to directly fill-in missing measurements. However, LOCF has been shown to introduce bias and is not recommended given the accessibility of more modern approaches [20, 21]. Given that other methods are generally less robust with respect to the goals of missing data techniques (discussed in Section 1.2) compared to the modern approaches discussed below, and are not often recommended, they are not discussed further in this dissertation.

1.2.2 Modern Methods

Maximum Likelihood

The basic principle of Maximum Likelihood (ML) is that parameter estimates are calculated such that if the estimates were in fact the true values, the probability of observing the actual observed values in the dataset is maximized [13]. This is accomplished by maximizing the likelihood function, which is an equation that expresses the probability of the data as a function of the unknown parameters [13]. Full Information Maximum Likelihood (FIML), which refers to ML estimation where missing data are present [22], utilizes the Expectation-Maximization (EM) algorithm to maximize the likelihood function [23]. In contrast to CCA, FIML uses all available information to estimate parameters and standard errors [15]. FIML is a convenient method for researchers as all estimation is done within the framework of a single model. FIML assumes multivariate normality and that data are MAR.

Multiple Imputation (MI)

Imputation refers to the 'filling-in' of missing values within a dataset, and Multiple Imputation (MI) involves many imputed datasets being produced. In general, MI is a three-step process:

1. Imputations are generated, usually based on one of the algorithms described below, resulting in the creation of multiple datasets

2. The scientific model of interest is estimated on *each* of the imputed datasets

3. Results from step 2 are combined according to Rubin's combination rules [24]

There are several options for imputation algorithms; two of the most common approaches are described below.

Markov Chain Monte Carlo (MCMC) The MCMC method is named after the algorithm used for generating imputed values. MCMC bases imputations only on observed data, meaning that imputed values are never used to impute other missing values [13]. MCMC requires that the imputation model be comprehensive, in that there is a joint probability distribution (multivariate normality) for all variables with missing data [13]. Therefore, in a situation where there is missing data on multiple variables that are not from the same probability distribution (e.g. a continuous variable and a categorical variable both with missing data), this assumption would be violated. However, some have suggested that this method is actually quite robust to departures from the multivariate normality assumption [25–27].

Fully Conditional Specification (FCS) Unlike MCMC, FCS (also known as sequential regression multiple imputation, or Multiple Imputation by Chained Equations (MICE)) does not require a joint probability distribution assumption, since each variable is imputed with separate regression equations [8]. This makes this method appealing for complex models with missing data on a variety of different non-normally distributed variables. Also unlike MCMC, FCS uses imputed values to impute other values through a program or user-specified hierarchy. Lastly, there is no mathematical guarantee of convergence with FCS, although this may not commonly lead to issues in practice [28]. Simulation studies have demonstrated that MCMC and FCS methods often yield similar results, and both are superior to CCA when there is a high fraction of missing data [29–31].

Auxiliary Variables

Whether using ML or MI methods to handle missing data, an important consideration is the use of auxiliary variables. Auxiliary variables are variables that are not included in an analysis model but are a part of the ML estimation or imputation model [15]. In the context of ML, auxiliary variables can reduce bias and increase power by regathering some information lost to missing variables and as such are always recommended for inclusion when available [15]. However, there are currently still computational and software limitations to incorporating auxiliary variables within ML.

For MI, auxiliary variables are necessary to improve the validity of the MAR assumption. The optimal choice of auxiliary variables are those which are associated with missingness in the variable(s) of interest. Allison [16] suggests that in order for auxiliary variables to be beneficial, their correlation with a missing variable needs to be relatively high at about 0.4 or greater. Notably, a good option for an auxiliary variable could be the same variable measured at a different time point. The presence or absence of good auxiliary variables should be determined in the exploratory phase of the analysis, as this will affect which model and associated assumptions will be most appropriate. In a dataset with few variables, auxiliary variable selection may be simple, since one can take the approach of including all available variables without much concern [32, 33]. However, in more complex datasets (e.g., multi-level, multiple time points, multiple variables within the same domain, etc.), this is not a feasible approach as it over-complicates the imputation model and can lead to issues with computation and collinearity [8]. In these situations, it is necessary to review current literature and perform exploratory data analyses to identify optimal variables for inclusion in an imputation model. Based on the exploratory analysis and literature review, iterative changes should be made to the model that are relevant and computationally feasible. This process may be challenging if there is little existing research on a topic, or if it is unclear which variables within a particular domain should be used.

1.2.3 Comparing Maximum Likelihood and Multiple Imputation

FIML and MI each have their own advantages and disadvantages as approaches to handle missing data. MI is more complex, requiring estimation of both an imputation model (i.e., the model used to impute the missing values), as well as the analysis model [13]. The time and effort required to understand and implement an MI procedure can be a barrier to its uptake. Because ML approaches are done within the analysis model, implementation is simpler. In turn, assessing model fit is also a simpler process compared to MI. In terms of efficiency, as long as sample sizes are sufficiently large and the procedures are implemented appropriately, FIML and MI approaches should produce very similar results [34].

Currently, a large downside to FIML is the lack of ubiquity in software to handle a variety of different scenarios, and the lack of consistency in how missing data are handled within and between programs. Many procedures don't allow for the implementation of FIML, and those that do may be inconsistent in their implementation. For example, in SAS, PROC CALIS (a general procedure for structural equation modelling) can implement FIML to handle missing data on outcome and predictor variables, while PROC MIXED or GLMMIX (which also do FIML but in the context of mixed models) will not account for data missing on predictor variables [22]. Specialized software (e.g., MPlus) is required in order to implemented FIML for categorical data or where missing data are assumed NMAR. It is also worth noting that the quality and quantity of details available with respect to missing data handling can differ substantially between softwares or even between procedures within a particular software. An additional disadvantage of the FIML approach is that iterative nature of the EM algorithm can translate to excessively long run-time for complex data or analyses. For example, data with a multi-level structure can take days or weeks to reach convergence on a single model.

Despite the complexity associated with MI, the fact that it is a more general approach that can be used in a variety of situations and can be implemented in most statistical software is a substantial advantage compared to FIML. In situations where there is missing data within predictor variables, auxiliary variables need be incorporated, or analyses are complex and may lead to computational difficulties, MI may be the more feasible approach. In theses circumstances, MI would be recommended over FIML.

1.2.4 The Status-Quo

Interestingly, methodologies that are considered more advanced are not necessarily new. MI and ML were both introduced in the mid-1970s and, although iterations have been made on these techniques since then, the core concepts have remained the same [12, 23]. Despite their age, these methods are still used sparingly in applied research. Systemic reviews of epidemiological studies have indicated that CCA is by far the most commonly used method [35, 36]. Moreover, there is an overall dearth of reporting on missing data in general: one review found that 20% of studies failed to report on the amount of missing data [36], and another found that nearly half failed to identify their assumptions about the type of missing data (MCAR, MAR, or NMAR) [35].

While it may be simple to criticize lack of uptake of sophisticated statistical methods for missing data, it is important to reflect on technological advances over the past couple decades. Present day researchers have a variety of options for software that can perform multiple imputation in a relatively uncomplicated way. Unlike researchers of the past, modern day researchers likely do not have to learn a new software to implement MI procedures, as they are now well integrated into popular analysis programs such as SAS, R, Stata, and SPSS. Notably, SAS only began integrated MI capability in 2002, with more rigid data assumptions and requirements compared to present day capabilities [37]. Moreover, present day researchers also have greater access to textbooks, research papers, and user-guides which make the knowledge required to use more advanced statistical techniques more accessible. As such, a present day researcher who wishes to use MI would not need to read Rubin's original statistically-dense publication when there are resources aiming to provide accessible frameworks for applied researchers which explain statistical concepts using applied language.

Nevertheless, there are great strides needed with respect to uptake of missing data methodologies, a point which has been echoed by numerous methodologists and researchers alike [11,17,19,38,39]. Many simulation studies have illustrated that CCA almost certainly produces bias results if data are MAR compared to ML or MI [18, 39–44]. However, simulation studies may not reach many applied researchers, and as such it is helpful to examine studies which have focused on how bias can impact research conclusions using real-world data. For example, in one such study, Becaria et al. [45] found that using MI identified two predictors of breast cancer that were missed with CCA, meaning that CCA introduced bias that manifested in the form of overlooked associations. The reverse is also possible; Ertel et al. [46] demonstrated where CCA found associations between perinatal depression and child BMI, MI found no such association.

1.2.5 Decision Trees and Missing Data

Although there is slow uptake in applied research of advanced missing data methods, these methods are constantly being improved. One of the ways in which missing data methods are progressing forward is through borrowed approaches from machine learning. Mostly notably, decision trees, or Classification and Regression Trees (CART), have been making their way into the missing data world. Decision trees are a type of prediction model which are advantageous because they are a visually straightforward and appealing way to display the results of a prediction algorithm. Decision trees begin at the root node, containing all data in the dataset of interest.

The data are then split into sub-nodes based on the variable of interest; in a regression tree, the variable is continuous, whereas in a classification tree, the variable is categorical [47]. Based off the all the available variables, the division that creates the most homogeneous sub-nodes is what determines the binary split. This process is recursive; splitting continues on each sub-node until some stopping criteria is reached (e.g., once there are less than 20 observations per group) [47]. Decision trees are particularly useful in situations where relationships are non-linear, or there are interactions present [48]. Decision trees have recently been used as an imputation algorithm, whereby instead of a typical regression based imputation approach, a decision tree is used. CART has even been incorporated into R's multivariate imputation package as a method that researchers can specify [49]. However, this is a relatively new approach with some identified limitations, and more research is needed on the use of this method [50–52].

Decision trees have also been used to help *understand* missing data; Tierny et al. [53] used CART to model the structure of missing data within a dataset (Figure 1.2). Their variable of interest was the proportion of missingness in the dataset; so, the tree was split based on what variables created the most homogeneous groups, in terms of degree of missing data. Through this technique, they discovered that their dataset had substantial missingness stemming from mismatched linking of individual smaller datasets. Despite Tierny et al.'s novel and illustrative use of this machine learning tool and the utility for understanding missing data, it does not appear this approach has been replicated or built on in the published literature to date.

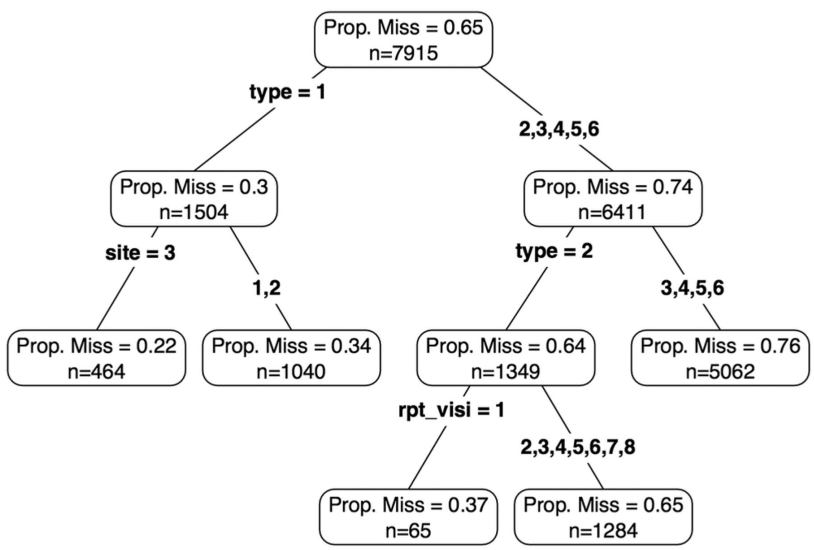


Figure 1.2: Figure from Tierney et al. [53] using CART to predict the structure of missingness in a medical dataset. The outcome variable is the proportion of missingness. The most important predictor of proportion of missingness was the type of data (type), shown through the first split. Data are further split illustrating the role of data collection site (site), and number of repeat visits (rpt_visi), on missingness.

1.3 Body Mass Index

Body Mass Index (BMI) is a widely used metric in public health as a population-based indicator of body adiposity. It is measured by dividing an individual's weight (kg) by their height (m^2), and is typically used to classify individuals as: underweight (>18.5), normal weight (18.5 - <25), overweight (25 - 30), or obese (>30). In the clinical context, BMI is objectively measured. However for a substantial portion of health research, height and weight are self-reported. BMI is missing if height, weight, or both are not reported. Self-report height and weight are unique in that they tend to be missing in high proportions that greatly exceed levels of missingness observed for other health indicators, particularly among youth populations [54, 55].

Global prevalence of individuals with Overweight or Obesity (OWOB) is rising, with little indication of improvement [56]. The proportion of Canadians with OWOB has reached an all-time high; 64% of adults and 30% of youth have OWOB [57,58]. OWOB is associated with many chronic diseases including cardiovascular diseases, type 2 diabetes, and cancer [59], as well as poorer mental and social wellbeing [60]. Estimates from 2010 indicated that Canada spends approximately $6 billion each year on direct healthcare costs for OWOB and related co-morbidities [61], and predictions have indicated that rising costs are likely to negatively impact the Canadian health care system [62]. Literature has indicated that OWOB trajectories are established at a young age, as the majority of youth with OWOB continue have OWOB during adulthood [63, 64]. Treating OWOB has proven difficult, and even when treatment begins during youth, programs are often difficult to implement and efficacy is inconsistent [65]. As such, there has been a public health shift in Canada from focus on treatment for OWOB, to a focus on addressing the upstream factors that lead to youth OWOB in order to inform prevention efforts [66, 67]. Many epidemiological studies have examined the environmental and behavioural factors associated with BMI among youth; a substantial portion of this research relies on survey based methods using self-report measures. Some of the main contributing factors to youth OWOB indicated in the literature are summarized below.

1.3.1 Correlates of youth BMI and OWOB

This section outlines some of the common factors that are related to OWOB among youth, but this is not an exhaustive list. The incredibly complex nature of OWOB means that the list of related physical, behaviour, and environmental factors is overwhelming, such that systems-thinking approaches only begin to encapsulate OWOB-related factors [68,69]. The list below focuses on evidence related to demographic and behavioural characteristics among youth.

Diet

It is well understood that dietary behaviours play a dominant role in health. For youth, numerous studies have found positive associations between OWOB and: lower fruit and vegetable intake [70, 71], lower frequency of breakfast intake [70, 72, 73], higher sugar-sweetened beverage consumption [74–76], and higher fast food consumption [77, 78]. Unfortunately, food environments (which play a large role in individual dietary patterns) are not typically supportive of positive behaviours [79–81]

Physical Activity, Sedentary Behaviour, and Sleep

Physical activity is parallel to diet in terms of known impact on physical health. Low physical activity [70, 82], and higher levels of sedentary behaviours (e.g., screen time) [82–85] have both been associated with OWOB among youth. Sleep duration and quality during adolescence have also been shown to be negatively associated with BMI both concurrently and over time [86–88].

Substance Use

Adolescence is a period where many youth experiment with substances. Several studies have indicated that alcohol consumption, cannabis use, and smoking are all associated with higher likelihood of OWOB [89–91]. While some mechanisms between BMI and substance use are unclear, associations with alcohol are likely due to the excess calories consumed, which is an often overlooked aspect of adolescent energy consumption [92]

Mental Health

Literature has begun to understand the implications of poor mental health during adolescence. Mental health indicators such as depression and anxiety have been associated with higher BMI, and research indicates that associations likely work in both directions [93–96]. Literature also indicates associations between OWOB and bullying victimization [89, 97].

Environmental Factors

Research has called for more emphasis on examining how the built environment influences health behaviours, including those that are associated with OWOB among youth [98, 99]. Positive aspects of the built environment have been associated with better OWOB outcomes in youth, such as the presence of parks and playgrounds [100, 101]. On the other hand, the presence of fast-food outlets has been associated with worse OWOB outcomes in youth [101, 102]. Other over-arching factors such as socioeconomic status are also associated with OWOB outcomes in youth [101, 103].

1.3.2 BMI and Missingness

Although BMI seems to be frequently used as a case study example throughout the missing data literature, very few studies have specifically examined the predictors of missingness for these self-report data. To date, it appears that only four studies have examined correlates of missing BMI or weight among youth [104–107]; these studies are summarized in Table 1.1. Based on the limited existing research, it seems that the missing data problem is more complex than lack of knowledge of these values, although lack of knowledge may be a factor for younger children [108]. This is supported by two of the four studies presented in Table 1.1 which found that younger age was associated with missing BMI values [105, 106].

For adolescents, there are other motivators in place influencing nonreporting of height and/or weight. It is well documented that there is substantial stigma surrounding weight, and that societal norms tend towards disapproval of individuals who have OWOB [109]. Moreover, concerns surrounding body image are heightened in adolescence [110–112]. The stigma surrounding body size is likely a contributor to the nonreporting of height and weight in youth survey-based research, and may be particularly salient for youth with OWOB. As such, youth with OWOB may be less likely to report their height and/or weight. Two of the four studies in Table 1.1 identified that poorer body satisfaction was related to missingness in BMI [106, 107], and three studies identified that poorer dietary and physical activity behaviours were associated with missing weight/BMI [104–106].

Given that height and weight data are missing for a variety of complex reasons, there are limitations to what can be done in terms of survey design to increase response rates, and as such it is essential to manage the missingness that occurs using appropriate statistical approaches. It is clear even from few existing studies that self-report BMI in youth is unlikely to be MCAR, which leaves either MAR or NMAR as plausible assumptions. Overall, research seems to indicate that missingness in BMI is likely NMAR, given that those with higher BMI may be less likely to report - although confirming this mechanism is not possible. The MAR assumption (which is more feasible to work with than NMAR) may still be valid if correlates of missingness can be appropriately incorporated into the statistical approach. However, given the sparsity of research on correlates of missing BMI, it is difficult to turn to literature for guidance.

A search of recent literature highlights that missingness in self-report BMI research remains poorly addressed, as it is not difficult to find relatively recent studies which fail to report the percentage of missing BMI data, methods used to address missing data, or both [113–116]. To further illustrate lack of missing data handling with respect to youth BMI literature, a review of the cited studies from Section 1.3.1 identifying youth OWOB risk factors (from diet to environmental factors, inclusive) finds that 11 studies used self-report data. Of these, only two reported the percent of missing BMI data [78, 102], and only one used a method other than CCA for the purpose of addressing missing BMI [102].

Table 1.1: Pre-existing research examining missing correlates of youth self-report weight or BMI.

Study	Sample (Ages)	% Missing	Correlates
Self-reported weight and predictors of missing responses in youth *Aceves-Martins et al. (2018)*	11,13,15	58.9% (weight)	• Low physical activity • Low fruits & vegetable consumption • High computer gaming
Learning from Non-Reported Data: Interpreting Missing Body Mass Index Values in Young Children *Arbour-Nicitopoulos & Faulkner & Leatherdale (2010)*	9-14	45.4% (BMI)	• Younger age • Female gender • Non-white ethnicity • Non-participation in sports • Low academic standing • Low energy expenditure
Emotional, behavioural and social correlates of missing values for BMI *Fonseca et al. (2008)*	11-16	10.8% (BMI)	• Younger age • Low physical activity • Poor body satisfaction • Absence of a father • Absence of opposite-sex friends • Poor perceived academic achievement
Self-reported weight and predictors of missing responses in youth *Tiggemann (2006)*	13.2-16.1	28.10% (BMI)	• Poor current figure rating • Poor weight & figure satisfaction • Increased drive for thinness

1.3.3 Validity of BMI

Given the wide use of BMI as an indicator of body adiposity across a variety of settings, the validity of the measure has rightfully been questioned. BMI cannot distinguish between lean body mass (i.e. muscle) and fat mass. Nevertheless, the underlying assumption when using BMI is that it is a reasonable indicator of adiposity, and that increases in BMI are representative of increases in body fat. Research has indicated that BMI as an indicator of adiposity has high specificity, but low sensitivity, for both adults [117], and children [118, 119]. High specificity indicates that the majority of OWOB classifications are correct, such that those identified as OWOB by BMI do have relevant excess adiposity. However, poor sensitivity indicates that there are many adults and children who have excess adiposity but are identified as normal by BMI criteria; a situation termed 'normal-weight obesity' in the literature [118]. A visual overview of specificity and sensitivity of BMI is given in Figure 1.3.

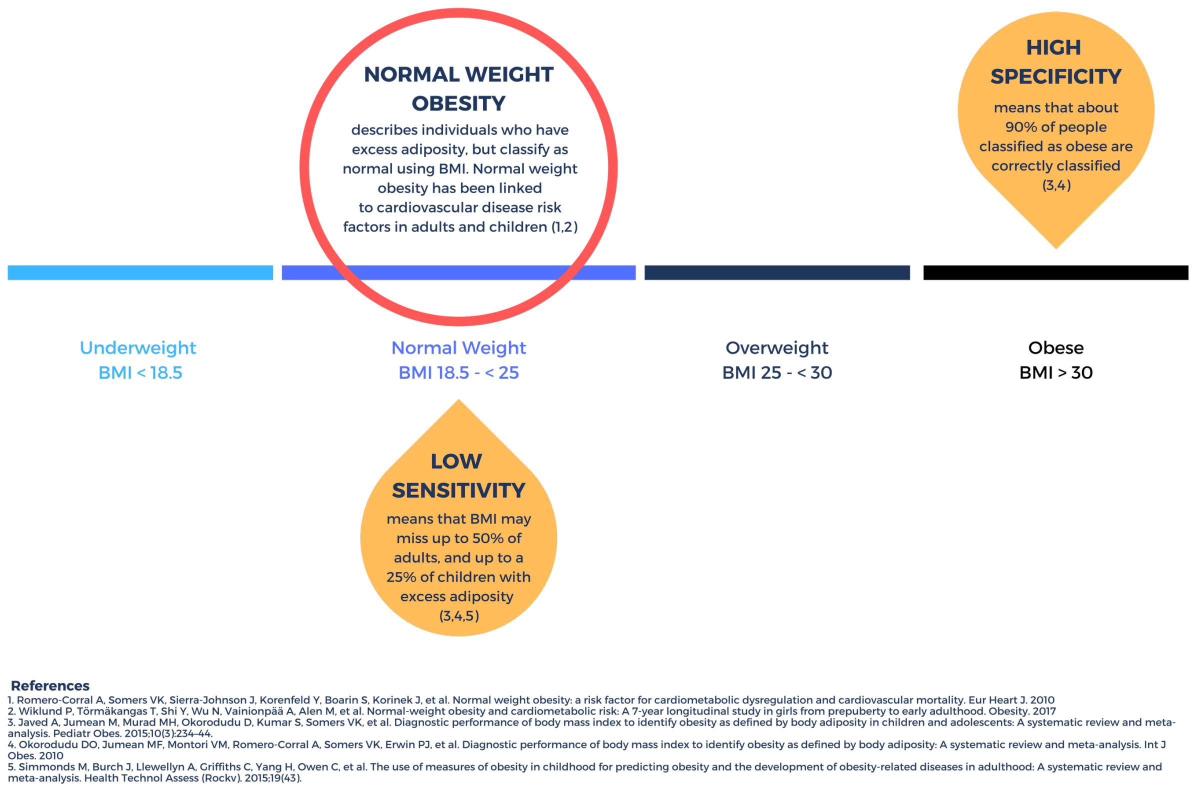


Figure 1.3: BMI as a Measure of Adiposity: Sensitivity, Specificity, and Normal Weight Obesity

Despite the fact that childhood obesity (identified by BMI) is linked to poor health outcomes in adulthood, including diabetes, coronary heart disease, and certain cancers, the majority adult occurrences of these conditions are in people who were *not* classified as OWOB as children [119]. This, coupled with research directly measuring cardiometabolic risk factors in adults and children with normal-weight obesity [120, 121], demonstrates that the real-world consequences of poor sensitivity are that a large proportion of individuals may be at risk for chronic disease, yet are not identified as such.

Another criticism of BMI is its accuracy when self-reported. Certain factors such as social desirability bias or lack of knowledge of measurements may cause individuals to under-report or over-report their height and/or weight. In terms of BMI, research suggests that weight is usually under-reported and height is usually over-reported [55, 122]. For adolescents, differences in reporting can vary by sex; a 2007 systematic review on this topic reports that females tend to under-report their weight by up to 4kg, while males tends to under-report by up to 2.6kg [55]. A more recent review found a more conservative range; under-reporting weight for US adolescents was in the 0.6kg - 1.7kg range (and overweight individuals were more likely to misreport), while height was over-reported by 1.0cm - 1.7cm [123]. Notably, children tend to less accurately report their weight and height compared to adolescents [108], likely related to lack of knowledge of these measurements. While inaccurate reporting is certainly a limitation of self-reported BMI, research indicates that reported numbers are not drastically different from their true values. The relatively consistent and smaller-scale misreporting that does exist is a limitation that need be kept in mind when the data are used, but does not necessarily indicate that BMI is an invalid measure.

Despite poor sensitivity, and some issues with misreporting, BMI remains the most feasible measure to assess OWOB in the population. Researchers generally agree that any measure that may be more accurate, or yield more information, is not feasible at a population level without extensive resources or compromising sample size [108, 119]. For large surveillance studies, and certainly research that is used for government surveillance of public health, BMI remains the only feasible option until more robust alternatives become cheaper and easier while maintaining precision. Meanwhile, it is important for researchers and knowledge-users to be aware of the limitations to the use of BMI: since it does not distinguish between types of body mass, misclassification is possible, and someone with "normal weight obesity" can be classified as healthy, or someone with a BMI greater than 25 but no cardio-metabolic risk factors can be classified as unhealthy. At the population level, the consequence of low-sensitivity compounded with any under-reporting of weight and/or over-reporting of height is that the true proportion of those with excess adiposity and associated health risks are underestimated.

Chapter 2

Research Questions and Rationale

The aim of this research was to improve understanding of the levels, patterns, and impacts of missingness in youth Body Mass Index (BMI) for survey-based research. This dissertation is divided into three separate studies, presented as manuscripts for academic publication. The results of each study were intended to inform the direction of the studies that followed.

2.1 Study 1: Learning from missing data: examining nonreporting patterns of height, weight, and BMI among Canadian youth

2.1.1 Rationale

Youth self-report BMI is frequently missing in high proportions, up to 59% [104]. It appears there are a combination of factors, including knowledge of these values and social stigma surrounding body image, which may influence youth not to report height or weight (or both). However, to date, it appears that only four studies have focused on the examination of predictors of missing youth self-report weight or BMI (see Table 1.1) [104–107]. These studies were limited in sample size as well as the variety of factors analyzed. The research objective of Study 1 was to gain in-depth insight into the predictors of missing youth BMI, adding to the limited existing research on this topic using a large sample size and a wide breadth of health behaviour variables.

There are several reasons why it is necessary to understand the characteristics and health behaviours of youth who do not report height and/or weight data. First, these findings can allow for researchers who use youth BMI from self-report measures to appropriately comment on how biases from missing data may impact study findings and interpretations. Next, advanced missing data approaches require some information on predictors of missingness in order to be properly implemented. Lastly, understanding what predicts missingness in youth BMI is essential to inform future survey methods; potential areas where likelihood of nonreporting can be reduced through modification of a survey or survey method is preferred over post-hoc handling of missing data.

2.1.2 Research Questions

1. What factors are identified as significantly contributing to missingness in youth BMI?
   1.1. What factors are identified as significantly contributing to missingness in youth height and weight, and how do these compare to the factors associated with missing BMI?
   1.2. How do the models for missing BMI, height, and weight, differ between females and males?

2.2 Study 2: Using classification and regression trees to model missingness in youth BMI, height, and body mass

2.2.1 Rationale

The complexity of the factors which influence youth BMI missingness are not fully captured by standard regression techniques used in existing research thus far [104–107]. Decision trees, a method borrowed from machine learning, are capable of identify complex structures and patterns within data. A novel use of decision trees was presented by Tierney et al. [53], who demonstrated the utility of using Classification and Regression Trees (CART) models to discern the structure of missing data. The research objective of Study 2 was to identify the structure of missingness in self-report youth BMI, height, and weight using CART models, as well as identify important subgroups of individuals with high propensity towards missing data.

The findings of this study have potential to contribute to both applied research as well as research methods literature in a few ways. First, using the CART approach can reveal previously uncaptured findings pertaining to the structure and predictors of missing youth BMI, height, and weight data, providing greater insight into youth reporting behaviours. Next, in some ways a CART model may be more useful (compared to parametric approaches) at identifying a parsimonious list of variables highly related to missing BMI, which may be particularly useful during the auxiliary variable selection required for application of advanced missing data methods. Lastly, given that using a CART model as opposed to standard parametric approaches can have added utility for examining missingness, future research may leverage the approach from this study as framework to examine variable-level missingness.

2.2.2 Research Questions

1. What combinations of factors are identified as contributing to missingness in youth BMI, height, and weight when missingness is modelled using CART?

    1.1. How do the CART models for missing BMI, height, and weight, differ between females and males?

    1.2. How do these identified predictors differ from those identified using the more traditional parametric approaches from Study 1?

2.3 Study 3: Assessing the impact of missing data in youth overweight and obesity research: complete case analysis versus multiple imputation

2.3.1 Rationale

Varying levels of missing data in youth BMI may bias the results of research based on self-report variables if appropriate methods are not used. Several simulation and applied studies have highlighted the bias that can be introduced when Complete Case Analysis (CCA) is used in situations with high levels of non-random missingness [18, 39–44], but to date it appears none have focused on youth BMI. If differing methods alter research results for youth BMI through differences in significance, magnitude, or direction of predictors, this will certainly impact potential research conclusions. As such, the objective of Study 3 was to compare the research conclusions from an analytical model predicting youth BMI where different missing data methodologies were used. More specifically, Study 3 compared the use of CCA and Multiple Imputation (MI) in separate models which contained the same data and predictors, only differing in their approach to missing data handling. The results and conclusions were compared across approaches, in order to identify the impact that missing data handling can have on research findings and recommendations for public health.

Study 3 contributions are at the intersection of methodology and practice. Through identifying the differences in research conclusions and recommendations that result from differing missing data methodologies, the importance of methods selection is illustrated. Moreover, this approach also provides greater context for those examining and evaluating the quality of current literature. Lastly, findings also allow for future research to better comment on the potential bias, and impact of that bias, that may stem from missingness in youth BMI for survey-based research.

2.3.2 Research Questions

1. How do the results of sex-stratified analytical models examining factors associated with youth BMI compare between CCA and MI approaches to handle missing data?

1.1. How do the significance of included factors, as well as the magnitude and direction of point estimates, compare between the two approaches?
1.2. What statistical inferences might be made from each respective analysis, and what implications might this have on the interpretation of findings?

Chapter 3

Methods

This section describes the analytical approach of each study, all of which use existing data from the COMPASS study. The COMPASS study (here out referred to as COMPASS) is a prospective cohort study of youth in Canada. COMPASS collects self-report data from students in grades 9-12 in Ontario, Alberta, British Columbia, and grades 7-11 in Quebec. An overview of COMPASS study features relevant to the work presented in this dissertation are described below, but a full description is published elsewhere [124], and is also available online www.compass.uwaterloo.ca.

3.1 The COMPASS Study

3.1.1 Research Funding and Ethics

COMPASS was initiated in 2012/13 and to date is funded until 2027. Funding details can be found in Appendix A. Ethical approval was obtained through the University of Waterloo Office of Research Ethics (reference #30118) as well as participating school boards.

3.1.2 Data Collection Overview

COMPASS uses school-based data collection and convenience sampling; as such, sample sizes may vary between yearly data collections. Schools are eligible to participate in COMPASS if they are in one of the aforementioned data collection provinces, have students in grades 9 through 12 (or 7-11 in Quebec) with at least 100 students per grade, and operate under a standard classroom setting. COMPASS schools must also allow for an active-information passive consent protocol, which operates an opt-out bases rather than opt-in. As such, parents of attending students are informed of the COMPASS study and asked to contact the school if they do not want their child/children to participate. Of course, students are also able to decline to participate on the day of the data collection as participation is voluntary. The use of this passive consent protocol is important for youth research, as it yields high participation and is less prone to response biases. This protocol has been proven to be especially important for health behaviours tied to social desirability, such as substance use [6, 7]. Notably, student questionnaires are anonymous and students are made aware of the anonymity prior to completing the survey.

3.1.3 Student Questionnaire

During 2018/19, the COMPASS Student Questionnaire (Cq) was administered to students during scheduled class time. In absence of extenuating circumstances, all participating students within a school completed the Cq at the same time. The Cq is a 15 page Scantron-style survey, which would take students on average 40 minutes to complete. Teachers administered the surveys, and were provided with detailed instructions including a script to read to students and a series of answers to common questions. Upon completing the Cq, students would put it into an unlabelled brown envelope. Individual envelopes were placed into a larger envelop and walked to the main office where the data collector received them. After arriving at the University of Waterloo, Cqs were visually scanned by staff, and subsequently machine scanned for data input. Machine scanning was monitored by staff for potential errors.

The Cq was divided into sections which focus on: demographics, diet and physical activity behaviours, substance use, and school related measures (e.g. math grade, educational aspirations, etc). In the 2016/17 school year, an additional mental health section was piloted among a sample of participating schools, and in 2017/18 the mental health section was implemented in all participating schools [125, 126]. Several changes were made to the Cq for the 2019/20 school year, but are not discussed here as this dissertation only uses data up to 2018/19. The full 2018/19 Cq can be found in Appendix B.

3.1.4 Sample

The **2018/19 cross-sectional sample** was used for all studies outlined in this dissertation. The same sample was used for all studies because each study was designed to build off of, and be informed by, the findings of the previous. The 2018/19 year sample was chosen because it was the most recently available dataset when this dissertation was initiated. Despite availability of a longitudinal sample, a cross sectional sample was chosen given the current lack of research in youth Body Mass Index (BMI) missingness, analytical benefits from a large sample size, and limitations surrounding feasibility of three-level approaches for the chosen methods (namely Classification and Regression Trees (CART) models and Multiple Imputation (MI) procedures).

3.1.5 Measures

The initial measures list for this dissertation was relatively extensive. This is for several reasons, including the breadth of variables available through the COMPASS study, the variety of variables that have been previously associated with youth BMI, and the exploratory nature of studies 1 and 2. Initially, all variables that could be feasibly related to BMI, height, or weight missingness were included. The methods for parsing down this list of initial variables differ between studies, and details can be found in Section 3.3. If variables were used as written in the Cq, they were not re-written below, and instead can be referred to in Appendix B. Where this is the case, the question number (identified with 'Q#') is indicated alongside the variable. For easy location, variables throughout this section are identified in **blue**.

BMI

BMI was determined by dividing weight (kg) by height (m^2). **Weight** was determined by asking students, "how much do you weigh without your shoes on? (please write your answer in pounds OR in kilograms, and then fill in the appropriate numbers for your weight.)" There was a space for students to write out their weight, as well as separate Scantron bubbles for them to fill in. There was an option for students to report, "I do not know how much I weigh." **Height** was similarly determined by asking, "how tall are you without your shoes on? (please write your height in feet and inches OR in centimeters, and then fill in the appropriate numbers for your height)." Examples for weight and height were given to demonstrate how to fill out the Scantron portion. There was an option for students to report "I do not know how tall I am." Inconsistent numbers between the open-ended answer space and Scantron bubbling were corrected during visual scanning, and unfeasible numbers (less than 45lbs or greater than 390lbs for weight, less than 4ft or greater than 6ft 11 for height) were marked as missing.

Demographics

Sex (Q3) and **Age** (Q2) were both used as written in the Cq. Grade and age are highly collinear and it would have been redundant to include both in analyses. As age has been identified as a predictor of missingness in previous studies, and it is more conceptually transferable to other populations, it was used in this research. **Ethnicity** was re categorized from multi-option question in the Cq. The Cq asked "How would you describe yourself? *(Mark all that apply)*" and the options were "White", "Black", "Asian", "Aboriginal (First Nations, Métis, Inuit)", "Latin American/Hispanic", "Other". Any option other than "White" (including anyone who selected multiple) was categorized as "racialized", where white was categorized as "non-racialized". Ethnicity was collapsed to binary for both computational reasons (due to small cell counts), as well as to ensure that statistical inference was appropriate.

Body Image and Weight Intentions

In addition to self-reported weight, youth were asked about their **weight perception** and **weight goals**. Youth were asked "How do you describe your weight?" where options were "Very underweight", "Slightly underweight", "About the right weight", "Slightly overweight", and "Very overweight". This measure was modified such that the "Very" and "Slightly" categories for both under and over weight were collapsed. Weight goals were assesed through the question "Which of the following are you trying to do about your weight?" where the options were "**Lose** weight", "**Gain** weight", "**Stay** the same weight", and "I am **not trying to do anything** about my weight". The latter two options were collapsed into a single category. Importantly, neither of these variables were considered a direct proxy for weight, since youth can misclassify themselves [127, 128].

Dietary Behaviours

The following 24-hour dietary recall variables were used as written in the Cq found in Appendix B:

- **Servings of meat and alternatives** (Q26)
- **Servings of fruit and vegetables** (Q27)
- **Servings of milk and alternatives** (Q28)
- **Servings of grain products** (Q29)

Notably, the 24-hour dietary recall questions pertain to the previous version of Canada's Food Guide established in 2007 [129]. Although there is now an updated version of the Food Guide [130], these measures remain relevant for research purposes, particularly given that no substantial guidance surrounding how these new guidelines should be assessed have been identified.

The Cq asked, "on how many days do you do the following?" for a number of dietary behaviours across weekdays ("in a usual school week (Monday to Friday)") (Q24)) and weekends ("on a usual weekend (Saturday and Sunday)" (Q25)). **Breakfast consumption** per week was obtained by aggregating across the two questions (see Q24a and 25a in Appendix B) to create a continuous indicator. A measure of **energy drink consumption** was similarly created by aggregating across the two questions (see Q24i and 25f in Appendix B). **Fast food consumption** was asked slightly differently between the two questions, whereby the weekday question read, "eat lunch purchased at a fast food place or restaurant," and the weekend question read, "eat foods purchased at a fast food place or restaurant." These questions were aggregated to create a continuous indicator of weekly fast-food consumption, keeping in mind that the weekday question specifically refers to lunch, and this is a limitation of the use of this variable.

Movement Behaviours

Consistent with existing research, the main indicator of physical activity was a continuous variable indicating time in minutes of **Moderate to Vigorous Physical Ac-**

tivity (MVPA). Vigorous physical activity was reported through the question, "mark how many minutes of HARD physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time." The Scantron format allowed students to indicate number of hours (0,1,2,3,4) and minutes (0,15,30,45) for each day of the week. Moderate physical activity was reported through the question "Mark how many minutes of MODERATE physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time. Do not include time spent doing hard physical activities." Scantron formatting was the same as for vigorous physical activity; examples for both questions were given demonstrating how to fill out these Scantron questions. Examples of "hard" and "moderate" activities were also given.

Sports participation was reported through three questions. The first asked, "do you participate in before-school, noon hour, or after-school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)." The second asked, "do you participate in competitive school sports teams that compete against other schools? (e.g., junior varsity or varsity sports)." The third asked "Do you participate in league or team sports outside of school?" The response options for each of the three questions were "Yes", "No", and a third option indicating if there were none offered at their school or available where they live. All three questions were aggregated into one measure of binary (i.e., yes, no) sports participation.

Indicators related to physical activity that were used as written in the Cq include:

- **Strength training** (Q22)
- **Physical activity of friends** (Q17)

The Cq asked students to indicate the collective number of hours (0-9) and minutes (0-15) spent on sedentary activities through asking, "how much time per day do you usually spend doing the following activities?" A continuous measure of total **Screen Time Sedentary Behaviour (STSB)** was derived by adding together reported time for: "watching/streaming TV shows or movies," "playing video/computer games," "surfing the internet," "texting, messaging, emailing (note: 50 texts = 30 minutes)." Values of STSB greater than 16.25 hours were considered outliers and reduced to missing. Number of hours and minutes of **sleep** (Q12g) was reported in the same section and manner as sedentary behaviours and was included as a continuous indicator. Values of sleep below 4 hours were considered outliers and reduced to missing. Details on how the STSB and sleep outlier cutoffs were chosen are given in Section 3.2.

Academic Indicators

English grade and **math grade** were two of the academic indicators considered. Math grade was self-reported and asked through the question, "in your current of most recent Math course, what is your approximate overall mark? (Think about last year if you have not taken math this year)" where the options were: "90%-100%," "80%-89%," "70%-79%," "60%-69%," "55%-59%," "50%-54%" and, "Less than 50%." The question was asked in the same format with the same response options for English grade. For both subjects, options were re-grouped into: less than 50% and greater than or equal to 50%.

Truancy was the third academic indicator considered, and was asked in the Cq through the question, "in the last 4 weeks, how many classes did you skip when you were not supposed to?" where the options were: "0 classes," "1 or 2 classes," "3 to 5 classes," "6 to 10 classes," "11 to 20 classes" and, "More than 20 classes." This variable was regrouped to be binary, such that the first option was considered no classes skipped, and all other options were considered 1+ classes skipped.

Mental Health

The Cq included several measures of mental health and wellbeing for population research purposes. The rationale for the use of each measure in the context of youth mental health, including rationale for the chosen scales, is described elsewhere [125]. As such, each measure is described here briefly. For each of the measures below that are aggregated to create a summed score representing a scale, if one item was missing, the entire aggregate score was set to missing. **Self-rated mental health** was included as written in the Cq (Q59).

The Cq used the Center for Epidemiologic Studies Depression Scale Revised (CESD-R-10) [131] as measure of **depression** among youth populations. Individuals were asked to report "on how many of the last 7 days did you feel the following ways?" where the response options were: "None or less than 1 day," "1-2 days," "3-4 days" and, "5-7 days" for the following 10 items: "I was bothered by things that usually don't bother me," "I had trouble keeping my mind on what I was doing," "I felt depressed," "I felt that everything I did was an effort," "I felt hopeful about the future," "I felt fearful," "My sleep was restless," "I was happy," "I felt lonely" and, "I could not get 'going'". Scores across the items were summed such that "none or less than 1 day" was 0, and "5-7 days" was 3. The two positively framed items ("I felt hopeful about the future" and "I was happy") were reversed scored. A score of 10 or higher is consistent with clinically relevant symptoms of depression.

The Cq used the Generalized Anxiety Disorder 7-item (GAD-7) scale [132] to identify clinically relevant symptoms of generalized **anxiety** and anxiety disorders. Individuals were asked to report "Over the last 2 weeks, how often have you been bothered by the following problems?" where the response options were "not all all," "several days," "over half the days," or "nearly every day" for the following 7 items: "feeling nervous, anxious, or on edge," "not being able to stop or control worrying," "worrying too much about different things," "trouble relaxing," "being so restless that it is hard to sit still," "becoming easily annoyed or irritable" and, "feeling afraid as if something awful might happen". Scores across the items were summed such that "not at all" was 0, and "nearly every day" was 3. A score of 10 or higher is consistent with clinically relevant anxiety symptoms.

The Cq used the Difficulties in Emotional Regulation Scale (DERS) [133] as an indicator of **socio-emotional skills**. Individuals were asked to "please indicate how often the following statements apply to you" where the response options were a 5-item Likert scale of "almost never" to "almost always" for the following statements: "I have difficulty making sense out of my feelings", "I pay attention to how I feel", "when I'm upset, I have difficulty concentrating," "when I'm upset, I believe there is nothing I can do to make myself feel better," "when I'm upset, I lost control over my behaviour," "when I'm upset, I feel ashamed for feeling that way." Scores across the items were summed such that "almost never" was 1, and "almost always" was 5. The positively framed item ("I pay attention to how I feel") was reverse scored.

The Cq used the Flourishing Scale [134] as a an indicator of **mental wellbeing**. Individuals were asked to report "how much do you agree with the following statements?" ranked on a 5-item Likert scale of "strongly agree" to "strongly disagree" to the following 8 items: "I lead a purposeful and meaningful life," "my social relationships are supportive and rewarding," "I am engaged and interested in my daily activities," "I actively contribute to the happiness and well-being of others," "I am competent and capable in the activities that are important to me," "I am a good person and live a good life," "I am optimistic about my future," and "people respect me." The total score of these 8-items was summed ("strongly agree" = 5, "strongly disagree" = 1) to create an overall Flourishing score, such that higher scores indicated better mental wellbeing.

The Cq used items from the Self Description Questionnaire II [135] to form an indicator of **self-concept**. Individuals were asked to "choose the answer that best describes how you feel" ranked on a 5-item Likert scale of "true" to "false" to the following 5 items: "in general, I like the way I am," "overall, I have a lot to be proud of," "a lot of things about me are good," "when I do something, I do it well," and "I like the way I look." The total score of these 5-items was summed ("True" = 1, "False" = 5) to create an overall self-concept score, such that higher scores indicated worse self-concept. Previous analyses have indicated that summing these scores is appropriate [126].

The Cq asked about both **bullying victimization** as well as **bullying perpetration**. Bullying victimization was reported through the question "in the last 30 days, how often have you been bullied by other students?" where the options were: "I have not been bullied by other students in the last 30 days," "less than once a week," "about once a week," "2 or 3 times a week", and "daily or almost daily." Bullying perpetration similarly asked "in the last 30 days, how often have you taken part in bullying other students?" where there response options were the same. Both variables were transformed into binary indicators whereby all options except the first were counted as a "yes."

Substance Use

To align with previous work on substance use and BMI [136], as well as the current substance use literature [137–140], binary measures of "current use" vs. "no current use" were created for each of the four measured substance use behaviours.

The Cq asked, "in the last 12 months, how often did you have 5 drinks of alcohol or more on one occasion?" where the options were, "I have never done this," "I did not have

5 or more drinks on one occasion in the last 12 months," "less than once a month," "once a month," "2 to 3 times a month," "once a week," "2 to 5 times a week," and "daily or almost daily." The binary measure of **current binge drinking** included anyone who reported binge drinking in the last 12 months.

The Cq also asked "in the last 12 months, have you had alcohol mixed or pre-mixed with an energy drink (such as Red Bull, Rock Star, Monster, or another brand)?" where the options were, "I have never done this," "I did not do this in the last 12 months," "Yes" and, "I do not know". A binary measure of **Alcohol Mixed with Energy Drinks (AmED)** use was created whereby all options except "Yes" were considered a "No."

The Cq asked "in the last 12 months, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)" where the options were "I have never used marijuana," "I have used marijuana but not in the last 12 months," "less than once a month," "once a month," "2 or 3 times a week," "4 to 6 times a week," and "every day." The binary measure of **current cannabis use** included anyone who reported using cannabis at least once a month.

The Cq asked "on how many of the last 30 days did you smoke one or more cigarettes?" where the response options were "none," "1 day," "2 to 3 days," "4 to 5 days," "6 to 10 days," "11 to 20 days," "21 to 29 days," "30 days (every day)." The binary measure of **current smoking** included anyone who reported smoking cigarettes at least once in the past 30 days.

The Cq asked "on how many of the last 30 days did you use an e- cigarette?" where the response option are "none," "1 day," "2 to 3 days," "4 to 5 days," "6 to 10 days," "11 to 20 days," "21 to 29 days" and, "30 days (every day)." The binary measure of **current e-cigarette use** included anyone who reported using e-cigarettes at least once in the past 30 days.

3.2 Data Preparation

Exploratory Data Analysis (EDA) included producing summary statistics, histograms (for continuous variables) and bar charts (for categorical variables). Based on the results of EDA, several variables which originally had many categories were collapsed to fewer categories, as outlined in Section 3.1.5. Most variables were collapsed to binary because of low frequency for some of the response options; re-categorizing these variables improved cell counts and allowed for clearer statistical inference. During EDA, continuous variables were examined for skewness. The sleep variable was found to left-skewed, while the STSB variables was found to be right-skewed. In the case of sleep, this was problematic because several unrealistically low (and therefore biologically unfeasible) values were reported. While these values might have been feasible if the sleep question asked about hours of sleep on the *previous night*, since the question asked youth to report their average hours *per night*, unrealistically low values were conceptually inappropriate. Similarly, some unrealistically high values were reported for STSB, which may have been due to misestimation by participants, or double-reporting of simultaneous activities. For both sleep and screen time, values were considered outliers if they fell outside 1.5x the interquartile range. As such, sleep values less than 4 (average hours per night) and STSB greater than 16.25 (average hours per day) were reduced to a missing value for sleep or STSB respectively. Importantly, height and weight (and thus BMI) outliers had already been removed upon receiving the dataset, as outlined in Section 3.1.5.

3.3 Analyses

3.3.1 Study 1

Descriptive statistics included examining the missingness levels (i.e., proportion of missing data) for all variables. A histogram and descriptive statistics table was used to explore and display missingness across the data. A stacked bar graph was used to explore and display missingness within BMI by sex.

Of the previous studies which have examined youth BMI missingness, the most recent used the approach of initial chi-square tests for independence between missing BMI and all hypothesized variables, and a subsequent logistic regression model based on the variables identified as significant [104]. For the COMPASS sample, using chi-square tests to identify variables which belong in a regression model would be inefficient; the large sample size and large portion of missing data leads to many separate significant associations and parses few variables. This is problematic because a model with too many explanatory variables can impact statistical inference. In order to address this challenge, this study used model selection criteria Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) within the context of a model selection framework. Using a model selection framework can provide a more objective approach to variable selection, and may be particularly useful in situations where there are many candidate variables. Model selection frameworks parse through each possible model and identify which models performs most optimally in terms of parsimony, goodness-of-fit, and generalizability [141]. For p predictor variables, the total number of candidate models arising from distinct combinations of variables is $2^p - 1$. Listed in Section 3.1.5, a total of 34 variables were selected from COMPASS that were potentially related to BMI missingness, either as hypothesized variables or variables that had been related to BMI missingness in previous literature. However $2^{34}-1$ yields over a billion possible models. Of course, parsing through all of these candidate models would be inefficient for a few reasons. First, the variables belong to distinct conceptual domains (which they are sorted by in Section 3.1.5), but model selection criteria does not take this into account. As such, an "optimal" model could be identified that is conceptually unrealistic. For example, applying a model selection framework to all 34 variables could result in an "optimal" model being identified that only contains mental health variables, but does not control for an individuals age or ethnicity. The second reason this approach would be inefficient is that it is computationally and time intensive to parse through a billion candidate models, in addition to the sample size being under-powered to allow model selection criteria (such as BIC [142]) to select models with large number of predictors [143, 144].

In order to address these concerns, the model selection algorithm used in this study was modified slightly to non-optionally include the control variables age, ethnicity, weight perception, and weight goal in all models. Next, variables were sorted into conceptual blocks which included: diet, movement, academic, mental health, and substance use (see Section 3.1.5 for which variables are included in each block). The modified model selection framework was applied to each block separately. The default inclusion of controls as well as sorting the variables into blocks was conceptually and computationally more efficient than applying the framework to the entire list of variables.

The model selection algorithm used in this study was developed by Ten Eyck and Cavanagh [141], who improved upon the existing pseudo-liklihood model selection process that is implemented for Generalized Linear Mixed Model (GLMM) in SAS's PROC GLMMIX function. The code to implement this improved model selection framework was obtained from author Patrick Ten Eyck in the form of a SAS macro. In order to select a model within each conceptual block, the most parsimonious combination of variables that had a comparably low AIC and BIC to other possible models was selected. Each aspect of this described approach was performed for females and males separately.

Lastly, the variables chosen by the model selection framework from each conceptual block were combined into sex-stratified GLMMs in order to examine associations with missing BMI, height, and weight. All analyses were conducted in SAS 9.4. A GLMM approach was chosen in order to account for clustering at the school level. The general form for a random-intercept multi-level model which accounts for spatial clustering is given in composite form by:

$$Y_{ij} = \gamma_{00} + \boldsymbol{\gamma_{01}} Z_j + \boldsymbol{\gamma_{10}} X_{ij} + \boldsymbol{\gamma_{11}} Z_j X_{ij} + \mu_{0j} + \epsilon_{ij} \tag{3.1}$$

In the context of students clustered within schools, in equation 3.1 $j = 1, ..., J$ schools, $i = 1, ..., I$ individuals, $X_{ij} = x_1, ..., x_p$, and $Z_j = z_1, ..., z_p$. In other words, the subscript "ij" represents the i^{ith} individual from the j^{ith} school, X is a matrix containing all individual-level predictors, and Z is a matrix containing all school-level predictors. $\boldsymbol{\gamma}$ represents the regression coefficients, which are fixed and unknown. The stochastic components of equation 3.1 are given by ε_{ij} and μ_{0j}, representing the residual error and the random residual error at the school level, respectively. Assumptions of the stochastic components are $\varepsilon \overset{iid}{\sim} N(0, \sigma_e^2)$, $\mu_{0j} \overset{iid}{\sim} N(0, \sigma_0^2)$, where $\boldsymbol{\epsilon} \perp \boldsymbol{\mu}$. Lastly, Y_{ij} denotes the outcome; in a Linear Mixed Model (LMM) the outcome would be continuous, whereas in a GLMM, the outcome is categorical. Study 1 used a GLMM as the outcome of interest was binary (missing vs. not missing). Notably, for completeness Equation 3.1 is written to include school-level variables as well as an interaction term between school and student-level variables, however no such effects were included in Study 1. Results of the GLMMs were presented in the form of odds ratios, and associations were examined for significance using a 95% Confidence Interval (CI).

3.3.2 Study 2

Study 2 built off of Tierny et al.'s approach [53] of modelling missing data using CART, using variable-level missingness in BMI, height, and weight as the outcome for each model. As the outcomes for all models were binary (either missing or not missing) classification trees were used; these are referred to throughout as the "CART" model(s) to align with typical terminology. For each outcome, two models were produced as analyses were sex-stratified, and as such, all steps described for this study below were also sex-stratified.

CART models initially included all variables mentioned in Section 3.1.5. Similar to Study 1, variable selection was systematic as the CART algorithm recursively splits data based on which variables predict the most homogeneous split into binary groups. In contrast to Study 1, variable selection was simpler and did not require variables be sorted into blocks, as CART models examine distinct subgroups within the data as opposed to regression approaches which focus on testing associations. Importantly, height nor weight was included in the CART model for missing BMI (and vise versa), as this would be redundant since the model would have certainly choose these variables as primary splits. In order to determine splits, CART models use the GINI index. The GINI index is a measure of impurity or uncertainty, and is calculated by subtracting the sum of squared probabilities for each class from 1, as outlined in equation 3.2, where p_i is the probability of being in a particular class.

$$GINI = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{3.2}$$

Following standard approaches for CART models [145], prior to analysis data were randomly divided into training data (80% of the sample) and testing data (20% of the sample). Models were created with the training data subset, and the testing subset was used to determine the accuracy of the created models. All CART models were run in R using the rpart package [146]. CART models were run with one pre-pruning restriction, which was that the size of terminal node be no smaller than the number of individuals within the smallest school. This was a cut-off of 14 individuals in the female models and 16 in the male models. Due to their overfitting nature, CART models require pruning; models were pruned using cost-complexity pruning and the 1-SE guideline [145]. Final pruned trees were presented visually.

Current standard CART modelling approaches do not account for hierarchy within data. Given the multilevel nature of the COMPASS study, best attempts were made to apply the newly developed multilevel CART (M-CART) algorithm for binary variables developed by Lin & Lou (2019) [147]. R code to run this algorithm was obtained from the authors. Unfortunately, constructing the M-CART models in this way was not feasible due to issues with model complexity and convergence. A full explanation of how this conclusion was made, as well as additional detail about the CART modelling, is available in Appendix C.

Table 3.1: Available Variables of Interest for youth BMI regression models from COMPASS 2018/19

Controls	Age Ethnicity
Movement	Sports Participation MVPA STSB Sleep
Diet	Fast Food consumption Breakfast consumption
Mental Health	Anxiety Depression Self-concept
Substance Use	Binge Drinking Cannabis Use E-cigarette Use

3.3.3 Study 3

Study 3 compared the results of MI and Complete Case Analysis (CCA) approaches to handle missing data in the context of analyses examining factors related to BMI among youth. CCA is known to produce biased results in situations where there are large amounts of non-random missingness [18, 39–44], which was clearly indicated by Studies 1 and 2 to be the case in this dataset. As such, if an MI model is appropriately constructed and assumptions are met, differences between MI and CCA models are demonstrative of bias from deleting cases. As a first step, an analytical model was hypothesized which examined factors associated with youth BMI; factors included dietary, movement, mental health, and substance use related variables supported by existing literature. Variables of interest for this analytical model are listed in Table 3.1.

For the CCA approach, any individual missing BMI or any factor listed in Table 3.1 was removed. For the MI approach, an imputation model was constructed using auxiliary variables from Study 2; variables identified in the sex-stratified CART models were included as auxiliary variables in the corresponding MI model. MI operates under the Missing At Random (MAR) assumption; the inclusion of auxiliary variables is necessary to increase the validity of this assumption, particularly for this study where patterns hinted at as possible Missing Not At Random (NMAR) mechanism [8]. Auxiliary variables were identified from not just the CART models for BMI, but from the height and weight models as well. The results from Study 2 were chosen over the regression findings from Study 1 because of the natural hierarchy of variables created in CART models, which was thought to be a valuable feature if any auxiliary variables needed to be parsed out of the MI model. All analysis variables (in Table 3.1) were also included in the MI models, as an MI model must be congenial with the analysis model to be appropriate. The predictor matrix, which organizes the analysis and auxiliary variables for multiple imputation, was arranged from least missing to most missing. The full predictor matrices used are available in Appendix D.

Data imputation was performed in R using the mice and miceadds packages [148, 149]. BMI was imputed through passive imputation, categorical variables were imputed through Predictive Mean Matching (PMM) and continuous variables were imputed using the MI for multivariate panel (pan) [150] method. The imputation procedure was set to produce 30 imputations over 10 iterations. There is no standard rule for setting number of imputations; the default for the mice algorithm is 5, which may be sufficient in some contexts. Other suggestions include using 20 imputation as a "rule of thumb" [15] or setting the number of imputations equal to the fraction of missing data [8, 25]. For this study, 30 imputations was chosen because it was similar the fraction of missing data for BMI. Diagnostic convergence plots indicated that the chosen specifications for the MI models were sufficient. These diagnostic plots, as well as more details about the diagnostic procedures, are presented in Appendix D. After the MI model was determined to be acceptable based on the diagnostics, the multiply imputed datasets were used for analysis.

Descriptive statistics for variables of interest were examined for both CCA data and the multiply imputed datasets, and presented in a descriptive table. Means and standard errors for MI datasets were combined using Rubins rules [24]. LMMs examining the associations between BMI and the hypothesized factors (listed in Table 3.1) were constructed for CCA and MI datasets, accounting for clustering at the school level. The general form for a two-level LMM was given in equation 3.1. For MI LMMs, estimates were pooled according to Rubins rules [24]. Sensitivity analysis was performed using a different modelling approach, generalized estimating equations (GEE) for all models. Point estimates and standard errors from MI were similar between LMM and GEE, so only LMM results were presented. The significance (determined through 95% CI) and directionality of associations were compared between corresponding CCA and MI models, and differing results between these approaches were considered to represent bias from deleting cases. In addition to being presented in traditional regression tables, point estimates and CIs were graphed to visually present where the results of CCA and MI models differed.

Chapter 4

Study 1: Learning from missing data: examining nonreporting patterns of height, weight, and BMI among Canadian youth

Authors: Amanda Doggett[1], Ashok Chaurasia[1], Jean-Philippe Chaput[2,3], Scott T. Leatherdale[1]

[1]School of Public Health Sciences, University of Waterloo, Waterloo, Ontario, Canada
[2]Department of Pediatrics, University of Ottawa, Ottawa, Ontario, Canada
[3]Healthy Active Living and Obesity Research Group, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada.

Status: Published in the International Journal of Obesity https://doi.org/10.1038/s41366-022-01154-8

4.1 Overview

Background: Youth body mass index (BMI), derived from self-reported height and weight, is commonly prone to nonreporting. A considerable proportion of overweight and obesity (OWOB) research relies on such self-report data, however little literature to date has examined this nonreporting and the potential impact on research conclusions. The objective of this study was to examine the characteristics and predictors of missing data in youth BMI, height, and weight.

Methods: Using a sample of 74 501 Canadian secondary school students who participated in the COMPASS study in 2018/19, sex-stratified generalized linear mixed models were run to examine predictors of missing data while controlling for school-level clustering.

Results: In this sample, 31% of BMI data were missing. A variety of diet, exercise, mental health, and substance use variables were associated with BMI, height, and weight missingness. Perceptions of being overweight (females: 95% CI (1.42,1.62), males: 95% CI (1.71,2.00)) as well as intentions to lose weight (females: 95% CI (1.17,1.33), males: 95% CI (1.13,1.32)) were positively associated with BMI missingness.

Conclusions: Findings from this study suggest that nonreporting in youth height and weight is likely somewhat related to the values themselves, and hint that social desirability may play a substantial role in nonreporting. The predictors of missingness identified in this study can be used to inform future studies on the potential bias stemming from missing data and identify auxiliary variables that may be used for multiple imputation approaches.

4.2 Background

4.2.1 Missing Data

Missing data is encountered in most applied research, particularly in epidemiological studies that utilize self-report surveys or questionnaires for data collection. Although there are certain elements of survey and study design that can help mitigate missing data issues [1] , missing data are often unavoidable due to the voluntary nature of research. Some observational research topics may be more impacted than others; for example, questions inherently tied to social desirability (e.g., questions about substance use or finances) are more likely to suffer from nonresponse [2, 3]. Missing data is problematic if survey responses of those who do not answer a particular question are systematically different than those who do, as this can introduce bias into the research findings since these individuals' characteristics are not captured. For this reason, the prevalence of many health risk behaviours could be over- or underestimating true population parameters. For example, rates of substance use can be underestimated as those who are more likely to engage substance use are also those who are more likely to not report those behaviours [4].

A prominent problem in applied research is the lack of consideration for how to appropriately handle missing data, or transparency in how it might impact research findings. A common default procedure for handling cases with missing data is case deletion (a method referred to as 'complete case analysis' (CCA)), which is known to produce biased results, with the nature and extent of bias being dependent on proportions and patterns of missing data [17]. Despite this, CCA remains the most commonly used method to handle missing data [35, 36]. This is a major concern in applied research because the bias that can result from CCA could also be reflected in research conclusions, and impact concomitant policy and programming decisions.

4.2.2 Missing BMI, height, and weight in youth health research

Missing data considerations seem to be lacking in public health research, which includes youth overweight and obesity (OWOB) research. Self-reported height and weight data among youth tend to be missing in high proportions, greatly exceeding levels of missingness observed for other health indicators [54, 55]. Height and weight are needed to calculate body mass index (BMI, kg/m^2), which is typically used to classify individuals as having: underweight, normal weight, overweight, or obesity. Although BMI is imperfect as it only proxies body adiposity rather than directly measuring it, researchers typically agree that BMI remains the most feasible and practical population level indicator of OWOB [108, 151–153]. Given that youth OWOB is a primary risk factor for chronic disease later in life including heart disease, diabetes, and cancer [59], population surveillance, disease monitoring, and prevention research fundamentally rely on these data. However, some studies have reported greater than 40% of youth weight or BMI data as missing [104, 105].

Addressing the missing data present in youth OWOB research is a complex issue. First, the high proportion of missing data means that the common approach of CCA is unlikely to be an appropriate option in many situations, as results are at risk of bias. However, missing data methods like maximum likelihood and multiple imputation that can help reduce this bias demand greater time, resources, and statistical expertise. It may be difficult for applied researchers working with OWOB data to leverage these approaches if there is a lack of literature in this field to reference how these methods might be used. In the least, it is beneficial for the OWOB literature to examine the predictors of missing data in order to understand nonreporting patterns in this field. However, despite the ubiquity of missing data problems in observational youth OWOB research, it appears that to date only four studies have examined predictors of missing self-report BMI and/or weight among youth [104–107]. Generally, these studies found that aspects such as age, body image, and activity levels were related to missingness in BMI or weight. However, much more research is needed in this area in order to understand nonreporting patterns, and how missing data may impact research findings in this field.

This study aims to address this knowledge gap in the literature by examining the patterns of missing BMI, height, and weight data, as well as factors associated with this missingness, among a large sample of Canadian youth who participated in the COMPASS study. The findings of this research may be useful to inform how missing data may impact youth OWOB research, as well as help inform future missing data decisions, such as auxiliary variable selection. This study may also be useful for the many researchers who use similar secondary data derived from the COMPASS study.

4.3 Methods

4.3.1 Sample

This study uses data from a larger host study known as the COMPASS study (herein referred to as 'COMPASS'). COMPASS is a longitudinal cohort study in Canada that tracks youth health behaviours over time. COMPASS operates in Ontario, Alberta, Québec, and British Columbia, and uses school-based purposive sampling to collect data on grades 9-12 students. Within these provinces, schools within school boards that support active-information passive-consent protocols are invited to participate in COMPASS. At participating schools, COMPASS collects self-report data on a variety of different youth demographics and health behaviours (e.g., diet, physical activity, substance use, mental health). All students present on data collection day are invited to complete the COMPASS questionnaire, which are paper-based and administered during class time using an active-information passive-consent protocol. Further information on COMPASS data collection tools and administration are detailed elsewhere [154]. The present research uses a cross-sectional sample from the 2018/19 COMPASS data collection year, consisting of 74 501 youth from 136 schools (Ontario: 61, Alberta: 8, Québec: 52, British Columbia: 15). Unit nonresponse for the 2018/19 data collection year was 15.7%, mostly attributable to absence from school or a scheduled spare during the data collection time. More information about the COMPASS study, including in-depth explanations of recruitment and data collection protocols can be found in print [124], or online www.compass.uwaterloo.ca.

The COMPASS Study received ethics approval from the University of Waterloo Office of Research Ethics as well as participating school boards, and informed consent was obtained from those participating in the study.

It is important to highlight that the present study will be using the term 'missing data' to refer specifically to item-nonresponse, whereby some, but not all, data are missing. Unit nonresponse, which refers to the complete lack of data at a particular time point (e.g. a student refusing to participate in the data collection during class time), is not examined in this study because the feasibility of addressing unit nonresponse in post-hoc analyses largely relies on there being some data which exists on those individuals (e.g. from other sources, previous time points, etc.) [5]. In the COMPASS study, the anonymous nature of participation means that no data are available on those who choose not to participate in a cross-sectional sample.

4.3.2 Variables

Weight, height, and Body Mass Index

BMI is a derived variable determined by dividing weight (kg) by height squared (m2). Weight was determined by asking students, "how much do you weigh without your shoes on? (please write your answer in pounds OR in kilograms, and then fill in the appropriate numbers for your weight.)" There is a space for students to write out their weight, as well as separate Scantron bubbles for them to fill in. Height is similarly determined by asking, "how tall are you without your shoes on? (please write your height in feet and inches OR in centimeters, and then fill in the appropriate numbers for your height)." While most missingness in the data occurred because of participant nonresponse, some was researched imposed; outliers less than 45lbs (20kgs) or greater than 390lbs (117kgs) for weight and less than 4ft (122cm) or greater than 6ft 11 (211cm) for height were marked as missing. Binary indicators of missingness (i.e., missing vs. not missing) were created for each height, weight, and BMI (for any cases where height and/or weight were missing).

Predictor Variables of Interest

Certain variables were required in all models, including: age (continuous), ethnicity (racialized, non-racialized), weight perception (underweight, overweight, about right), and weight goal (lose, gain, stay the same/not trying anything). Ethnicity was collapsed into a binary measured due to insufficient counts for appropriate statistical inference and small cell counts impacting the computational feasibility of model convergence. Weight perception and weight control intention were considered important to include because these variables are the closest available variables that have the potential to somewhat proxy a missing value for BMI.

Given limited existing literature and the wide breadth of variables that could feasibly be related to BMI missingness, a number of variables of interest were included in initial variable selection analyses. All the hypothesized variables from the COMPASS study questionnaire were screened for possible association with missingness in BMI. If variables overlapped or were similar to those used in previous studies [104–107], or if they could feasibly be related to BMI itself, they were included in this initial step. These variables were then categorized into blocks whereby variables that represented concepts within a similar domain were grouped. These blocks were: diet, movement, academic, mental health, and substance use. For an explanation of how final variables were chosen for each model, see Section 4.3.3.

- The diet block included 24-hours recall for servings of: fruits and vegetables, grain products, meat and alternatives, and milk and alternatives as well as number of days per week of breakfast, energy drink, and fast-food consumption.
- The movement block included hours of moderate-to-vigorous physical activity, sports participation (yes/no), number of days per week of strength training, number of physically active peers, minutes of screen time sedentary behaviour, and average hours of sleep. Notably, the self-report physical activity and sedentary behaviour measures used in this study have been shown to have high test-retest reliability in a youth population [124].
- The academic block included: English grade ($< 50\%$, $\geq 50\%$), math grade ($< 50\%$, $\geq 50\%$), and 4-week truancy (no classes skipped, more than 1 class skipped).
- The mental health block included clinically relevant symptoms of depression (CESD-R-10 scale [131])), anxiety (GAD7 scale [132]), socio-emotional skills (DERS scale [133]), as well as self-reported wellbeing (flourishing scale [134]), self-concept (self-description questionnaire II [135]), self-rated mental health, and reported status as a bullying victim (yes/no) or perpetrator (yes/no).
- The substance use block included binary indicators of binge drinking, smoking, e-cigarette use, cannabis use, and alcohol mixed with energy drink use.

4.3.3 Analysis

Three of the four existing studies that have previously examined missingness in BMI or weight among youth conducted initial bivariate analyses (chi-squared and ANOVA), and subsequently built a logistic regression model using only the variables which were significant according to bivariate comparisons [104–106]. This approach was not used in our study because the large sample size and large proportion of missing data yielded significance for nearly all bivariate comparisons and including all variables in a logistic regression model would be neither statistically nor inferentially practical. Instead, we used the model selection criteria AIC and BIC to facilitate the decision around variable selection. AIC and BIC are statistical model selection criteria which aid researchers in determining which model is the most plausible [155, 156]. The pseudo-likelihood model selection framework developed by Ten Eyck & Cavanaugh [141] was used to compare all candidate variables for each variable block. When selecting the top candidate variable(s) for each block, a combination of minimum AIC, BIC, and model parsimony were considered, whereby the model with the fewest number of variables that still had low AIC and BIC comparable to higher order models was chosen. The SAS macro for this framework was obtained from Ten Eyck & Cavanaugh, which was modified for our purposes, helping identify essential variables predictive of missingness observed in BMI, weight and height, while controlling for the required variables listed in Section 4.3.2. This was an important step, as the goal was for the model selection algorithm to select the top variables from each block after the variance explained by the identified essential variables had already been accounted for.

Descriptive statistics were used on our study sample to explore the missingness in BMI, height, and weight. All components of the analyses were stratified by a binary indicator of sex to examine any differences in the reporting of height and weight between males and females. In order to examine predictors of missingness, separate regression models were created where the binary outcomes were missingness in: 1) BMI, 2) height, and 3) weight. This yielded 6 total models as each were stratified between males and females. A generalized linear mixed model approach (via PROC GLMMIX in SAS 9.4) was used in order to account for data clustering at the school level, whereby school was specified as a random effect in the model. Although the province in which a school is located could represent an additional level of clustering, due to computational limitations, only school-level clustering was controlled for in the modelling.

4.4 Results

4.4.1 Descriptive Analyses

Figure 4.1 shows the degree of nonresponse in this sample for each main variable included in the COMPASS student questionnaire. Of all variables, BMI had the highest degree of missingness (31.3%, n=23 329), followed by weight (21.3%, n=15 849) and height (17.5%, n=13 049). Mental health related measures demonstrated the next highest degree of missingness, the top being the CESD-R-10 scale used to measure clinically relevant symptoms of depression, where 14.6% (n=10 854) of data were missing for this variable.

Of the participants missing BMI, 32.2% (n=7520) were missing just weight, 20.2% (n=4720) were missing just height, and 35.7% (n=8329) were missing both height and weight. The remaining 11.8% (n=2760) provided height and weight data but were marked as missing post-hoc due to extreme outlier values (BMI <10 or >50).

Figure 4.2 shows the breakdown on missing BMI by sex and types of non-reporting. Among those missing weight only, 59.1% (n= 4471) were female, whereas 39.6% (n=2977) were male. Among those missing height only, 41.6% (n=1962) were female and 57.1% (n=2696) were male. Of those missing both height and weight, 41.7% (n=3477) were female and 52.5% (n=4371) were male. Table 4.1 presents stratified descriptive statistics for all variables included in at least one of the final mixed models.

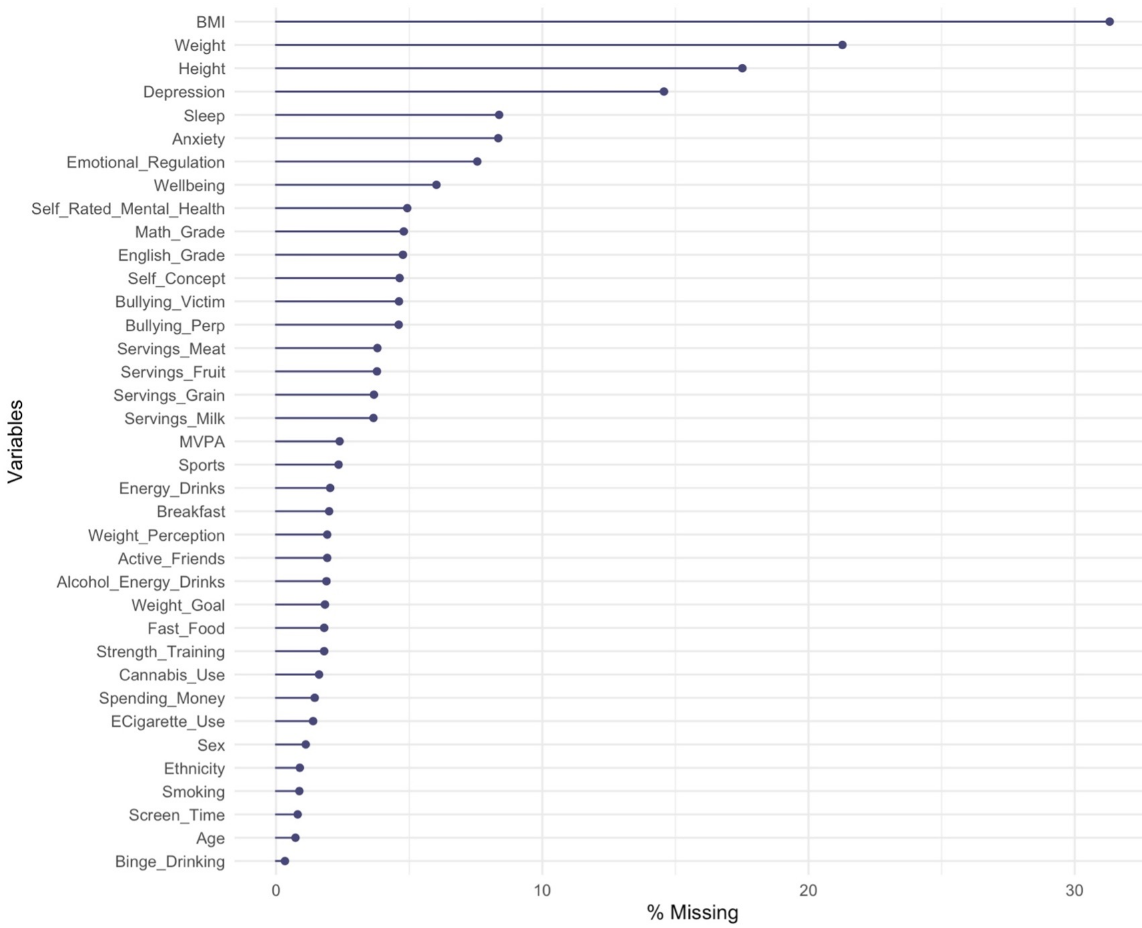


Figure 4.1: Degrees of item nonresponse across a sample of COMPASS variables (2018-19)

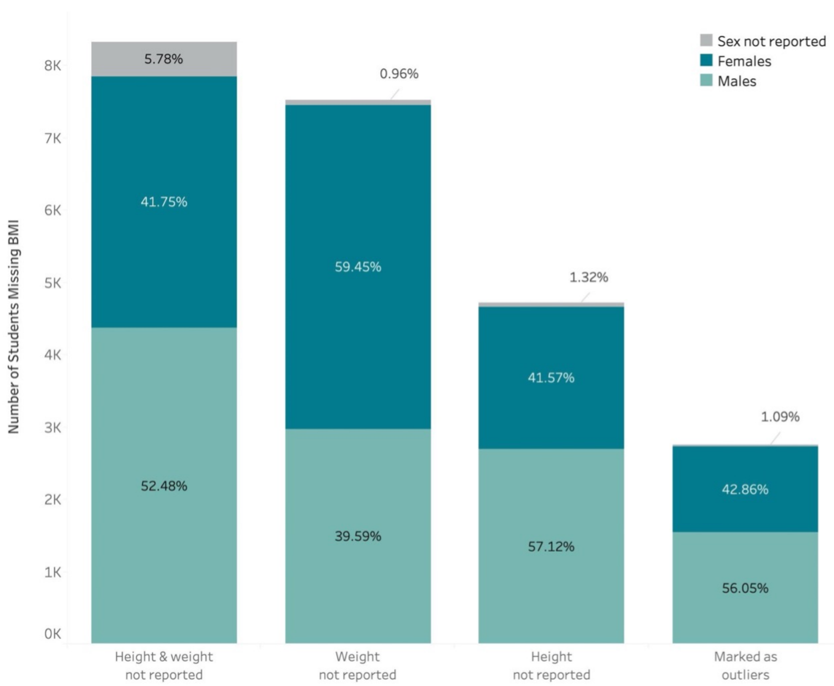


Figure 4.2: BMI missingness categories by reported sex (COMPASS 2018-19)

Table 4.1: Descriptive statistics of COMPASS study sample (2018/19)

Variable[1]	**Category**	% (n)[2]		
		Females n=36546	**Males N=37126**	**Total**[3] **N=74501**
Age	Average (SD)	15.14 (1.50)	15.18 (1.51)	15.16 (1.51)
	Missing	0.08 (31)	0.19 (69)	0.73 (541)
Ethnicity	Racialized	69.45 (25383)	68.62 (25477)	68.48 (51017)
	Non-Racialized	30.27 (11063)	30.99 (11505)	30.63 (22822)
	Missing	0.27 (100)	0.39 (144)	0.89 (662)
Weight Perception	Underweight	11.47 (4190)	21.00 (7795)	16.30 (12140)
	Overweight	25.85 (9448)	19.93 (7398)	22.87 (17038)
	About Right	61.14 (22343)	57.19 (21233)	58.92 (43893)
	Missing	1.55 (565)	1.89 (700)	1.92 (1430)
Weight Goal	Gain	5.98 (2186)	24.57 (9121)	15.43 (11492)
	Lose	40.26 (14715)	20.71 (7690)	30.32 (22587)
	Stay the same or not trying anything	52.38 (19144)	52.83 (19612)	52.42 (39055)
	Missing	1.37 (501)	1.89 (703)	1.83 (1367)
Dietary Block				
Fruit/Vegetable Servings	Average (SD)	2.89 (1.89)	3.06 (2.11)	2.98(2.01)
	Missing	2.44 (890)	4.74 (1759)	3.79 (2822)
Meat/ Alternatives Servings	Average	1.88 (1.03)	2.41 (1.20)	2.15 (1.15)
	Missing	2.45 (896)	4.76 (1766)	3.80 (2833)
Breakfast Consumption	Average	4.67 (2.37)	5.05 (2.33)	4.85 (2.36)
	Missing	1.31 (479)	2.30 (855)	1.99 (1484)
Energy drink Consumption	Average	0.24 (0.90)	0.53 (1.34)	0.388 (1.16)
	Missing	1.16 (424)	2.49 (925)	2.03 (1512)
Movement Block				
Sports Participation	Yes	56.70 (20720)	62.05 (23036)	59.24 (44135)
	No	41.70 (15241)	35.25 (13088)	38.41 (28618)
	Missing	1.60 (585)	2.70 (1002)	2.35 (1748)
Strength Training	Average	2.24 (2.02)	2.77 (2.27)	2.51 (2.16)
	Missing	1.29 (473)	1.93 (717)	1.80 (1344)

Table 4.1 continued from previous page

Friend Physical Activity	Average	3.03 (1.68)	3.52 (1.69)	3.28 (1.71)
	Missing	1.35 (494)	2.13 (789)	1.92 (1430)
Screen Time Sedentary Behaviour	Average	6.54 (4.59)	7.21 (5.08)	6.91 (4.93)
	Missing	0.42 (152)	0.88 (325)	0.81 (602)
Academic Block				
English Grade	<50%	1.09 (399)	2.44 (907)	1.83 (1362)
	$\geq$ 50%	95.39 (34862)	91.92 (34128)	93.41 (69590)
	Missing	3.52 (1285)	5.63 (2091)	4.76 (3549)
Math Grade	<50%	3.36 (1162)	4.10 (1521)	3.70 (2757)
	$\geq$ 50%	92.82 (32065)	90.28 (33518)	91.50 (68172)
	Missing	3.61 (1319)	5.62 (2087)	4.79 (3572)
Mental Health Block				
Self-rated Mental Health	Average	2.76 (1.21)	2.21 (1.15)	2.49 (1.21)
	Missing	3.37 (1230)	6.05 (2245)	4.93 (3670)
Wellbeing	Average	31.78 (5.75)	32.64 (5.60)	32.19 (5.72)
	Missing	4.84 (1770)	6.78 (2518)	6.02 (4486)
Anxiety	Average	7.69 (5.75)	4.58 (4.91)	6.15 (5.59)
	Missing	7.66 (2801)	8.61 (3196)	8.34 (6216)
Substance Use Block				
Smoking	Yes	6.64 (2425)	8.00 (2969)	7.43 (5532)
	No	92.89 (33949)	91.01 (33790)	91.70 (68320)
	Missing	0.47 (172)	0.99 (367)	0.87 (649)
E-cigarette use	Yes	25.48 (9312)	30.34 (11264)	27.99 (20852)
	No	73.75 (26951)	67.98 (25237)	70.62 (52614)
	Missing	0.77 (172)	1.68 (625)	1.39 (1035)

[1] Table only includes variables selected for inclusion in at least one of the models by the variable selection framework outlined in section 4.3.2
[2] Average (SD) rather than %(n) where indicated in category column
[3] Includes those who did not report sex, so sex-stratified counts may not add to total counts

4.4.2 Regression Models

Results of the models examining associations with BMI, height, and weight missingness are presented in Tables 4.2, 4.2, and 4.2, respectively. Each table presents two models, as analyses were stratified by sex. Although covariates are highly similar between models, they are not identical, as the pseudo-likelihood model selection algorithm identified different explanatory variables for each. The dietary and mental health blocks were the only blocks where the algorithm selected different variables between males and females.

Table 4.2: Regression model predicting BMI missingness among youth in the COMPASS study (2018/19)

Variable[1]	**Category**	**OR (95% CI)**	
		Model 1.1: Females N= 36546	**Model 1.2: Males N=37126**
Age		0.80 (0.78, 0.82)	0.85 (0.83, 0.87)
Ethnicity	Racialized	1.36 (1.27, 1.44)	1.40 (1.32, 1.50)
	Non-Racialized	—	—
Weight Perception	Overweight	1.51 (1.42, 1.62)	1.85 (1.71, 2.00)
	Underweight	1.00 (0.91, 1.09)	1.17 (1.09, 1.27)
	About right	—	—
Weight Goal	Gain	1.07 (0.95, 1.22)	0.68 (0.63, 0.74)
	Lose	1.25 (1.17, 1.33)	1.22 (1.13, 1.32)
	Stay the same or not trying anything	—	—
Meat/Alt. Servings		N/A	0.95 (0.92, 0.97)
Fruit/Vegetable Servings		0.98 (0.96, 0.99)	0.99 (0.98, 1.01)
Breakfast Consumption		0.99 (0.98, 1.00)	0.99 (0.98, 1.01)
Energy Drink Consumption		1.08 (1.05, 1.11)	1.05 (1.03, 1.08)
Sports Participation	Yes	0.80 (0.75, 0.85)	0.64 (0.60, 0.68)
	No	—	—
Strength Training		0.95 (0.93, 0.96)	0.95 (0.94, 0.96)

Table 4.2 continued from previous page

Friends Physical Activity		0.95 (0.94, 0.97)	0.94 (0.92, 0.96)
Screen Time Sedentary Behaviour		1.05 (1.04, 1.05)	1.05 (1.05, 1.06)
English Grade	<50%	1.77 (1.36, 2.31)	1.44 (1.21, 1.72)
	$\geq$ 50%	—	—
Math Grade	<50%	1.31 (1.13, 1.52)	1.15 (1.00, 1.33)
	$\geq$ 50%	—	—
Self-rated Mental Health		0.93 (0.90, 0.96)	N/A
Wellbeing		0.98 (0.98, 0.99)	0.99 (0.98, 0.99)
Smoking	Yes	1.18 (1.05, 1.33)	1.15 (1.03, 1.29)
	No	—	—
E-cigarette Use	Yes	0.69 (0.65, 0.74)	0.63 (0.59, 0.68)
	No	—	—

[1] Variables differ slightly between models, as the variable selection framework described in section 4.3.3 was conducted separately for each model. N/A is present if a particular variable was not selected for inclusion in a model.

Table 4.3: Regression model predicting height missingness among youth in the COMPASS study (2018/19)

Variable[1]	Category	OR (95% CI)	
		Model 1.1: Females N= 36546	**Model 1.2: Males N=37126**
Age		0.71 (0.69, 0.73)	0.76 (0.74, 0.77)
Ethnicity	Racialized	1.46 (1.35, 1.59)	1.45 (1.34, 1.56)
	Non-Racialized	—	—
Weight Perception	Overweight	0.77 (0.70, 0.85)	0.97 (0.88, 1.07)
	Underweight	1.14 (1.02, 1.28)	1.16 (1.06, 1.26)
	About right	—	—
Weight Goal	Gain	1.12 (0.96, 1.30)	0.69 (0.63, 0.76)
	Lose	0.98 (0.89, 1.06)	0.94 (0.85, 1.03)
	Stay the same or not trying anything	—	—
Meat/Alt. Servings		N/A	0.92 (0.89, 0.95)
Fruit/Vegetable Servings		0.95 (0.93, 0.97)	0.97 (0.95, 0.99)
Breakfast Consumption		0.99 (0.97, 1.01)	N/A
Energy Drink Consumption		1.11 (1.07, 1.16)	1.03 (1.00, 1.06)
Sports Participation	Yes	0.74 (0.69, 0.80)	0.58 (0.54, 0.62)
	No	—	—
Friends Physical Activity		0.94 (0.92, 0.96)	0.91 (0.90, 0.93)
Screen Time Sedentary Behaviour		1.04 (1.03, 1.05)	1.05 (1.04, 1.05)
English Grade	<50%	1.98 (1.48, 2.66)	1.50 (1.23, 1.82)
	≥ 50%	—	—
Math Grade	<50%	1.39 (1.16, 1.67)	1.06 (0.90, 1.25)

Table 4.3 continued from previous page

	≥ 50%	—	—
Self-rated Mental Health		0.83 (0.80, 0.86)	N/A
Wellbeing		0.97 (0.96, 0.98)	0.99 (0.98, 0.99)
Smoking	Yes	1.24 (1.06, 1.46)	1.28 (1.12, 1.47)
	No	—	—
E-cigarette Use	Yes	0.67 (0.61, 0.73)	0.60 (0.55, 0.65)
	No	—	—

[1] Variables differ slightly between models, as the variable selection framework described in section 4.3.3 was conducted separately for each model. N/A is present if a particular variable was not selected for inclusion in a model.

Table 4.4: Regression model predicting weight missingness among youth in the COMPASS study (2018/19)

Variable[1]	**Category**	**OR (95% CI)**	
		Model 1.1: Females N= 36546	**Model 1.2: Males N=37126**
Age		0.82 (0.80, 0.84)	0.87 (0.85, 0.89)
Ethnicity	Racialized	1.33 (1.24, 1.43)	1.55 (1.44, 1.67)
	Non-Racialized	—	—
Weight Perception	Overweight	1.09 (1.01, 1.17)	1.13 (1.03, 1.23)
	Underweight	0.82 (0.73, 0.91)	1.12 (1.02, 1.22)
	About right	—	—
Weight Goal	Gain	1.02 (0.88, 1.18)	0.67 (0.61, 0.74)
	Lose	1.27 (1.18, 1.36)	1.04 (0.95, 1.14)
	Stay the same or not trying anything	—	—
Meat/Alt. Servings		N/A	0.89 (0.87, 0.92)
Fruit/Vegetable Servings		0.98 (0.96, 0.99)	N/A
Breakfast Consumption		0.99 (0.97, 1.00)	0.99 (0.98, 1.01)
Energy Drink Consumption		1.06 (1.03, 1.10)	1.03 (1.01, 1.06)
Sports Participation	Yes	0.76 (0.71, 0.81)	0.60 (0.56, 0.64)
	No	—	—
Strength Training		0.95 (0.93, 0.96)	0.94 (0.93, 0.96)
Screen Time Sedentary Behaviour		1.05 (1.05, 1.06)	1.05 (1.05, 1.06)
English Grade	$<50\%$	1.21 (0.91, 1.60)	1.24 (1.03, 1.51)
	$\geq 50\%$	—	—
Math Grade	$<50\%$	1.33 (1.13, 1.57)	1.27 (1.09, 1.48)
	$\geq 50\%$	—	—

Table 4.4 continued from previous page

Anxiety (GAD7)		0.99 (0.98, 0.99)	N/A
Wellbeing		0.98 (0.98, 0.99)	0.98 (0.98, 0.99)
Smoking	Yes	1.04 (0.91, 1.19)	1.06 (0.92, 1.21)
	No	—	—
E-cigarette Use	Yes	0.69 (0.64, 0.75)	0.59 (0.54, 0.64)
	No	—	—

[1] Variables differ slightly between models, as the variable selection framework described in section 4.3.3 was conducted separately for each model. N/A is present if a particular variable was not selected for inclusion in a model.

4.5 Discussion

This study examined the patterns and predictors of missing BMI, height, and weight data among a large sample of Canadian youth. The extent of missingness for these indicators was substantial; nearly 1 in 3 students were missing BMI because they did not report their height, weight, or both. Compared to all other variables examined in this study, which included a large breadth of health behaviours, BMI, height, and weight showed the highest rate of missingness. Females were more likely to be missing BMI because they did not report their weight, whereas males were more likely to not report their height. Consistent with previous research, younger male and female participants were less likely to report their height and weight, which has been hypothesized to be at least in part related to not knowing what their current height and/or weight is [105, 106, 108].

Racialized youth were more likely to be missing data on height, weight, and BMI. Although some previous research has examined missing response patterns across racial or ethnic groups [157], there is very little existing literature that has attempted to explain nonreporting patterns. It is important, however, to acknowledge that where analyses show differences in nonreporting patterns by race/ethnicity, the missing data methodology used should be carefully considered. If these differences in missingness go unaccounted for, misestimated effects can lead to inappropriate conclusions and recommendations, in particular in the health domain [158–160].

4.5.1 Weight perceptions and goals

In this study sample, those who reported perceiving themselves as overweight were more likely to be missing weight and thus BMI. This association was quite prominent for BMI missingness, whereby females who reported perceiving themselves as overweight were 51% (and males 85%) more likely to be missing BMI compared to youth who perceive themselves as about the right weight. Interestingly, females who report perceiving themselves as underweight were less likely to be missing weight compared to those who perceive themselves as about right, but the opposite was true for underweight males, as they were still less likely to report the weight than their male counterparts who reported perceiving themselves as about the right weight. Reported intentions to gain weight was significantly inversely associated with likelihood of missing BMI, height, and weight for males only, meaning that males who reported intention to gain weight were more likely to report their height and weight compared to male counterparts who said they wanted to stay the same or weren't trying anything. On the other hand, females who indicated a desire to lose weight were more likely to not report their weight.

Given the social stigma that surrounds body image during adolescence [112] these patterns of missingness are somewhat expected, since social desirability can influence nonresponse to survey questions [161,162]. Having a higher BMI is known to be associated with greater weight-related concerns among youth, in particular among females [112]. Moreover, male youth are more likely to express a desire to gain weight or muscle mass [163, 164]. This study indicated overweight perceptions and weight loss intentions predicted missingness among females whereas underweight perceptions predicted missingness among males. These findings suggest that at least some of the nonreporting present in this study is related to body image concerns. Descriptive statistics may have hinted that lack of knowledge of height and weight values might be a main driver of non-reporting in this sample due to the relatively low missingness level for the weight perception variable; however if this was the case, it would be expected that weight perception and weight loss intentions would not be significantly related to missingness in BMI, height, or weight. It is possible that for youth, there is greater social stigma surrounding the quantitative value itself compared to reporting self-perception, leading to the missingness patterns observed in this study. Notably, unlike weight, measures of height perception were not recorded in this study, so findings surrounding height missingness are more difficult to interpret. However, since the majority of those missing BMI were missing both height and weight, it is possible that body image related concerns lead participants to skip this particular question in its entirety.

Although missingness mechanisms are not the focus of this paper, it is worth highlighting that this study suggests that, in this sample, BMI and weight data are not missing at random (MNAR), as the findings related to weight perception imply that the likelihood of missing BMI and weight appear to be at least in part associated with the true unreported weight value itself. This further emphasizes the importance of the present study, as well as missingness investigations in applied research in general, as researchers need to not only be aware of these missingness patterns, but also need to be aware of potential auxiliary variables that can be leveraged to increase the validity of the missing at random (MAR) assumption. Notably however, it is important to avoid making the assumption that those who perceive themselves as overweight, or indicate a desire to lose weight, are in fact clinically classified with OWOB. The authors caution that neither weight perception nor weight goal be considered a direct proxy for weight or BMI, since research has indicated that sizable percentages of youth misclassify themselves; many youth who are in the normal BMI range perceive themselves as overweight, and vice versa [127, 128].

4.5.2 Diet

Consumption of meat/alternatives was inversely associated with missingness across BMI, height, and weight for males. Recalling that weight gain intentions were also negatively associated with missingness, this finding makes sense in the context of existing literature which has indicated that males who desire to gain weight are more likely to consume more protein-rich foods such as those found in this food group [165,166]. On the other hand, energy drink consumption was associated with greater likelihood of missingness (compared to no energy drink consumption), in particular for females. Given that sugar-added drinks are known to be associated with greater BMI [76], this is also consistent with the findings surrounding weight perceptions. Notably, although the significance of dietary variables varied in this study and may somewhat reflect those already observed for weight perception, given the well-established associations between diet and BMI [70–73,75] these variables are likely to be useful as auxiliary variables in most analyses using BMI.

4.5.3 Movement

Across all models, those who participated in sports were less likely to be missing BMI, height, and weight, compared to those who did not participate in sports. Similarly, models indicated that greater number of days spent strength training was associated with less likelihood of missingness. On the other hand, screen time showed the opposite pattern, where greater number of screen time hours was associated with greater likelihood of missingness. Similar to the dietary variables, these findings appear to be consistent with what might be expected based on the weight perception findings, given that low physical activity and high screen time have been shown to be associated with greater BMI among youth [70,82].

4.5.4 Academic

Across most models, youth with poorer academic achievement were more likely to be missing BMI, height, and weight. Previous research has identified an 'obesity achievement gap,' whereby students with OWOB may internalize the stigma and biases they face, which can lead to poorer academic outcomes [167,168]. As such, the academic-related findings in this study are consistent with the findings that overweight weight perception also predicted missingness.

4.5.5 Mental Health

Females with higher self-rated mental health were less likely to be missing weight, height, or BMI. Similar associations were observed in the wellbeing scale across males and females. These associations may be in part related to body image, as research indicates that body image concerns are associated with poorer mental health [169–171]. Notably, this study showed that mental health data were also prone to missingness and, as such, the associations observed here do not reflect those who did not report mental health data in addition to nonreporting of their height/weight data. It is clear that mental health is a domain which can also suffer greatly from missing data, and future research should consider focussing on missingness in mental health indicators among youth.

4.5.6 Substance Use

While smoking was associated with greater likelihood of missingness among BMI and height compared to not smoking, e-cigarette use was associated with less likelihood of missingness in height, weight, and BMI. Those who reported e-cigarette use were 32% less likely to be missing weight compared to those who don't use e-cigarettes. While cigarette smoking in Canada is generally perceived as a negative behaviour which is declining in popularity [172, 173], e-cigarette use perceptions are drastically different. Research has found that positive perceptions surrounding e-cigarette use are common among youth, including perceptions of being a socially desirable or 'cool' behaviour [174–176]. Just as social stigma can negatively influence reporting, it is possible that the opposite is true, whereby those who engage in behaviours considered more socially desirable are more willing to report health-related metrics, including height and weight as indicated by this study.

4.5.7 Strengths & Limitations

This study has several strengths to highlight, including the large sample size, the active information passive consent procedure, and the novel focus on missing youth BMI, height, and weight data. This study adds to very limited existing research on this topic and has the potential to directly inform the reporting and handling of missing data in this field. Moreover, this study highlights the importance of missing data considerations for applied research in general. This study also has some limitations to note. The study sample was purposive and as such may not necessarily be generalizable to all Canadian youth. Replication of these analyses within representative population samples of youth is warranted. Moreover, although this manuscript highlighted the importance of missing data considerations, it must be acknowledged that this study itself was impacted by missing responses on predictor variables, which may have impacted the associations observed. Lastly, it should be noted that this study was observational and as such the results do not identify causal mechanisms for missingness in BMI, height, and weight, but rather identify different variables that are statistically associated with missingness in these data.

4.6 Conclusions

The variables associated with missing BMI, height, and weight in this study seems to indicate two overarching findings. First, it is likely that missing values are at least in part related to the values themselves, given that males and females who perceive themselves as overweight are less likely to report their height and weight, and that several known correlates of BMI, such as diet and physical activity variables, were also associated with missingness in BMI. Second, missingness in this sample seemed to hint that social desirability played a large role in nonreporting, as many observed associations in this study were congruent to literature which has examined social desirability of health indicators, such as those surrounding body image concerns in youth.

This study fills an important gap in the literature by examining missingness patterns and predictors for BMI, height, and weight among youth. Not only does this study add to our existing understanding of survey nonresponse in this domain, but also serves as a useful tool for other researchers who work with similar large epidemiological datasets. This study can inform auxiliary variable selection for those who wish to use methods such as multiple imputation, as well as help researchers identify potential sources and direction of nonresponse bias in research which uses youth BMI, height, or weight.

Chapter 5

Study 2: Using classification and regression trees to model missingness in youth BMI, height, and body mass

Authors: Amanda Doggett[1], Ashok Chaurasia[1], Jean-Philippe Chaput[2,3], Scott T. Leatherdale[1]

[1]School of Public Health Sciences, University of Waterloo, Waterloo, Ontario, Canada
[2]Department of Pediatrics, University of Ottawa, Ottawa, Ontario, Canada
[3]Healthy Active Living and Obesity Research Group, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada.

Status: Under Review at Health Promotion and Chronic Disease Prevention

5.1 Overview

Background: Research suggests that there is often a high degree of missingness in self-reported body mass index (BMI) data among youth. Although public health relies on these data to survey body adiposity and conduct research surrounding overweight and obesity in this population, very few studies have focused on examining missingness in this domain.

Methods: This study used classification and regression tree (CART) models to examine missingness in youth height, body mass, and BMI among 74 501 youth in Canada who participated in the COMPASS study in 2018/19.

Results: Findings suggest that social desirability played a large role in nonreporting among both males and females. Findings also suggest that and that those who perceived themselves as overweight were more likely to be missing their height and body mass values.

Conclusions: This study adds to limited examination of missing data related to youth BMI. This study also demonstrated the utility of CART models for examining missingness, highlighting how they may be used as an initial step to the appropriate handling missing data.

5.2 Introduction

5.2.1 Missing Data in OWOB (overweight and obesity) literature

OWOB remains one of the top health concerns globally, being one of the strongest predictors of future chronic diseases [59]. Many studies that examine OWOB use body mass index (BMI) derived from self-report measures of height and body mass to provide a proxy measure of body adiposity. Self-report measures can be less accurate compared to direct anthropomorphic measurements; there is a tendency for individuals to underreport their body mass and overreport their height [55, 123, 177, 178]. However, researchers generally acknowledge that self-report BMI measures are substantially more feasible (both logistically and financially) than other alternative approaches to population surveillance [55, 123, 178]. Overall, self-report BMI measures have utility when used in the appropriate context where the limitations of the data are understood.

A less-discussed methodological issue associated with self-reported height and body mass is nonresponse (i.e., missing data). Among youth, who are a primary target in the OWOB prevention literature, self-report height and body mass tend to be missing in large proportions; sometimes over 50% [104, 105]. Deleting these missing cases (a method called 'complete case analysis (CCA)') can be a problematic approach, as simulation studies have established that this method tends to produce biased results, in particular when there is a large proportion of missingness and when data are not missing at random [18]. Despite this, reviews have indicated that CCA remains the most common approach in epidemiological literature [35, 36]. Given the high degree of missingness in youth self-report height and body mass data, this raises concerns about the missing data methods used in this field and how mishandling of missing data may impact research findings and conclusions.

Sophisticated statistical approaches to handle large proportions of non-random missingness are available to researchers, however, they generally require more time and expertise which may be a barrier to their use more generally. That being said, an important initial step towards selecting a reasonable and practical method for handling missing data is understanding the extent and patterns of missingness in a dataset. This is important to not only understand potential sources of nonreporting bias, but may also be a necessary step to identify inputs for certain missing data approaches (e.g., multiple imputation). Identifying various sources of missingness is especially important in large datasets with many variables, as methods for handling missingness can become exponentially complicated. Moreover, given missingness is generally unique to studies, there is no clear framework for the process for identifying sources of missingness.

5.2.2 Regression Approaches

Existing research examining BMI or body mass missingness has employed regression approaches [104–106]. However, regression approaches may not be ideal for examining missingness because missingness models may be more complex than a simplistic regression approach allows. Moreover, the process for variable selection in regression models can be ambiguous. Typically, when building a regression model, an initial first step to variable selection might be to review the literature for similar analyses; this poses a challenge in the context of examining BMI missingness because the literature is sparse in this area. Bivariate comparisons are also sometimes used to decide on regression inputs; however, for large datasets with substantial missingness, this may not be useful for elimination purposes as many bivariate associations may be statistically significant. Common model selection procedures such as AIC or BIC can be used for variable selection, but this can be challenging in practice; the authors previously examined BMI, height, and body mass missingness using model selection procedures for generalized linear mixed models [179], but this required many additional modelling decisions and a customized algorithm suitable for pseudo-likelihood methods [141]. Lastly, in situations where variable selection processes yield a large number of relevant variables, the decision process for what to exclude in order to produce a parsimonious model may not be clear. In such cases, identifying some hierarchy of variable importance would be beneficial; not only may it help with parsimony and clearer interpretation, but it may also be a necessary step to employ certain missing data approaches like multiple imputation.

5.2.3 Decision Trees

Decision trees are a type of machine learning approach which have been leveraged in applied research, including some uptake in public health [180, 181]. Decision trees are useful not only for analyzing primary data, but may also be used to examine missing data, and can be a solution to some of the variable selection problems described above. Decision trees recursively split the data by predictor variables and can handle large datasets with multiple predictors measured on different scales with relative ease. Once pruned, decision trees present a parsed selection of predictor variables in a hierarchical format, allowing for some inference on variable importance. Decision trees are also advantageous because unlike regression, the entire model can be easily visualized, which may aid their interpretation. In 2015, Tierney et al. published work demonstrating the utility of using decision trees to examine missing data [53], but to our knowledge it does not appear that any applied publications thus far have leveraged this approach.

The purpose of this study is threefold: (i) add the limited literature on missing data in youth self-reported height and body mass, (ii) identify potential areas of bias stemming from nonreporting in the youth OWOB domain, and (iii) to demonstrate the use of decision trees to model missing data, which builds on the work of Tierney et al. [53] who first identified the utility of this approach.

5.3 Methods

5.3.1 Sample

This study uses a large cross-sectional dataset from the 2018/19 wave of the COMPASS study, which is a cohort study that collects data on a variety of different health behaviours among youth over time. The 2018/19 cross-sectional wave of COMPASS data consists of 74,501 youth, representing an 84.3% participation rate. COMPASS uses an active information passive consent protocol which yields high participation rates, and non-participation is usually due to absence on the data collection day or being in a scheduled spare during the data collection time.

5.3.2 Variables

This study focuses on missingness in BMI, as well as missingness in the height and body mass variables used to derive BMI. Binary indicators of missingness (i.e., missing vs. not missing) were created for each of these variables. Body mass was reported by asking students, "how much do you weigh without your shoes on? (please write your answer in pounds OR in kilograms, and then fill in the appropriate numbers for your weight.)" Height is similarly reported by asking, "how tall are you without your shoes on? (please write your height in feet and inches OR in centimeters, and then fill in the appropriate numbers for your height)." BMI is a derived variable determined by dividing body mass (kg) by height squared (m2).

A benefit of decision tree approaches is the feasibility to include many variables. In this study, a variety of diet, movement, academic, mental health, and substance use variables were included. Diet-related variables included servings of fruits and vegetables, grain products, meat and alternatives, and milk and alternatives, as well as number of days per week of breakfast, energy drink, and fast-food consumption. Movement-related variables included moderate-to-vigorous physical activity, sports participation, strength training, physically active peers, screen time sedentary behaviour (STSB), and sleep. Academic-related variables included English grade, Math grade, and truancy. Mental health variables included clinically relevant symptoms of depression (CESD-R-10 scale [131]), anxiety (GAD7 scale [132]), socio-emotional skills (DERS scale [133]), as well as self-reported wellbeing (flourishing scale [134]), self-concept (self-description questionnaire II [135]), self-rated mental health, and reported status as a bullying victim or perpetrator. Substance use-related variables included binge drinking, smoking, e-cigarette use, cannabis use, and alcohol mixed with energy drink use. Although all these variables were input into analyses, only a subset of variables appeared in the final models.

5.3.3 Outliers

In some cases, missingness was imposed onto the data. Weights less than 45lbs or greater than 390lbs were marked as missing. Height less than 4ft or greater than 6ft 11 were marked as missing. These values were considered biologically improbable and were marked as missing to ensure integrity of the data. Sleep and STSB were two variables that had a number of unfeasible outliers in the dataset. The 1.5x(IQR) method was used to identify statistical outliers, and these cut-offs were considered alongside biological plausibility in order to determine how to handle these cases. Those who reported regularly sleeping less than 4 hours a night, as well as those who reported a collective STSB greater than 16.25 hours, were marked as missing. Notably, missingness was only imposed for that particular variable; for example, those who reported less than 4 hours of sleep only had their sleep value marked as missing, but all other reported remained the same.

5.3.4 Analysis

The decision tree approach used for this study was classification and regression trees (CART) as the outcome was binary (i.e., missing vs. not missing). All models were stratified by reported sex (female, male). Consistent with decision tree approaches [145], the data were split into training and testing datasets, which contained 80% and 20% of the data, respectively. The training dataset was used to fit the tree, while the testing dataset was used to assess the prediction accuracy of the training tree. Cost complexity pruning was used alongside the 1-SE rule [145] to prune the tree, which helps correct for overfitting and yields a more parsimonious final tree. Decision tree analyses were conducted in R using the 'rpart' package and final pruned trees were visualized using the 'rattle' package. A pre-pruning restriction was set so that final nodes had to contain a minimum number of individuals. The minimum number of individuals in a school for each stratified sample was used to determine these cut-offs; which was 14 for females and 16 for males.

5.4 Results

5.4.1 Descriptive Statistics

Stratified descriptive statistics are present in Table 5.1 for any variable that appeared in at least one of the CART models. Among the whole sample (n=74,501), 31% were missing BMI. Height missingness was slightly higher among males (19%) compared to females (15%), whereas body mass missingness was slightly higher among females (22%) compared to males (20%).

5.4.2 Interpreting the CART models

Gender stratified results of the CART models are visualized in Figures 5.1 (BMI), 5.2 (Body Mass), and 5.3 (Height). All CART models can be read starting from the root node (node 1) at the top of the tree, which contains all the training data for that particular dataset. Nodes underneath node 1 represent splits in the tree, whereby a split to the left is always a 'yes' and a split to right is always a 'no'; this applies to continuous and categorical variables. The label and colour of each node, 'present' (green) or 'missing' (blue) represents which situation is more probable for data contained in that node. The shade of colour reflects the probabilities (darker colours indicate higher probability); probabilities are also included in each node, where left side shows the probability of being present, and the right side shows the probability of being missing. Variables that appear higher up in the tree (i.e., closer to node 1) and those that appear more often, can be considered more relevant criteria compared to variables that only appear once further down in the tree.

For example, in the female BMI missingness CART model (Figure 5.1), the data is first split by weight perception. If individuals in this sample indicated their weight perception as 'about right' or 'underweight', they are in node 2. Node 2 contains 74% of the sample, and in this node the probability of BMI being missing is 0.27. If individuals indicated a weight perception of 'overweight' (i.e., the other remaining category for this variable), then they are in the second node which contains 26% of the data and where the probability of BMI being missing is 0.38.

5.4.3 CART Model Accuracy

Accuracy testing using the test partition of the dataset showed that all models gained accuracy after pruning. Pruned accuracy for CART BMI models was 69% for females and 70% for males. Accuracy for CART body mass models was 78% for females and 80% for males. Accuracy for CART height models was 85% for females and 81% for males.

Table 5.1: Descriptive statistics of COMPASS study sample (2018/19)

		% (n)[2]		
Variable[1]	**Category**	**Females n=36546**	**Males N=37126**	**Total[3] N=74501**
Body Mass Index (kg/m2)	Average (SD)	20.98 (3.02)	21.21 (3.24)	21.10 (3.14)
	Missing	30.35 (11093)	31.22 (11591)	31. 31 (23329)
Height (cm)	Average (SD)	163.4(7.50)	174.2 (10.24)	168.7 (10.47)
	Missing	14.88 (5439)	19.04 (7067)	17.52 (13049)
Body mass (kg)	Average (SD)	57.42 (13.13)	66.59 (17.74)	62.16 (16.44)
	Missing	21.75 (7948)	19.79 (7348)	21.33 (15894)
Age (years)	Average (SD)	15.14 (1.50)	15.18 (1.51)	15.16 (1.51)
	Missing	0.08 (31)	0.19 (69)	0.73 (541)
Ethnicity	Racialized	69.45 (25383)	68.62 (25477)	68.48 (51017)
	Non-Racialized	30.27 (11063)	30.99 (11505)	30.63 (22822)
	Missing	0.27 (100)	0.39 (144)	0.89 (662)
Weight Perception	Underweight	11.47 (4190)	21.00 (7795)	16.30 (12140)
	Overweight	25.85 (9448)	19.93 (7398)	22.87 (17038)
	About Right	61.14 (22343)	57.19 (21233)	58.92 (43893)
	Missing	1.55 (565)	1.89 (700)	1.92 (1430)
Diet Variables				
Fruit/Vegetable Servings (24-hour recall)	Average (SD)	2.89 (1.89)	3.06 (2.11)	2.98(2.01)
	Missing	2.44 (890)	4.74 (1759)	3.79 (2822)

Table 5.1 continued from previous page

Meat/Alternatives Servings (24-hour recall)	Average (SD)	1.88 (1.03)	2.41 (1.20)	2.15 (1.15)
	Missing	2.45 (896)	4.76 (1766)	3.80 (2833)
Breakfast Consumption (Days per week)	Average (SD)	4.67 (2.37)	5.05 (2.33)	4.85 (2.36)
	Missing	1.31 (479)	2.30 (855)	1.99 (1484)
GrainServings (24-hour recall)	Average (SD)	2.41 (1.52)	2.98 (1.93)	2.69 (1.77)
	Missing	2.33 (851)	4.61 (1711)	3.67 (2737)
Milk/Alternatives Servings (24-hour recall)	Average (SD)	1.77(1.32)	2.39(1.54)	2.08 (1.47)
	Missing	2.33 (853)	4.57 (1697)	3.66 (2724)
Fast Food Consumption (Days per week)	Average (SD)	1.19(1.34)	1.43(1.61)	1.31 (1.49)
	Missing	1.03 (380)	2.16 (801)	1.81 (1345)
Movement Variables				
Sports Participation	Yes	56.70 (20720)	62.05 (23036)	59.24 (44135)
	No	41.70 (15241)	35.25 (13088)	38.41 (28618)
	Missing	1.60 (585)	2.70 (1002)	2.35 (1748)
Strength Training (Days per week)	Average (SD)	2.24 (2.02)	2.77 (2.27)	2.51 (2.16)
	Missing	1.29 (473)	1.93 (717)	1.80 (1344)
Physically Active Friends (number)	Average (SD)	3.03 (1.68)	3.52 (1.69)	3.28 (1.71)
	Missing	1.35 (494)	2.13 (789)	1.92 (1430)
Screen Time Sedentary Behaviour (Hours per day)	Average (SD)	5.92 (3.35)	6.37 (3.37)	6.15 (3.37)
	Missing	4.41 (1613)	5.94 (2206)	5.44 (4056)
Moderate to Vigorous Physical Activity (Hours per day)	Average (SD)	1.60(1.23)	2.00(1.47)	1.80(1.38)
	Missing	1.87 (683)	2.56 (949)	2.39 (1777)
Sleep (Hours per night)	Average (SD)	7.47 (1.30)	7.60 (1.28)	7.54 (1.29)
	Missing	7.33 (2679)	8.92 (3310)	8.38 (6241)
Academic Variables				
English Grade	<50%	1.09 (399)	2.44 (907)	1.83 (1362)
	$\geq$ 50%	95.39 (34862)	91.92 (34128)	93.41 (69590)
	Missing	3.52 (1285)	5.63 (2091)	4.76 (3549)
Mental Health Variables				
Self-rated Mental Health	Average (SD)	2.76 (1.21)	2.21 (1.15)	2.49 (1.21)
	Missing	3.37 (1230)	6.05 (2245)	4.93 (3670)
Wellbeing	Average (SD)	31.78 (5.75)	32.64 (5.60)	32.19 (5.72)
	Missing	4.84 (1770)	6.78 (2518)	6.02 (4486)
Self-Concept	Average (SD)	11.79(4.69)	9.76(4.19)	10.79(4.58)
	Missing	3.34 (1221)	5.51 (2045)	4.64 (3455)
Substance Use Variables				
Smoking (last 30 days)	Yes	6.64 (2425)	8.00 (2969)	7.43 (5532)
	No	92.89 (33949)	91.01 (33790)	91.70 (68320)
	Missing	0.47 (172)	0.99 (367)	0.87 (649)

Table 5.1 continued from previous page

E-cigarette use (last 30 days)	Yes	25.48 (9312)	30.34 (11264)	27.99 (20852)
	No	73.75 (26951)	67.98 (25237)	70.62 (52614)
	Missing	0.77 (172)	1.68 (625)	1.39 (1035)
Cannabis Use (last 30 days)	Yes	10.95 (4001)	14.70 (5458)	12.97 (9662)
	No	88.06 (32183)	83.36 (30950)	85.42 (63637)
	Missing	1.00 (362)	2.32 (718)	1.61 (1202)

[1] Table only includes variables present in at least one of the final CART models.
[2] Average (SD) rather than %(n) where indicated in category column
[3] Includes those who did not report sex, so sex-stratified counts may not add to total counts

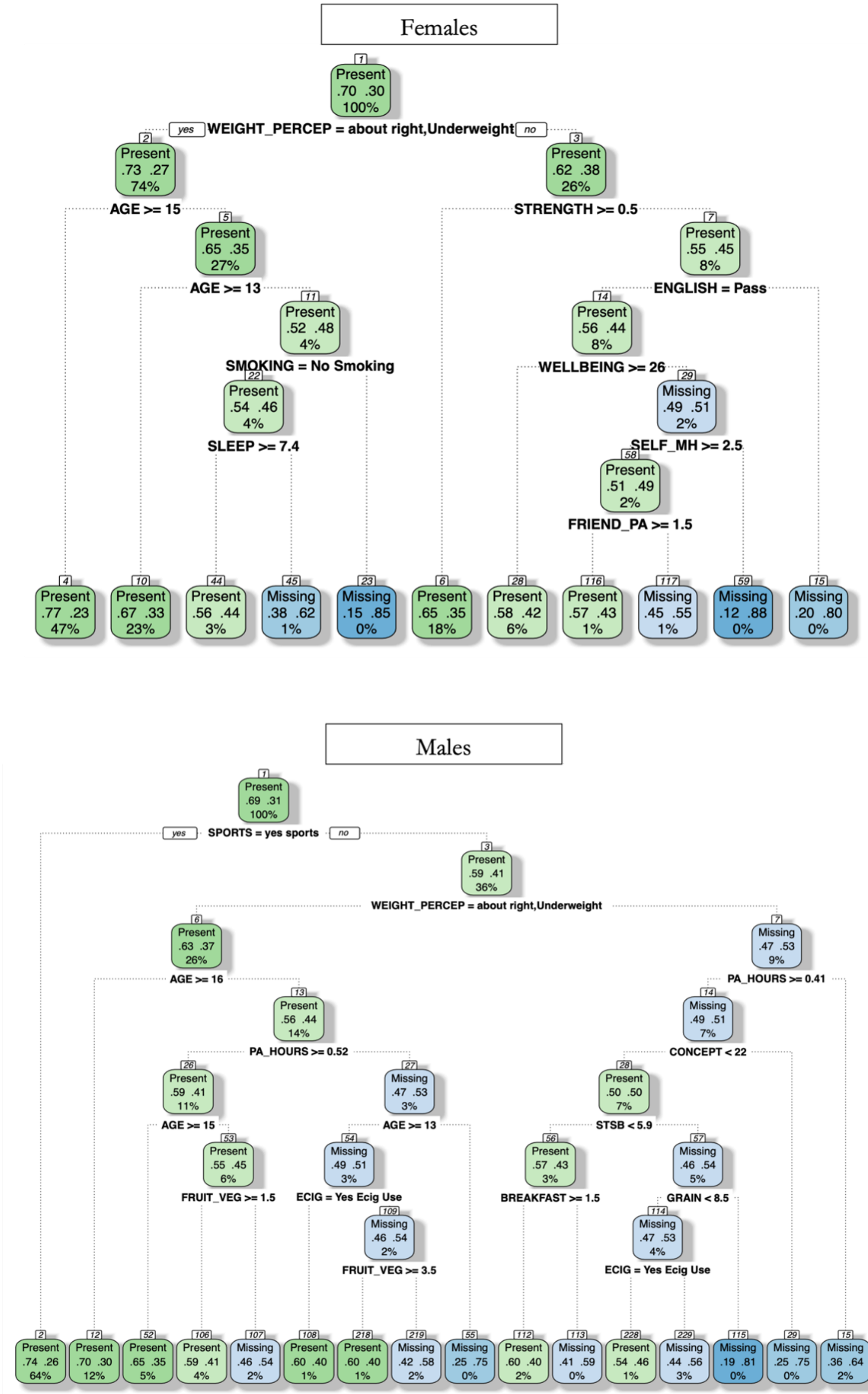


Figure 5.1: BMI Missingness CART Models for Females and Males (COMPASS 2018/19)

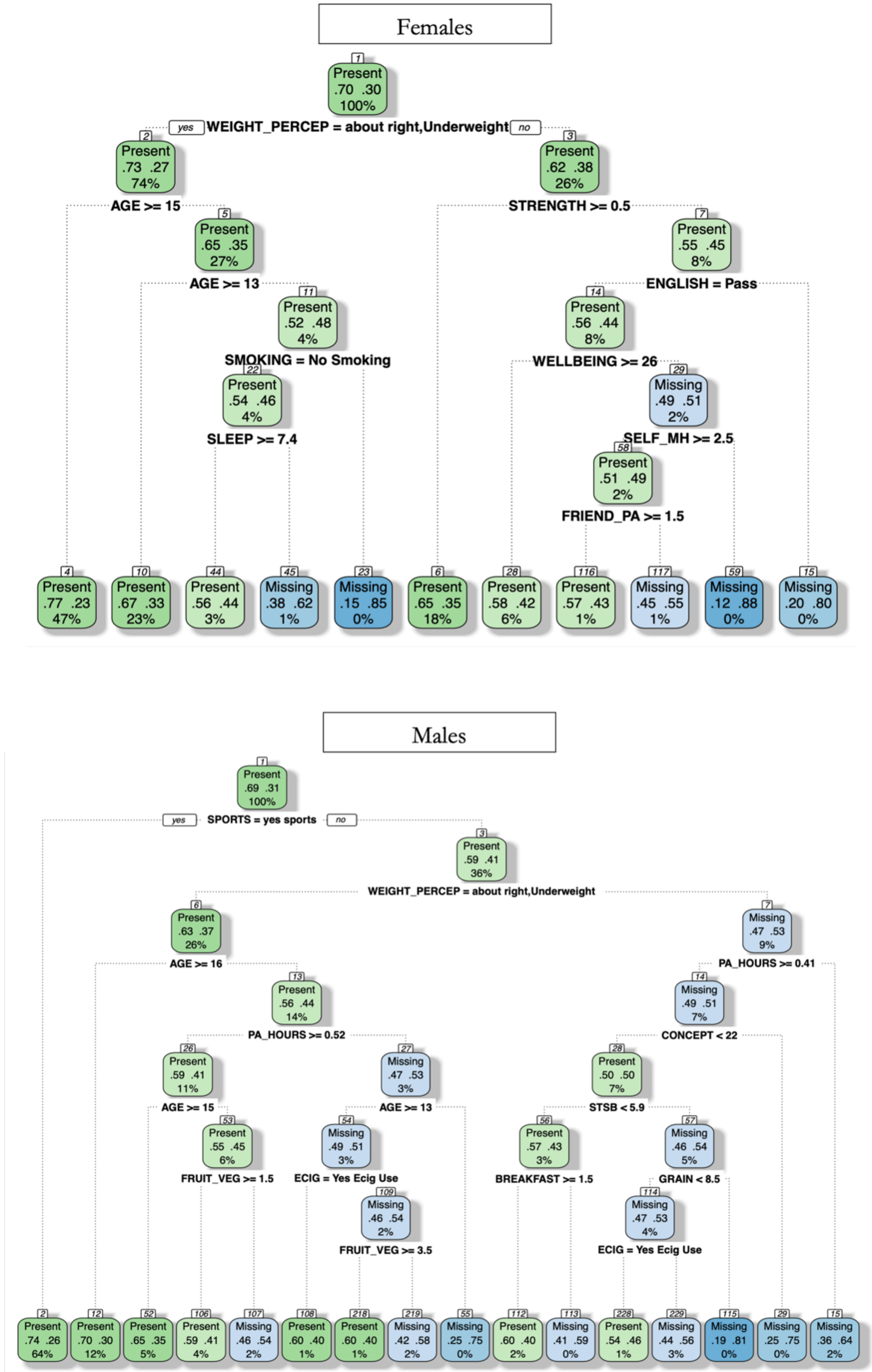


Figure 5.2: Body Mass Missingness CART Models for Females and Males (COMPASS 2018/19)

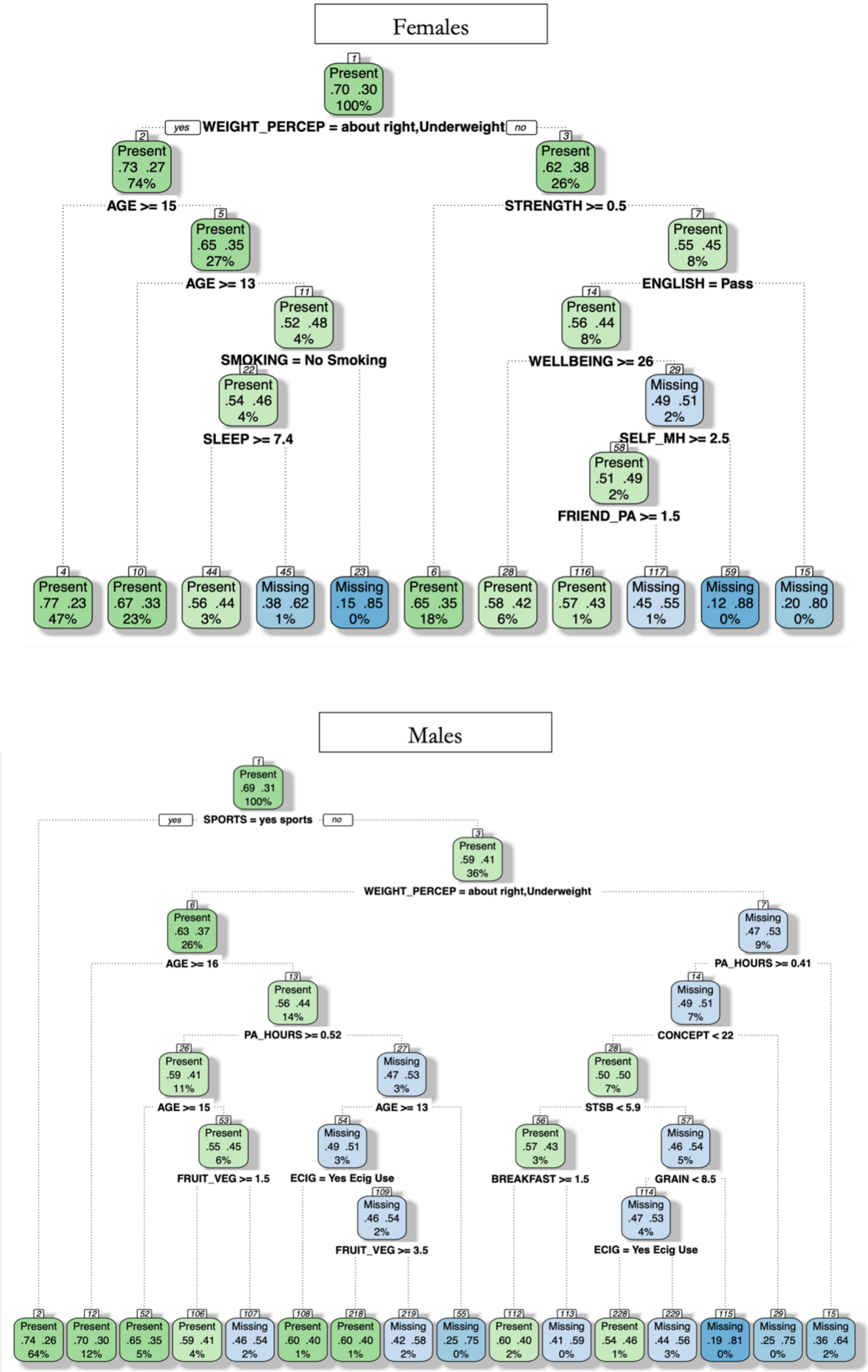


Figure 5.3: Height Missingness CART Models for Females and Males (COMPASS 2018/19)

5.5 Discussion

This study used a decision tree approach to examine missingness in BMI, height, and body mass in a large sample of Canadian youth. One of the aims of this study was to inform the structure of missingness in these variables, as youth self-reported height and body mass are problematically missing in high proportions but the literature lacks examination of this missingness. The other aim of this study was to employ a newer decision tree method to examining missingness in a dataset in order to overcome some of the barriers of regression-based approaches. The decision tree approach used in this study proved a useful approach, yielding many insights about the mechanisms of missingness present in this sample which can be useful to inform future studies on youth OWOB.

5.5.1 Mechanisms of BMI, height, and body mass missingness

In our developed BMI missingness CART models, age and weight perception were among the first few primary splits for both males and females. Previous research has suggested that those who are younger are more likely to be missing BMI values due to lack of knowledge of their own height and body mass [108], and this is consistent with what is observed in the CART models, as each split by age led to a node with a higher likelihood of missingness for the younger groups. Weight perception consistently split those who perceived themselves as 'overweight' separate from their 'about right' and 'underweight' counterparts, whereby there was a higher likelihood of missingness in the group who said they perceived themselves as overweight. Previous studies examining BMI missingness mechanisms have not included a measure of weight perception, but two studies have found that poorer body satisfaction was associated with greater likelihood of missing BMI [106, 107].

Physical activity was also one of the first few splits in both the male and female models. In the female model, strength training was identified as important split criteria, where those who did not do any strength training on average were more likely to be missing BMI. A similar mechanism was observed for males, but with sports as well as hours of physical activity; not playing sports or having less average physical activity per day led to splits where there was higher likelihood of BMI being missing. This is consistent with previous research which has included some measures of physical activity [104–106]. Mental health related variables also appeared in both male and female models. For females, wellbeing and self-rated mental health were used for splitting, and for males, self-concept was used. For all these mental health-related variables, worse scores (i.e., scores indicating poorer mental health) were associated with a higher likelihood of missing BMI. Some diet-related, substance use, and screen time or sleep behaviours were also present in the BMI missingness CART models further down the tree.

The consistent splitting of individuals who perceived themselves as overweight into a separate group more likely to be missing BMI, if one assumes at least some level of accuracy in self-perception, suggests that those with a higher BMI are more likely to be non-reporters. Findings related to physical activity support this, as those with less physical activity were also split into groups more likely to be missing BMI, and inverse associations between physical activity and BMI are well established [70,82] . These findings, taken alongside what is already known about heightened body image concerns during adolescence [112], demonstrate a clear pattern that social desirability is likely playing a large role in youth nonreporting of height and body mass in this sample.

Height and body mass missingness CART models showed some similar split criteria to the BMI missingness models, where age was a common partitioning variable, and physical activity, diet, mental health, and substance use variables were also observed. One finding exclusive to the body mass missingness models was ethnicity, whereby the model indicated individuals from racialized populations were more likely to be missing body mass values. Interestingly, although weight perception was identified as a key variable for BMI missingness, it was not identified as important in the body mass missingness CART models for males and females.

5.5.2 Utility of CART in examining BMI, height, and body mass missingness

The decision tree approach used in this study appears to have several advantages over using traditional regression approaches to examine missingness. The visual nature of decision tree models make them particularly useful for researchers to understand how missingness might be influenced by other variables. For example, the focus on weight perception and physical activity in the CART models hints that the missingness in BMI is not missing at random (NMAR [12]), and is probably more likely among those who are of a higher BMI. In the missing data literature, NMAR indicates that missingness is related to value itself [15]. Future OWOB research should consider where mechanisms and degree of missingness are observed as in this study (i.e., there is indication that NMAR missingness mechanisms are present), certain statistical approaches (e.g., complete case analysis) may not be ideal due to the risk of bias [39].

While a regression model could similarly highlight the associations between predictor variables and BMI missingness (e.g., observing a positive odds ratio for overweight weight perception), one advantage of the CART models is the hierarchy of variable importance that can be observed. In the BMI CART model, weight perception being among the top two splits for males and females indicates that this variable is of primary importance in predicting BMI missingness. The authors previously examined BMI missingness using regression [179], and while weight perception was significantly associated with missingness, it was only one of many significant variables and relative importance couldn't be discerned. Another advantage of CART models is that one can follow through a decision tree order to identify potentially important subgroups. For example, in the male BMI missingness tree, the 9% of this sample who did not participate in sports and perceived themselves as overweight were more likely to be missing BMI. Moreover, following subgroups to the bottom of the trees reveals that overall, those who perceive themselves as overweight and who were worse off in terms of their physical activity, dietary behaviours, academics, and mental health are almost certain to be missing BMI.

Missing data examinations are often the first step in applying certain statistical approaches, such as multiple imputation. Such examinations are needed to identify auxiliary variables that can inform reasonable imputed values; however, the selection of these variables can be somewhat difficult if there are many variables related to missingness. This was the case with the authors previous work using regression; almost all variables were significantly associated with missingness in BMI, height, and body mass, and comparing the effective sizes would not be appropriate as these variables are all on different scales. However, the hierarchical nature of CART models could help make the auxiliary variable selection process more systematic. For example, CART models can parse out redundant variables; while previous regression work identified weight goal as significantly related to BMI missingness [179], the CART models in this study did not perform any splits based on this variable; this may be because BMI missingness is sufficiently explained by weight perception variable alone.

This study has demonstrated the potential utility of using CART models to examine missingness in youth height, body mass, and BMI. However, missingness is a pervasive problem throughout the social sciences, and a similar approach may be useful in many other applied research domains. Moreover, public availability of machine learning packages in R as well as a wealth of online resources makes this approach reasonably accessible and feasible for applied researchers.

5.6 Conclusion

This study adds to the limited existing research examining missingness in BMI, height, and body mass among youth. CART models demonstrated that age, overweight weight perception, and indicators of lower physical activity were among the top variables partitioning variables, meaning that these were important variables related to missingness in BMI, height, and body mass. The direction of model partitioning for these variables suggests some social desirability in nonreporting of height and body mass, whereby it is likely that those who have a higher BMI are more likely to have missing height and body mass data. Importantly, this suggests an MNAR mechanism of missingness for BMI, and future research using these types of self-reported data among youth should closely examine missingness to select an appropriate statistical approach for handling missing data.

Chapter 6

Study 3: Assessing the impact of missing data in youth overweight and obesity research: complete case analysis versus multiple imputation

Authors: Amanda Doggett[1], Ashok Chaurasia[1], Jean-Philippe Chaput[2,3], Scott T. Leatherdale[1]

[1]School of Public Health Sciences, University of Waterloo, Waterloo, Ontario, Canada
[2]Department of Pediatrics, University of Ottawa, Ottawa, Ontario, Canada
[3]Healthy Active Living and Obesity Research Group, Children's Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada.

Status: Submitted to International Journal of Social Research Methodology

6.1 Overview

Youth overweight and obesity (OWOB) surveillance often uses body mass index (BMI) derived from self-reported height and weight, but these measures can suffer from high proportions of missing data. Complete case analysis (CCA) is the most common approach to handle missing data, but this approach can introduce bias if missing data are not missing completely at random. Using BMI and related covariate data from 36,546 female and 37,126 male youth aged 12-19 years who participated in the COMPASS study in 2018/19, where approximately 30% of BMI data were missing, results and inference were compared between CCA and multiple imputation (MI) approaches to examine associations with youth BMI. Results of regression joint models showed contrasting findings between MI and CCA, highlighting that appropriate methodological choices on the handling of missing data are essential in youth OWOB research, and that choices can impact research inference and thereby associated policy and programming recommendations.

6.2 Introduction

6.2.1 Youth Overweight and Obesity

Global prevalence of individuals with overweight or obesity (OWOB) is high, with little indication of improvement [56]. OWOB is associated with many chronic diseases including cardiovascular diseases, type 2 diabetes, and cancer [59], as well as poorer mental and social wellbeing [60]. Literature has indicated that OWOB trajectories are established at a young age, as the majority of youth with OWOB continue to have OWOB during adulthood [63, 64]. Treating OWOB has proven to be challenging, and even when treatment begins during childhood, programs are often difficult to implement, and efficacy is inconsistent [65]. As such, there has been a public health shift from a focus on treatment for OWOB, to a focus on addressing the upstream factors that lead to youth OWOB in order to inform prevention efforts [66, 67]. Many factors have shown to be associated with youth BMI or OWOB, including diet-related factors, such as breakfast consumption [72], movement related behaviours such as physical activity and sleep [82,86], substance use behaviours [89], and mental health indicators [93].

6.2.2 Missing Data in Youth OWOB

A sizable portion of the youth OWOB literature examining upstream factors relies on surveillance data using traditional survey-based methods, as self-report measures are substantially more feasible and cost-effective compared to objective measurements. Although the use of BMI has drawbacks, researchers generally agree that in absence of better alternatives, self-report measures of height and weight are acceptable for the purposes of population surveillance and epidemiological research [108, 119]. However, a downside of these measures is the propensity for missing data. Self-report height and weight are unique in that they tend to be missing in high proportions that greatly exceed levels of missingness observed for other health indicators [54, 55], particularly among youth. In this domain, the high proportions of missing data likely stem from a combination of factors including lack of knowledge of one's height and weight, and social desirability bias surrounding these measures that leads to nonreporting [106, 108].

6.2.3 Missing Data Mechanisms and Methods

Due to the voluntary nature of research, the presence of at least some missing data is largely unavoidable for measures that might be linked to social desirability. As such, researchers must develop and use appropriate methods for handling missingness post-data collection through statistical techniques. The most common approach to handling missing data in applied research is complete case analysis (CCA) [35, 36], which involves deleting missing cases such that only complete cases are included in statistical analyses [13]. Alternatively, methods such as multiple imputation (MI) algorithmically impute missing data using other available information. The appropriate choice of a missing data approach largely depends on the degree and mechanism of missingness in a dataset. Broadly, missing data can occur under three main mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). MCAR assumes that missing data follow no systematic pattern, and that the data set with missingness is a simple random sample of the hypothetical complete sample. MAR assumes that missing data depend on observed/measured values of variables within the study, but not the missing value itself, and NMAR assumes that the missing value depends on the missing value itself. For example, an individual not reporting their weight because it is higher than average would be a NMAR scenario, whereas missing weight due to their observed/measured gender (regardless of what their weight was) would be a MAR scenario. In practice, research can often face many challenges related to data collection, and the MCAR mechanism is unlikely, particularly in situations where missing data are highly prevalent or could be linked to reporting biases (e.g., social desirability).

6.2.4 Applied Research Methods

In the statistical literature, simulation studies have demonstrated that CCA can produce biased estimates where mechanisms of missing data are not MCAR, whereas MI (if used under appropriate circumstances) can produce unbiased and efficient estimates [18, 39, 43]. Yet, systematic reviews have demonstrated that CCA remains the most common approach to deal with missing data, and that missing data reporting across studies is generally poor, and frequently non-existent [35, 36]. These reviews demonstrate that when it comes to missing data, there is a gap between 'best practices' actual practices in applied research. Simulation studies are a robust approach to test statistical methods, but a drawback is that they may not reach applied research audiences, and often don't discuss potential impacts of incorrect inference in the context of any specific domain. In contrast, application of different methods using real-world data may be more relevant and accessible for those who apply these statistical methods. For example, one real-world study that examined risk factors for breast cancer found that the use of MI identified two predictors of breast cancer that were missed with CCA, demonstrating that CCA had likely led to a type II error [45]. Another study examining missing data methods on perinatal depression found that where CCA indicated associations with child BMI, MI found no such association - indicating that the CCA result is likely a type I error [46]. If the CCA results from these studies had been presented alone, the associated findings and recommendations would likely have been inappropriate, and would potentially misinform future decisions surrounding practice and research.

6.2.5 Study Aims

The high proportions of missingness observed in youth-reported height and weight, in combination with historically poor reporting and handling of missing data within cohort studies in general, highlights that there is a gap in this field related to missing data that requires thoughtful solutions. This study aims to use real-world data on youth BMI data to explore the different results produced between CCA and MI. This study also aims to illustrate how these differences may result in contrasting research conclusions.

6.3 Methods

6.3.1 Sample

This study uses a cross-sectional sample consisting of 36 546 female and 37 126 male youth aged 12-19 years who participated in the 2018/19 wave of the COMPASS study. Briefly, the COMPASS study (COMPASS) is a longitudinal cohort study of youth in Canada, which collects data across several provinces. COMPASS uses purposive sampling at the school-level and applies an active-information passive consent protocol. All information collected about students and their behaviours is self-reported; for the 2018/19 data collection, the questionnaire was paper-based and administered during class time. Full details describing the COMPASS study are available elsewhere in print [124] and online www.compass.uwaterloo.ca.

6.3.2 Measures

The outcome of interest, BMI (kg/m^2), was measured via self-reported height and weight using a validated approach [182]. Weight was determined by asking students, "how much do you weigh without your shoes on? (please write your answer in pounds OR in kilograms, and then fill in the appropriate numbers for your weight.)" Height was similarly determined by asking, "how tall are you without your shoes on? (please write your height in feet and inches OR in centimeters, and then fill in the appropriate numbers for your height)."

This study selected several variables available from COMPASS which have been associated with youth BMI or OWOB in previous literature. Control variables included age (continuous [12-19 years old]) and ethnicity (racialized, non-racialized). The diet-related variables of interest included average fast-food consumption (times per week) and average breakfast consumption (times per week). Movement-related variables of interest included average physical activity (hours per day), current participation in sports (yes, no), average screen time sedentary behaviour (hours per day), and average sleep duration (hours per day). Mental health variables of interest were measured via validated self-report scales and included depression (via the Center for Epidemiologic Studies Depression Scale Revised (CESD-R-10) [131] , anxiety (via the Generalized Anxiety Disorder 7-item (GAD-7) [132]) and self-concept (via the Self Description Questionnaire II [135]). For the mental health measures, higher scores indicated poorer outcomes; for depression and anxiety, scores higher than 10 indicated what may be considered as clinically relevant symptoms of depression or anxiety. Substance use variables of interest included current binge drinking, current cannabis use, and current e-cigarette use. Substance use measures were binary (yes, no) and consistent with previous research, 'current' was defined as use at least once in the past 30 days [136].

6.3.3 Data Preparation

Missing data have been previously examined in depth in this sample [179]. For this study, degree of missing data is presented for each variable of interest graphically. While most missing data were present in the dataset due to non-response, some missingness was researcher applied; unfeasible numbers for height or weight (less than 45 lbs or greater than 390 lbs, or less than 4 ft or greater than 6 ft 11) were marked as missing. Sleep duration also had several biologically unfeasible values, and outliers were identified and marked as missing using the interquartile range. The complete case analysis sample was created by removing any case with missing data on any of the variables listed in section 6.3.2.

6.3.4 Imputation

Multiply imputed datasets were created in R using mice and miceadds packages [148, 149]. The imputation procedure was multi-level, accounting for clustering at the school level and produced 30 imputations over 10 iterations. Auxiliary variables for the imputation (variables which inform the imputation but are not present in the analysis model) were selected based on findings from a previous study which identified available variables most associated with missing BMI, height, or weight (Doggett et al., n.d.). The predictor matrix was organized such that variables with the least amount of missingness were imputed first. Passive imputation was used to impute BMI, predictive mean matching was used to impute categorical variables, and the pan method was used to impute all continuous variables. Imputations were checked for acceptable convergence, and other checks to assess the quality of the imputed values were also performed, as outlined by van Burren [8]. Congenial with the analysis models, imputations for males and females were performed separately.

Notably, MI as it was performed in this study operates under the MAR assumption. Previous research examining associations between variables and missing BMI has suggested that missing BMI is NMAR because those not reporting BMI likely have a higher BMI [179]. However, NMAR models require information that is not often available, such as a close estimate of the degree of difference between missing and non-missing values. Standard practice in this scenario is to appropriately include enough auxiliary variables to increase the plausibility of the MAR assumption [8], which is the approach that was used in this study.

6.3.5 Analyses

Descriptive statistics in the form of mean and standard errors were calculated for CCA and MI samples. Linear mixed effects models (LMM) examining associations between variables of interest (listed in Section 6.3.2) and BMI were conducted on CCA and MI samples, specifying clustering at the school level. All analyses were performed separately for males and females. For imputation models, standard practice was followed for descriptive and mixed models, whereby modelling was performed separately on each imputed dataset and estimates were pooled according to Rubin's rules [24]. Additionally, sensitivity analysis was performed by running GEE models of imputed data as an alternative to LMM; findings were similar and as such only LMM results are presented.

6.4 Results

The stratified samples for this study consisted of 36 546 females and 37 126 males. Figure 6.1 presents the missingness by case and variable for each sample. Height and weight (and subsequently derived BMI) show the greatest proportions of missing data across females and males; 30% of BMI data were missing for females, and 31% for males. Depression and anxiety scores as well as sleep duration also showed high levels of missingness. Looking horizontally across Figure 6.1 demonstrates the overlap in missing variable data for each case.

Table 6.1 presents descriptive statistics for each sample. After deleting cases with missing data (on any variable), CCA samples dropped to 19 116 females and 17 979 males, representing 52% and 48% of the original samples, respectively. MI sample sizes are the same as original samples sizes since all variables were imputed. For both females and males, MI descriptive statistics indicate that mean BMI was over half a unit higher compared to statistics from CCA.

Tables 6.2 (females) and 6.3 (males) present the results of LMMs examining associations between BMI and variables of interest from CCA and MI. Differences between MI and CCA are visually presented in Figure 6.2, which shows point and confidence interval estimates from LMM. The purpose of Figure 6.2 is to illustrate the difference in point estimates and confidence intervals between the CCA and MI models; the estimates between different variables should not be directly compared as all variables are not on the same scale. For females, MI identified no association between anxiety score or binge drinking on BMI, whereas CCA indicated an association. Conversely, MI identified a significant association between cannabis use and female BMI, whereas CCA found no association. For males, MI identified significant associations between anxiety score as well as screen time and BMI, while CCA found no associations. MI found no association of depression score, or e-cigarette use on male BMI, whereas CCA indicated an association. Sports participation for males was significantly associated with BMI in both MI and CCA models, but directionality of the effect size was reversed. Notably, many effect sizes in the models presented in Tables 6.2 and 6.3 are small; however, this is to be expected given that continuous BMI scores (as opposed to OWOB status) is being modelled.

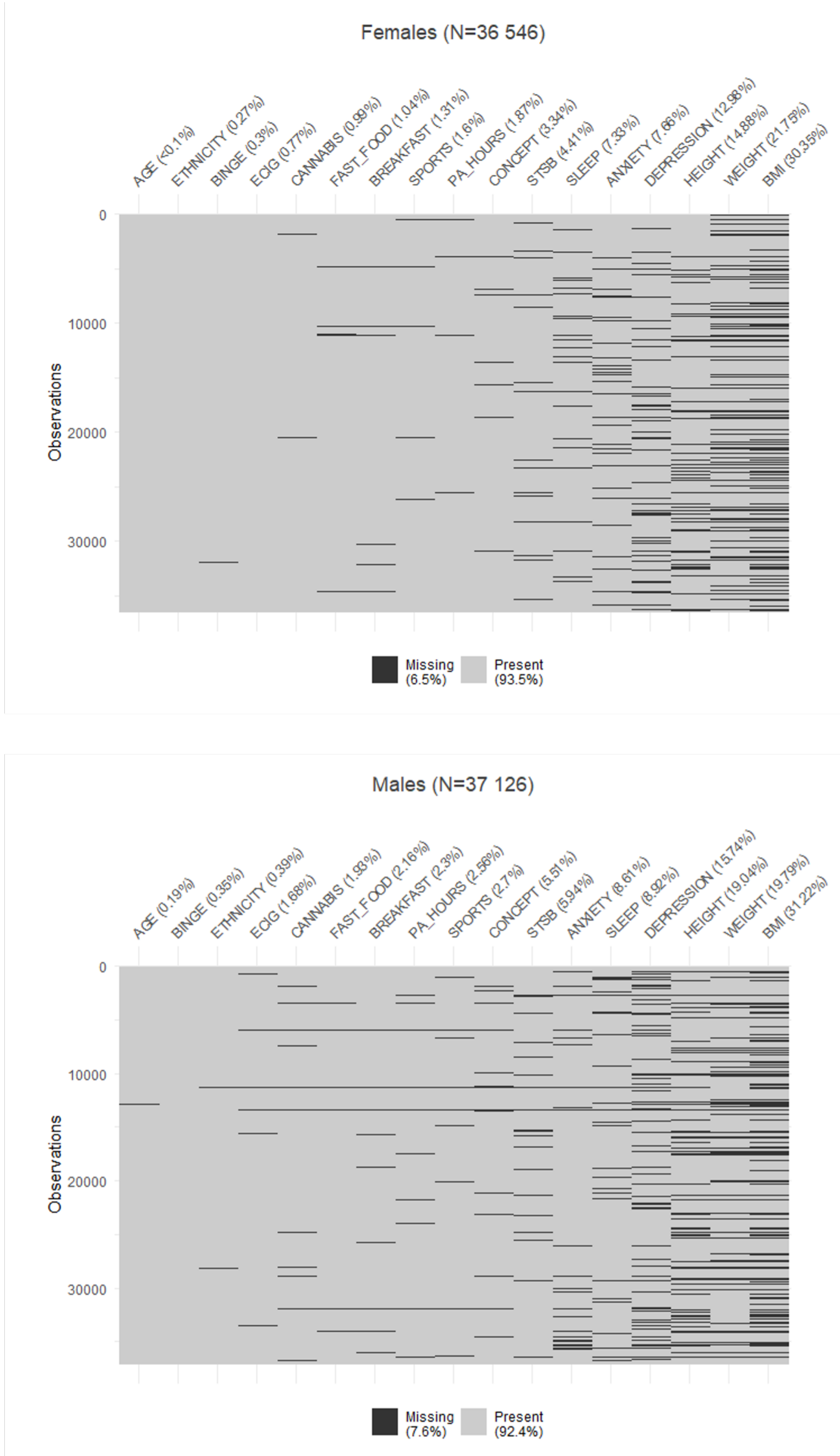


Figure 6.1: Missingness in COMPASS 2018/19 study sample by observation and variable. Variables are sorted from least missing to most missing, left to right
BMI, body mass index; BINGE, binge drinking; ECIG, e-cigarette use; PA_HOURS, physical activity hours; STSB, screen time sedentary behaviour

Table 6.1: Mean Estimates of Complete Case Analysis and Multiple Imputation Procedures (COMPASS 2018/19, Females and Males)

	Females		Males	
Variable	**Complete Case Analysis N = 19 116**	**Multiple Imputation N= 36 546**	**Complete Case Analysis N = 17 979**	**Multiple Imputation N= 37 126**
	Mean (SE)		**Mean (SE)**	
BMI (kg/m2)	21.0 (0.02)	21.5 (0.03)	21.2 (0.02)	21.9 (0.03)
Height (cm)	163.5 (0.05)	163.3 (0.04)	174.2 (0.07)	173.7 (0.06)
Weight (kg)	56.3 (0.07)	57.6 (0.08)	64.9 (0.1)	66.6 (0.1)
Age (years)	15.3 (0.01)	15.1 (0.01)	15.3 (0.01)	15.2 (0.01)
Ethnicity[1]				
Racialized	0.3 (0.003)	0.3 (0.01)	0.3 (0.003)	0.3 (0.01)
Non-racialized	0.7 (0.003)	0.7 (0.01)	0.7 (0.003)	0.7 (0.01)
Sports[1]				
Participates in sports	0.6 (0.004)	0.6 (0.01)	0.7 (0.003)	0.6 (0.01)
Does not participate in sports	0.4 (0.004)	0.4 (0.01)	0.3 (0.003)	0.4 (0.01)
Moderate to vigorous physical activity (average hours per day)	1.6 (0.01)	1.6 (0.01)	2.0 (0.01)	2.0 (0.01)
Screen time sedentary behaviour (average hours per day)	5.7 (0.02)	6.0 (0.02)	6.2 (0.02)	6.4 (0.02)
Sleep (average hours per night)	7.5 (0.01)	7.5 (0.01)	7.6 (0.01)	7.6 (0.01)
Fast food consumption (times per week)	1.1 (0.01)	1.2 (0.01)	1.4 (0.01)	1.4 (0.01)
Breakfast consumption (times per week)	4.9 (0.02)	4.7 (0.01)	5.3 (0.02)	5.1 (0.01)
Anxiety (GAD7 Scale)	7.5 (0.04)	7.7 (0.03)	4.3 (0.03)	4.6 (0.03)
Depression (CESD-R-10 Scale)	9.7 (0.04)	10.1 (0.03)	6.9 (0.04)	7.4 (0.03)
Self-Concept (Self-Description Questionnaire II)	11.4 (0.03)	11.8 (0.02)	9.4 (0.03)	9.8 (0.02)
Binge Drinking[1]				
Binge drinking in past 30 days	0.2 (0.003)	0.2 (0.01)	0.2 (0.003)	0.2 (0.01)
No binge drinking in past 30 days	0.8 (0.003)	0.8 (0.01)	0.8 (0.003)	0.8 (0.01)
Cannabis Use[1]				
Cannabis use in past 30 days	0.1 (0.002)	0.1 (0.01)	0.1 (0.003)	0.2 (0.01)
No cannabis use in past 30 days	0.9 (0.002)	0.9 (0.01)	0.9 (0.003)	0.9 (0.01)
E-cigarette Use[1]				
E-cigarette use in past 30 days	0.3 (0.003)	0.3 (0.01)	0.3 (0.003)	0.3 (0.01)
No e-cigarette use in past 30 days	0.7 (0.003)	0.7 (0.01)	0.7 (0.003)	0.7 (0.01)

[1] The mean of binary variables can be interpreted prevalence.
SE, standard error; BMI, body mass index; GAD7, generalized anxiety disorder 7-item; CESD-R-10, center for epidemiologic studies depression scale revised.

Table 6.2: Linear Mixed Model Estimates of Complete Case Analysis and Multiple Imputation Procedures (COMPASS 2018/19, Females)

	Complete Case Analysis N = 19 116		Multiple Imputation N = 36 546	
Predictor	**Estimate (95% CI)**	**p-value**	**Estimate (95% CI)**	**p-value**
Age (years)	0.47 (0.44 – 0.50)	<0.001	0.54 (0.5 - 0.58)	<0.001
Ethnicity				
Racialized	0.07 (-0.04 – 0.17)	0.208	0.33 (0.2 - 0.45)	<0.001
Non-racialized	-	-	-	-
Sports				
Participates in sports	-0.01 (-0.10 – 0.08)	0.847	-0.07 (-0.18 - 0.04)	0.198
Does not participate in sports	-	-	-	-
Moderate to vigorous physical activity (average hours per day)	0.01 (-0.03 – 0.05)	0.542	0.03 (-0.01 - 0.08)	0.156
Screen time sedentary behaviour (average hours per day)	0.04 (0.02 – 0.05)	<0.001	0.05 (0.03 - 0.07)	<0.001
Sleep (average hours per night)	0 (-0.04 – 0.04)	0.918	0.02 (-0.03 - 0.07)	0.436
Fast food consumption (times per week)	-0.09 (-0.12 – -0.05)	<0.001	-0.08 (-0.12 - -0.03)	<0.001
Breakfast consumption (times per week)	-0.07 (-0.09 – -0.05)	<0.001	-0.08 (-0.11 - -0.06)	<0.001
Anxiety (GAD7 Scale)	-0.02 (-0.03 – -0.01)	0.003	-0.01 (-0.03 - 0)	0.125
Depression (CESD-R-10 Scale)	0 (-0.01 – 0.01)	0.641	0 (-0.01 - 0.02)	0.634
Self-Concept (Self-Description Questionnaire II)	0.08 (0.07 – 0.09)	<0.001	0.1 (0.09 - 0.12)	<0.001
Binge Drinking				
Binge drinking in past 30 days	0.13 (0.01 – 0.26)	0.041	0.05 (-0.11 - 0.21)	0.539
No binge drinking in past 30 days	-	-	-	-
Cannabis Use				
Cannabis use in past 30 days	-0.03 (-0.18 – 0.13)	0.74	0.40 (0.21 - 0.59)	<0.001
No cannabis use in past 30 days	-	-	-	-
E-cigarette Use				
E-cigarette use in past 30 days	0.06 (-0.05 – 0.17)	0.249	-0.05 (-0.18 - 0.09)	0.499
No e-cigarette use in past 30 days	-	-	-	-

CI, confidence interval; GAD7, generalized anxiety disorder 7-item; CESD-R-10, center for epidemiologic studies depression scale revised

Table 6.3: Linear Mixed Model Estimates of Complete Case Analysis and Multiple Imputation Procedures (COMPASS 2018/19, Males)

	Complete Case Analysis N = 19 116		Multiple Imputation N = 36 546	
Predictor	**Estimate (95% CI)**	**p-value**	**Estimate (95% CI)**	**p-value**
Age (years)	0.52 (0.49 - 0.56)	<0.001	0.58 (0.54 - 0.62)	<0.001
Ethnicity				
Racialized	0.17 (0.05 – 0.28)	0.004	0.34 (0.22 - 0.48)	<0.001
Non-racialized	-	-	-	-
Sports				
Participates in sports	0.24 (0.13 - 0.34)	<0.001	-0.18 (-0.3 - -0.06)	0.003
Does not participate in sports	-	-	-	-
Moderate to vigorous physical activity (average hours per day)	0.11 (0.08 - 0.15)	<0.001	0.07 (0.03 - 0.11)	<0.001
Screen time sedentary behaviour (average hours per day)	0.01 (0 - 0.03)	0.079	0.05 (0.03 - 0.07)	<0.001
Sleep (average hours per night)	-0.08 (-0.12 - -0.04)	<0.001	-0.08 (-0.14 - -0.03)	0.002
Fast food consumption (times per week)	-0.08 (-0.11 - -0.05)	<0.001	-0.05 (-0.09 - -0.02)	0.004
Breakfast consumption (times per week)	-0.12 (-0.14 - -0.09)	<0.001	-0.14 (-0.16 - -0.11)	<0.001
Anxiety (GAD7 Scale)	0 (-0.01 - 0.01)	0.951	-0.02 (-0.04 - 0)	0.024
Depression (CESD-R-10 Scale)	-0.02 (-0.04 - -0.01)	0.002	0 (-0.01 - 0.02)	0.627
Self-Concept (Self-Description Questionnaire II)	0.04 (0.02 - 0.05)	<0.001	0.08 (0.07 - 0.1)	<0.001
Binge Drinking				
Binge drinking in past 30 days	0.33 (0.2 - 0.46)	<0.001	0.47 (0.3 - 0.63)	<0.001
No binge drinking in past 30 days	-	-	-	-
Cannabis Use				
Cannabis use in past 30 days	-0.01 (-0.16 - 0.15)	0.943	0.04 (-0.14 - 0.22)	0.643
No cannabis use in past 30 days	-	-	-	-
E-cigarette Use				
E-cigarette use in past 30 days	0.25 (0.14 - 0.37)	<0.001	0.11 (-0.03 - 0.25)	0.125
No e-cigarette use in past 30 days	-	-	-	-

CI, confidence interval; GAD7, generalized anxiety disorder 7-item; CESD-R-10, center for epidemiologic studies depression scale revised

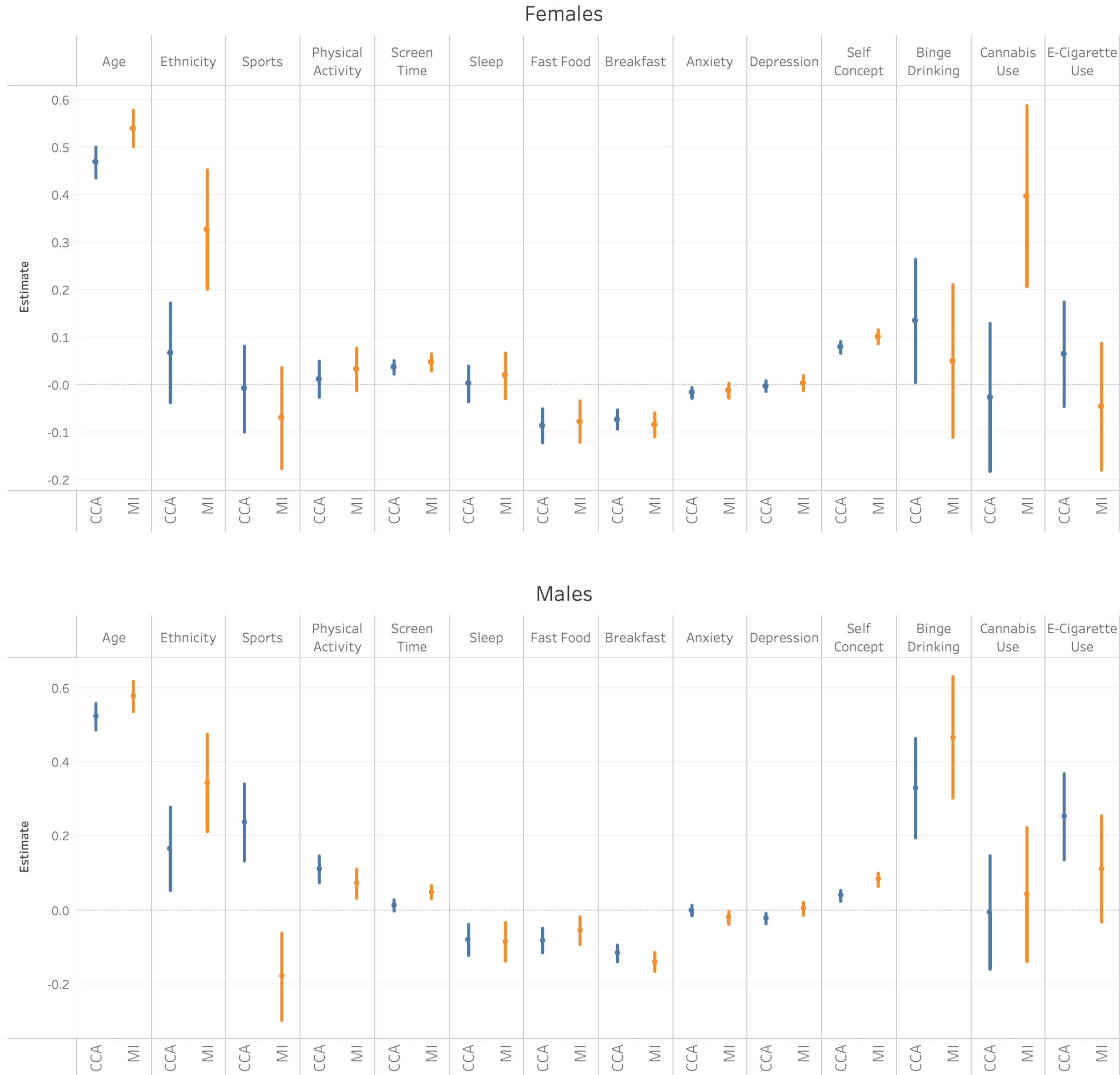


Figure 6.2: Point and confidence interval estimates for complete case analysis (CCA) and multiple imputation (MI) models (COMPASS, 2018/19). Point estimates are represented by dots, while confidence intervals are represented by vertical bars. Findings related to significance are evident in effects where their respective confidence intervals do not include a null value.

6.5 Discussion

This study compared findings between CCA and MI methods for handling missing data within models examining factors associated with youth BMI. Under the assumption that models have been properly specified and assumptions hold true, contrasting results between MI and CCA are illustrative of bias introduced by deleting cases; several instances of such bias were found in this study. First, in the descriptive statistics, deleting cases biased the average BMI of the female and male samples downwards, which aligns with expectations based on previous research that has hinted that the true values of missing BMI are likely to be higher than observed values [104,106,179]. Next, regression models showed contrasting results between MI and CCA methods. For females, MI results indicated a significant association between cannabis use and BMI, whereas CCA found no evidence of association. For males, e-cigarette use was not identified as significant by MI, but CCA indicated a significant association. Comparing these results to previous studies, the female MI model was consistent with a previous study finding a positive association between frequent cannabis use and obesity among girls only [89]; however, results of the male MI model contradict a similar study analyzing associations between obesity and e-cigarette use [183]. For males, sports participation was identified as significant in both MI and CCA models, but MI indicated a negative association, while CCA indicated a positive association. Notably, the CCA sample was made up of more sports playing individuals with a lower BMI, whereas the MI models include more non-sports playing individuals with higher average BMI. It is difficult to compare these results to existing literature, as research on youth participation in sports and body weight is mixed, likely in part due to heterogenous definition of "sports" [184]. Conflicting findings are not uncommon in youth OWOB research, and as such, if the CCA findings from this study were presented on their own, they may not necessarily have stood out as contradictory. However, when presented alongside MI models, the impact of missing data methodology choice is clear. Statistical literature indicates that CCA can introduce bias in situations where there is a substantial amount of non-random missingness [18, 39, 43], which was the case in this sample. The findings presented in this study highlight how this bias can propagate into research conclusions, and demonstrates the impact of inappropriately applying CCA under these circumstances.

Findings from this study also identified that contrasting conclusions regarding sex-differences would have been reached between the CCA and MI models. MI models indicated that anxiety was significantly associated with BMI for males but not for females. However, CCA models indicated the opposite; anxiety was significantly associated with BMI for females, but not for males. Limited research has analyzed sex-stratified associations between anxiety and youth BMI, but MI results are consistent with a previous study which found that presence of an anxiety disorder was prospectively associated with overweight status for males, but not females [185]. Notably, in this aforementioned study, objective measurements of height and weight, as well as clinical diagnosis to examine anxiety, increases confidence in these findings; such measurements would be considered more robust compared to self-report measures, and also less likely to suffer from nonreporting (although missing data was unfortunately not reported). The present study demonstrates how the bias introduced by CCA can not only change research conclusions, but also confounds our understanding of how equity-related indicators such as sex play a role in BMI and youth OWOB.

Missing data will manifest differently across studies; aspects of design, sample, data collection, etc. may all influence nonreporting. Comparing between studies can be generally difficult because of aspects such as differing data collection strategies or model inputs; the results of this study highlight that this difficulty can be compounded by differing levels, types, and handling of missing data. Unfortunately, missing data are often not reported at all in studies [36], making proper methodological assessment of these studies near impossible. Given the high degree of missing data that can be present in self-report youth height and weight, missing data and use of appropriate (or inappropriate) methods may partly explain some of the conflicting findings in youth OWOB research. As such, future reviews in this field should consider the impact that missing data methodology has when comparing results between studies. Of course, this is not to say that all disparities between studies arise from missing data, nor that the present study is a guide for true associations. However, this study has demonstrated that missing data methods are highly important in this field, and has illustrated how bias introduced by deleting cases can impact research findings in one scenario. While this study advocates for better reporting and handling of missing data, particularly in the field of youth OWOB research, it must be acknowledged that such recommendations are not necessarily simplistic to implement. MI is not recommended in all cases [39]; advanced understanding of missing data mechanisms is needed to understand which methods are appropriate in which scenarios. Moreover, in scenarios where MI is an appropriate choice, practical implementation is complex, and the simpler option of maximum likelihood estimation is more limited in options to handle missing covariate data. Given the ubiquity of missing data in much of applied research, and the impact it can have on research conclusions, there is an argument to be made that missing data methodologies should be a fundamental component of applied statistical training.

This study has some limitations to note. This study focused on item non-response (individuals participating but not answering specific questions). Unit-nonresponse, referring to youth who didn't participate in data collection, were not included, and these individuals may be systematically different than those who do participate. Unfortunately, little can be done about unit non-response in a cross-sectional context; future research in this area could consider the application of a longitudinal MI procedure in attempt to impute these cases, although multi-level imputation with three levels is not yet a fully developed practice. Another limitation is that this study focused on behavioural factors associated with youth BMI; biological data were not collected, and as such there may be important factors missing from these models. Lastly, this study used self-report data, and as such reported values may not reflect true values; however, research has suggested that the degree of misreporting for adolescent height and weight is not substantial [55, 123].

This study also has several strengths and implications for the field. This study used data from the COMPASS study, which benefits from a large sample size and application of an active information passive consent protocol. Next, this study used an applied research lens to illustrate the impact of properly applying MI, or, conversely, the impact of inappropriately applying CCA, in the context of this sample where the degree of missing data was high, and patterns indicated that data were not MCAR. Given the numerous studies that rely on similar self-reported data, coupled with high propensity for missing data in this domain, this study may have broad implications for research practice.

6.5.1 Conclusions

This study found that among a large sample of Canadian youth who participated in the COMPASS study in 2018/19, the substantial and systematic levels of missing data present in the self-report height and weight measures greatly impacted results depending on the missing data approach used. CCA and MI models produced contrasting results, which highlighted the bias of using CCA in this sample. Future studies that use self-reported youth height and weight should thoroughly examine the degree and patterns of missing data in order to determine what missing data approaches are appropriate. This type of methodological research, which uses domain-specific approach, is needed to reach applied scientists working with real-world data in order to build capacity and understanding of how statistical bias can propagate into research findings and concomitant recommendations.

Chapter 7

General Discussion

7.1 Overview

This dissertation explored the degree, patterns, and impacts of missing youth Body Mass Index (BMI), height, and weight data among participants in the 2018/19 wave of the COMPASS study. Study 1 focused on exploring the level of missing data across youth BMI, height, weight, and related factors, as well as examined associations with missing BMI, height, and weight. Study 2 examined subgroups most likely to have missing information for height, weight, and BMI by using findings of Study 1. Studies 1 and 2 illustrated that complex patterns of missing data existed, and nonreporting was not missing at random and likely influenced in part by social desirability. These findings strongly suggested that deleting cases with missing data would introduce bias into analyses, making results unreliable for statistical inference. As such, Study 3 explored the impact that different methodologies used to address missingness could have on data analyses. Study 3 illustrated that in a model predicting factors associated with youth BMI, Complete Case Analysis (CCA) produced substantially different results to Multiple Imputation (MI), highlighting the bias resulting from deleting cases. Taken together, the findings from this research indicate that improper handling of missing data can greatly impact research findings, and, in turn, impact associated policy and programming recommendations in behavioural research among youth which uses self-report measures.

7.2 Summary of Key Findings

7.2.1 Study 1

Examination of the COMPASS 2018/19 cross-sectional data set revealed that 31% (23 329 of 74 501 participants) were missing BMI. As BMI is a derived variable, missingness occurred in four ways: 1) 32% of BMI missing cases were caused by nonreporting of weight only; 2) 20% were caused by nonreporting of height only; 3) 36% were caused by nonreporting of both weight and height, and; 4) 12% were set to missing due to extreme outlier values. More females were missing weight only, while more males were missing height only. Height and weight showed the greatest level of nonreporting compared to other variables in this dataset, followed by depression, where 15% of the data were missing.

Perceiving oneself as overweight was significantly related to missing BMI for both females and males. Females who indicated they wanted to lose weight were more likely to omit reporting their weight, and males who said they wanted to gain weight were less likely to omit reporting their weight. Several other factors related to diet, movement, academic, mental health, and substance use showed significant association with BMI, height, and weight missingness. Specifically, missing BMI was significantly associated with variables that indicated poorer diet and higher levels of physical inactivity; aspects which are well established to be associated to Overweight or Obesity (OWOB) in youth [70,72,73,82]. These results, considered alongside the results that those who perceive themselves as overweight were less likely to report their weight, strongly hint that on average, those not reporting their weight likely have a higher weight than those who do report their weight. Notably, of the variables that were comparable to measures used in the four existing studies which have examined BMI or weight missignness among youth, results were consistent [104–107].

Findings from Study 1 are consistent with the literature surrounding body image during adolescence. It is known that body image concerns are heightened during adolescence, which is likely influenced by the combination of biological and social changes that occur during this life stage [112]. Taken together, results indicate the potential for a strong social desirability effect that may have contributed to some of the nonreporting of height and weight in this sample. It might be expected that social desirability would lead participants to simply alter the values they report, however the few studies that have examined this have found that degree of misreporting is modest [55, 123]. As such, considering this previous research alongside the present study, it appears that social desirability for height and weight doesn't lead to drastic misreporting as much as it may contribute to nonreporting. Interestingly, because many youth reported perceiving themselves as overweight, but neglected to report their actual weight, it is possible that this indicates a greater social desirability influence over the numerical value of ones weight compared to weight perception. Of course, lack of knowledge of ones height and/or weight likely also influenced nonreporting in this sample, particularly for younger youth.

From a research standpoint, it is not unreasonable to expect that comparable self-report survey-based studies among youth may find similar issues with missing height, weight, and BMI as observed in the COMPASS study. If this is the case, missing data need to be appropriately handled in any subsequent analysis which uses these variables, as such a high degree of non-random missing data could certainly bias research findings. For example, if the present sample was to conduct a simple analysis examining average BMI using the most common missing data approach (CCA), average BMI would be biased downwards. If these results were then to be used for population surveillance, a CCA approach would underestimate the prevalence of youth with OWOB. This study has not only contributed to better understanding of the degree and patterns of missing youth BMI, height, and weight data, but has also illustrated the importance of thorough examination of missing data in youth OWOB research. Findings from this study may also extend to other domains; many other self-report metrics are inherently tied to social desirability, including mental health, substance use, finances, etc.

7.2.2 Study 2

Through several significant associations and differential results between males and females, Study 1 hinted that the mechanisms of missing BMI, height, and weight data were complex and warranted further exploration. Results of standard regression approaches can be difficult to synthesize and interpret where there are a large number of significant variables without a hierarchy of importance. As such, in order to further explore the missingness in BMI, height, and weight, Study 2 leveraged a Classification and Regression Trees (CART) model approach, first identified for use to explore missing data by Tierney et al [53].

CART models identified weight perception as a key predictor of BMI missingness for females and males, whereby those who perceived themselves as overweight distinctly belonged to a different group more likely to be missing BMI compared to their counterparts who perceived themselves as underweight or 'about right'. These results were consistent with the regression performed in Study 1, which showed significant association between perceptions of overweight and BMI missingness. Two previous studies have found that poorer satisfaction with one's body was associated with greater likelihood of missing BMI [106,107]. Although this is a different indicator compared to weight perception, given what is known about body image during adolescence these findings from Studies 1 and 2 are congruent with the previous research. Also consistent with Study 1 as well as previous literature [105, 106, 108], younger groups were more likely to be missing BMI. Unlike the regression approach from Study 1, the CART models specifically identified which ages were most likely to split those who reported BMI versus those who didn't. For example among females, 15 years was the age that most differentiated those who reported their BMI from those who did not, whereas for males this age was 16 years. Several other variables were also consistent with Study 1, including movement-related and mental health-related variables. However, unlike Study 1, CART models identified a clear hierarchy of importance among the variables related to missing BMI, height, and weight, that could be very useful for constructing a MI model.

A major finding from Study 2 was the identification of many important subgroups related to missingness. For example, the 9% of the male sample who didn't play sports and perceived themselves to be overweight were more likely to have missing BMI. Subgroups with the highest likelihood of missingness can also be identified; among females, BMI was highly likely to be missing for those who perceived themselves as overweight, didn't participate in strength training, did not pass their last English class, had a wellbeing score less than or equal to 26, and had self rated mental health scores less than 2.5. Examining these subgroups informs us that it is not simply these variables in isolation contributing to nonreporting, but that it is a combination of: weight perception, low physical activity, poor academics, and poor mental health. Broadly speaking, CART models highlight that youth who are struggling with their physical, mental, and emotional health are nearly guaranteed to be missing BMI. This is a major concern for research which then uses youth

BMI in any capacity, since this clearly indicates a systematic pattern of missingness that will bias research results towards the youth who are physically, mentally, and emotionally healthier.

7.2.3 Study 3

Studies 1 and 2 thoroughly examined the missing BMI, height, and weight data among youth in the 2018/19 COMPASS study sample, and found that the missingness was severe; observed levels of missing data were high and likely not missing at random. CCA is the most common approach to handle missing data, but these findings indicated that CCA would be an inappropriate choice. As such, the final step was to explore the impact that missing data methodology choice had on research findings surrounding youth BMI. Specifically, Study 3 compared the results between CCA and MI approaches for an analytical model examining a variety of factors associated with youth BMI. Based on the results of Studies 1 and 2, deleting cases would bias results while a properly constructed MI model could produce unbiased results. As such, differences in results between CCA and MI were assumed to illustrate bias resulting from deleting missing cases. Study 2 directly informed the auxiliary variable selection procedure required for creating the MI model in Study 3.

Across the sex-stratified models examining the association between youth BMI and a variety of diet, movement, mental health, and substance use behaviours, MI and CCA approaches produced very different results. In descriptive statistics, CCA biased average BMI for both females and males downwards, which is consistent with what was expected from the results of Studies 1 and 2, as well as previous research [104,106]. In the regression models, several contrasting results between MI and CCA were identified; for females, MI indicated that cannabis use was significant (consistent with a previous study [89]), whereas CCA found no such association. For males, MI identified depression and e-cigarette use as not significant (inconsistent with a previous study [183]), whereas CCA found significant associations. Other differences were also observed; the direction of association for sports participation was reversed between MI and CCA male models, and sex differences with respect to associations between anxiety and BMI were opposite between CCA and MI models.

If the results of the CCA models in this study were presented without the context of the missingness levels and patterns from Studies 1 and 2, the findings might not appear unusual, since associations in the literature can often be conflicting. There are countless reasons why it can be difficult to compare results between studies, such as being conducted in different populations, using different measures (e.g. objective vs. subjective, self-reported vs. parent reported, different phrasing of questions), or different methods. Study 3 has highlighted that different amounts and handling of missing data can greatly impact research findings, and this may in fact be a contributor to conflicting findings in the literature.

Study 3 is not a guide for conducting analyses related to youth BMI nor does it claim to represent true associations, as this study itself certainly has its own limitations. However, Study 3 does illustrate the importance of examining, reporting, and appropriate handling of missing data, particularly in similar contexts where data suffers from severe missingness.

Study 3 has also demonstrated how comparing methodological practices using real-world data may be beneficial to reach applied scientists working within that domain.

7.3 Implications

7.3.1 Overall

This dissertation has illustrated patterns and mechanisms of missing youth BMI, height, and weight data, as well as the impact that missing data methodology can have on research findings and conclusions in this domain. The COMPASS study sample used in this dissertation is unique in its breadth of variables examined and large sample size, but it is not unlike other cohort studies in the use of a survey-based approach to data collection. Moreover, some of the questions in the COMPASS study survey are validated for use in youth populations and may be identical to those used elsewhere. As such, while the potential social desirability bias in nonreporting observed in this study cannot be naively assumed to be true for studies in other study contexts, it is not unreasonable to suggest that similar survey-based research may face some of the same challenges with nonreporting. At minimum, the missing data patterns related to height, weight, and BMI observed in this research certainly imply that robust examination of missing data in this domain is necessary. In particular, research that relies on self-report measures to examine youth OWOB needs to be transparent about missing data and use a robust approach that is informed by thorough examination of the missingness.

7.3.2 Interpreting the Current Literature

This dissertation also provides context with which to interpret current literature. The findings from this dissertation make clear that differences in missing data and their handling between studies can make comparing results difficult, if not almost inappropriate. Comparisons between existing studies in the youth OWOB domain should look closely for differences in missing data handling, and researchers should be very skeptical of the results of studies that fail to report on missing data at all. This may be especially true for studies that use data collection procedures that will almost inevitably be impacted by nonreporting in some way (e.g., survey-based data collection, longitudinal designs [due to potential for attrition]).

7.3.3 Statistical Training

The results of the research presented in this dissertation certainly lead to recommendations of better examination and handling of missing data in studies which use self-report measures for youth height and weight as well as applied research that uses survey-based approaches more broadly. However, this dissertation is not the first to advocate for improved missing data handling in the literature; it is clear that there is a gap between what is promoted as "best practice", compared to typical practices in applied research. A major limitation to these recommendations is that they are unlikely to be simple or quick to implement. Some recommendations are arguably quite simple, such as reporting percentages of missing data or performing basic examinations. However, the tasks of understanding mechanisms of missing data or implementing MI are not simplistic and require a great deal of training. There is certainly an argument to be made that missing data be a fundamental part of statistical training for applied researchers, particularly in public health. Idealistic scenarios with no missing data ill-prepare students for the challenges of properly working with real data.

7.3.4 Future Directions

This dissertation has highlighted the importance of missing data considerations in future research, both in the youth OWOB domain and survey-based research more broadly. This robust examination and handling of missing data is still far from being the status-quo, however in many ways applied research is progressing. There are some journals that require key information about missing data be included in submissions; for example, the instructions for authors for Epidemiology mentions that authors should report response rates and that quantitative analyses of missing data is recommended [186]. Considering that journals are key academic gatekeepers, this is an important step and it would be beneficial if more journals required robust missing data reporting.

While researchers can examine recent publications to establish an idea of current approaches to missing data reporting and handling in their respective fields, there is substantial value in systematic reviews that explore this topic. A valuable area of future study would be to take inventory of current missing data handling by updating previously published systematic reviews that have examined this in applied research in general [35, 187]. In additional to general reviews, domain-specific reviews of missing data handling may also be valuable. For example, a 2015 review examined missing data handing for research examining predictors of type II diabetes. The authors found the majority of studies did not report missing data at all and MI methods were seldom used [188]. Given the domain-specific focus, this review likely has better opportunity to reach researchers focusing on risk factors for type II diabetes compared to a general review.

This dissertation focused on missing data in the youth OWOB domain using real-world data and aimed to publish in journals that would reach applied scientists. This type of approach can be applied to other fields, and help contextualize the reasons for missingness, as well as how statistical bias from missing data can infiltrate research findings and subsequent policy and programming recommendations. Lastly, an ideal scenario would be to avoid missing data before it occurs. This dissertation focuses on post-processing of missing data, but future research could further dive into reasons for missingness with the aim of preventing it from occurring. For example, the youth OWOB domain would benefit from qualitative research examining how youth perceive these survey questions, and the reasons they might not be reporting these values.

7.4 Strengths and Limitations

7.4.1 Sample

There are a number of strengths and limitations of this dissertation related to the sample used. Data from the COMPASS study used in this research did not follow a representative sampling frame. As such, research conclusions cannot be generalized to all Canadian youth. However, other aspects of the COMPASS study greatly increase utility of the findings and applicability to other research, potentially beyond what would be possible in a representative sample. For example, the large sample size allowed for sufficient power to conduct complex multi-level analyses and reduced concerns surrounding small cell counts, singularity, and convergence. The passive consent protocol used by COMPASS is also a large strength of this research, as this procedure helps limit self-selection bias as well as limit unit non-response [6, 7]. Limiting unit non-response allowed this dissertation to focus on item non-response.

All measures in this research, including BMI, were self-reported. Of course, the methodological lens of this research was built upon this feature. However, the self-report nature of these variables limits the breadth of research conclusions, as well as comparisons to other research (particularly that which uses objective measurements). Moreover, there are a few key variables absent from this research that would be beneficial for future studies to consider. First, because COMPASS is survey-based and focused on health behaviours, biological data are not collected. Objectively measured biological indicators (e.g. blood pressure, cholesterol, vascular changes, etc.) are robust predictors of future chronic disease among youth [189]. Since this dissertation focused on self-report measures and no objective biological data were examined, no substantial conclusions surrounding diagnostics or health status of this sample could be made. In terms of missing behavioural predictors, existing literature had suggested that parental variables may be related to youth BMI [190–193], as well as youth likelihood to report BMI [106]. COMPASS did not collect parental information, but parental influence would be a beneficial area for future research to explore.

Although this research is missing some measures, the breadth of factors that were included extends beyond the majority of existing research on the topic of missing youth BMI, height, and weight data. No previous studies which focused on examining missing BMI data among youth had included mental health variables, which findings from this research suggested were important predictors of missingness. Also missing from previous studies was a measure of weight perception, which was highly beneficial to include in this research as it was the only variable that could somewhat proxy weight data where missing.

Lastly, the use of a cross-sectional sample in this research has associated strengths and limitations. With a cross-sectional sample, temporal trends cannot be established, and as such no conclusions could be made about order of occurrence for behaviours. An examination of temporal trends may have been particularly interesting for Study 3, as more concrete conclusions could be made about factors associated with youth BMI, and changes in BMI could have been examined. However, there was insufficient foundation in the literature to warrant that this dissertation focus on longitudinal examination of missing data and methods. Instead, the following benefits of a cross-sectional approach were determined to outweigh the drawbacks. First, a cross-sectional approach establishes a foundation for future longitudinal work and may be simpler for other researchers to interpret; this was considered important given that researchers were a target audience for this work. Next, this sample is hierarchical in nature (students clustered within schools), and a longitudinal approach would add an additional level in the model (i.e., students clustered within schools, over time) that would not have been feasible for some of the methods used in this research. Namely, three-level CART models are not an established method at present time, nor is three-level MI. Although the cross-sectional nature of this research may limit its use for researchers working with longitudinal data, a similar approach could be adopted for longitudinal analyses in which spatial clustering (i.e., by school) is switched to temporal clustering (i.e., over time).

7.4.2 Chosen methods

Some of the statistical techniques used in this dissertation have their own associated strengths and limitations. Subsections below outline strengths and limitations of CART and MI.

CART models

The CART approach used in Study 2 has several strengths. First, CART approaches address the issue of systematic variable selection, as by nature they recursively split the data by variables identified as most predictive of a split. In contrast to Study 1, no specialized variable selection framework was required to handle the large list of potential predictors. The process of pruning to reduce the number of factors in the CART models was systematic and simple to implement, as the cross-validation approach used is typical in CART modelling [145]. Another strength of CART models is that they are visual in nature, which can facilitate the presentation and interpretation of results. Lastly, a major strength of CART models is that they produce a hierarchy of variables, which can be used to identify variable importance or relevance. In a regression approach, in order to directly compare estimates to try and establish variable importance, variables would need to exist on the same scale or be re-scaled, which may decrease their interpretability. In contrast, CART models are hierarchical by nature so no such rescaling is required.

CART models also have some associated limitations to note. Namely, CART models tend to prefer split criteria that have more categories (a feature known as “greedy splitting”). In other words, variables may be identified as “more important” simply based on the fact that there are more categorical options. As such, care was taken in re-defining variables in Section 3.1.5 to limit the number of categories, and where possible variables were transformed to binary. The loss of information that results from grouping categories was a necessary trade-off for improved interpretability and better cell counts across all studies, as well as avoiding greedy splitting in Study 2. Another limitation to note is that

CART models are considered inferior to their ensemble-based cousins, bagging and random forests [194]. While ensemble methods produce a set of trees as opposed to a single tree and thus have better predictive performance, they are substantially more difficult to interpret. In the context of this research, ease of interpretation was paramount given that the results from Study 2 were to aid in the understanding of missing data as well as directly inform inputs for Study 3. As such, a CART approach was more appropriate in this case.

Multiple Imputation

MI has one key strength that rationalizes its use, which is the potential to produce efficient and unbiased estimates where it is properly implemented [13, 15]. However there are some limitations surrounding MI that need to be mentioned. One key limitation of applying MI is that researchers are limited in their ability to check the fit of an imputation model in comparison to the available tools to check analysis model fit. In Study 3 there was an underlying assumption that discrepancies between the CCA and MI model stemmed from bias introduced by deleting cases; if the MI model was properly constructed, this assumption is valid. However, as with nearly any statistical approach, model misspecification is always possible. Moreover, there is no absolute way to check the validity of a MI model. Several steps throughout Study 3 were taken to address this limitation. First, auxiliary variable selection was a robust process, as the results from Study 2 directly informed the auxiliary variables for Study 3. This process was already beyond the typical auxiliary variable selection procedure, which suggests the use of correlation coefficients for selection [195]. Also, several model checking procedures were performed in order to increase confidence in the MI model, including monitoring convergence and comparing the distribution of observed and imputed variables, as outlined in van Buuren [8]

7.4.3 Novelty

A major strength of all three studies is novelty. Prior to Study 1, it appeared that only 4 studies had focused on examining missingness in BMI or weight among youth. Study 1 added to this limited existing literature and included many variables not previously examined, such as the mental health and substance use related variables. Moreover, none of the previous studies had stratified analyses by sex in order to examine how reporting patterns differed between females and males, and none had examined height missingness directly. Study 1 filled these gaps in the literature. Study 2 was novel in the use of a CART model to examine patterns of missigness; this approach was previously identified by Tierney et al [53], however it did not appear to have been leveraged in the literature until Study 2. Lastly, Study 3 addressed the topic of bias resulting from missing data methodology; a limited number of previous studies have also compared missing data methodologies using real-world data [18, 41, 45], but it appears that Study 3 was the first to explore this topic in the context of a model examining factors associated with youth BMI.

7.5 Conclusions

Findings from all three studies included in this dissertation demonstrate the importance of missing data examination and handling in research focusing on youth BMI or OWOB. This research found that in a large cohort study of Canadian youth, there was substantial non-random missing data in height, weight, and BMI. The factors identified as significantly related to missing youth BMI implied that social desirability was likely contributing in part to nonreporting of height and/or weight, and that females more frequently neglected to report their weight, while males more frequently neglected to report their height. Overall, based on observed associations with missingness in this sample it appeared that those who were missing BMI were more likley to have a higher BMI. Missing BMI was highly likely for those who: perceived themselves as overweight, were younger, were less physically active, and had poorer mental health. Overall, removing cases for those missing BMI biased the sample towards youth who were physically, mentally, and emotionally healthier. In analyses examining youth BMI, comparing multiple imputation and complete cases analysis approaches demonstrated that missing data methodology greatly impacted research findings. Bias introduced by deleting missing cases led to the two approaches yielding contrasting conclusions with respect to the factors associated with BMI and sex-differences in the analytical models. All together, findings suggest that missing data must be a key consideration for survey-based research among youth, and that research with poor missing data reporting or handling be interpreted cautiously.

References

[1] Lin JY, Lu Y, Tu X. How to avoid missing data and the problems they pose: Design considerations. Shanghai Archives of Psychiatry. 2012 6;24(3):181–185. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25324625.

[2] Kelly N, Harpel T, Fontes A, Walters C, Murphy J. An Examination of Social Desirability Bias in Measures of College Students' Financial Behavior. College Student Journal. 2017 3;51(1):115–128. Available from: https://www.ingentaconnect.com/content/prin/csj/2017/00000051/00000001/art00013.

[3] Latkin CA, Edwards C, Davey-Rothwell MA, Tobin KE. The relationship between social desirability bias and self-reports of health, substance use, and social network factors among urban substance users in Baltimore, Maryland. Addictive Behaviors. 2017;73:133–136. Available from: http://dx.doi.org/10.1016/j.addbeh.2017.05.005.

[4] Maclennan B, Kypri K, Langley J, Room R. Non-response bias in a community survey of drinking, alcohol-related experiences and public opinion on alcohol policy. Drug and Alcohol Dependence. 2012 11;126(1-2):189–194. Available from: https://www.sciencedirect.com/science/article/pii/S0376871612001779.

[5] Haziza D, Lesage A discussion of weighting procedures for unit nonresponse. Journal of Official Statistics. 2016;32(1):129–145. Available from: http://dx.doi.org/10.1515/JOS-2016-0006.

[6] White VM, Hill DJ, Effendi Y. How does active parental consent influence the findings of drug-use surveys in schools? Evaluation Review. 2004 6;28(3):246–260. Available from: https://www.doi.org/10.1177/0193841X03259549.

[7] Rojas NL, Sherrit L, Harris S, Knight JR. The Role of Parental Consent in Adolescent Substance Use Research. Journal of Adolescent Health. 2008 2;42(2):192–197. Available from: https://doi.org/10.1016/j.jadohealth.2007.07.011.

[8] van Buuren S. Flexible Imputation of Missing Data. CRC Press; 2018. Available from: https://stefvanbuuren.name/fimd/.

[9] Dong Y, Peng CYJ. Principled missing data methods for researchers. SpringerPlus. 2013 12;2(1):1–17. Available from: http://www.springerplus.com/content/2/1/222.

[10] Gad AM, Ahmed AS. Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm. Computational Statistics and Data Analysis. 2006 6;50(10):2702–2714. Available from: https://www.sciencedirect.com/science/article/pii/S0167947305000678.

[11] Horton NJ, Kleinman KP. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician. 2007 2;61(1):79–90. Available from: http://www.tandfonline.com/doi/abs/10.1198/000313007X172556.

[12] Rubin DB. Inference and missing data. Biometrika. 1976 12;63(3):581–592. Available from: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/63.3.581.

[13] Allison PD. Missing data. vol. 11. SAGE Publications; 2002.

[14] Little RJA. A Test of Missing Completely at Random for Multivariate Data with Missing Values. Journal of the American Statistical Association. 1988 12;83(404):1198. Available from: https://www.jstor.org/stable/2290157?origin=crossref.

[15] Enders CK. Applied missing data analysis. Guilford Press; 2010.

[16] Allison P. Missing Data Using SAS; 2019.

[17] Schafer JL, Graham JW. Missing data: Our view of the state of the art. Psychological Methods. 2002;7(2):147–177. Available from: https://psycnet.apa.org/doi/10.1037/1082-989X.7.2.147.

[18] Hallgren KA, Witkiewitz K. Missing Data in Alcohol Clinical Trials: A Comparison of Methods. Alcoholism: Clinical and Experimental Research. 2013;37(12):2152–2160. Available from: https://journals.scholarsportal.info/pdf/01456008/v37i0012/2152_mdiactacom.xml.

[19] Harris AHS, Boden MT, Finlay AK, Rubinsky AD. The Challenges of Improving Statistical Practice in Alcohol Treatment Research. Alcoholism: Clinical and Experimental Research. 2013;37(12):1999–2001. Available from: https://doi.org/10.1111/acer.12316.

[20] Streiner DL. Missing data and the trouble with LOCF. Evidence-Based Mental Health. 2008 2;11(1):3–5. Available from: http://ebmh.bmj.com/cgi/doi/10.1136/ebmh.11.1.3-a.

[21] Gadbury GL, Coffey CS, Allison DB. Modern statistical methods for handling missing repeated measurements in obesity trial data: Beyond LOCF. Obesity Reviews. 2003;4(3):175–184. Available from: https://www.doi.org/10.1046/j.1467-789X.2003.00109.x.

[22] Yung YF, Zhang W. Making Use of Incomplete Observations in the Analysis of Structural Equation Models : The CALIS Procedure's Full Information Maximum Likelihood Method in SAS/STAT; 2011. Available from: https://support.sas.com/resources/papers/proceedings11/333-2011.pdf.

[23] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society: Series B (Methodological). 1977 9;39(1):1–22. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x.

[24] Rubin DB. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.; 1987. Available from: http://doi.wiley.com/10.1002/9780470316696.

[25] Allison P. Why You Probably Need More Imputations Than You Think — Statistical Horizons; 2012. Available from: http://statisticalhorizons.com/more-imputations.

[26] Demirtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. Journal of Statistical Computation and Simulation. 2008 2;78(1):69–84. Available from: http://www.tandfonline.com/doi/abs/10.1080/10629360600903866.

[27] von Hippel PT. Should a Normal Imputation Model be Modified to Impute Skewed Variables? Sociological Methods and Research. 2013 2;42(1):105–138. Available from: http://journals.sagepub.com/doi/10.1177/0049124112464866.

[28] van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006 12;76(12):1049–1064. Available from: http://www.tandfonline.com/doi/abs/10.1080/10629360600810434.

[29] De Silva AP, Moreno-Betancur M, De Livera AM, Lee KJ, Simpson JA. A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: A simulation study. BMC Medical Research Methodology. 2017;17(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5526258/pdf/12874_2017_Article_372.pdf.

[30] Karangwa I. Using the Markov Chain Monte Carlo Method to Make Inferences on Items of Data Contaminated by Missing Values,. American Journal of Theoretical and Applied Statistics. 2013;2(3):48. Available from: http://www.sciencepublishinggroup.com/journal/paperinfo.aspx?journalid=146&doi=10.11648/j.ajtas.20130203.12.

[31] Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. American Journal of Epidemiology. 2010 3;171(5):624–632. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwp425.

[32] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods. 2001 12;6(4):330–351. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11778676.

[33] Yoo JE. The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. Educational and Psychological Measurement. 2009 12;69(6):929–947. Available from: http://journals.sagepub.com/doi/10.1177/0013164409332225.

[34] Schafer JL, Olsen MK. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behavioral Research. 1998 10;33(4):545–571. Available from: http://www.tandfonline.com/doi/abs/10.1207/s15327906mbr3304_5.

[35] Eekhout I, De Boer RM, Twisk JWR, De Vet HCW, Heymans MW. Missing data: A systematic review of how they are reported and handled. Epidemiology. 2012 9;23(5):729–732. Available from: https://www.doi.org/10.1097/EDE.0b013e3182576cdb.

[36] Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Medical Research Methodology. 2012 12;12(1):96. Available from: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-12-96.

[37] Yucel RM. State of the multiple imputation software. Journal of Statistical Software. 2011 12;45(1):1–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22289957.

[38] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ (Online). 2009;339(7713):157–160. Available from: https://www.doi.org/10.1136/bmj.b2393.

[39] White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010 12;29(28):2920–2931. Available from: http://doi.wiley.com/10.1002/sim.3944.

[40] Young R, Johnson D. Methods for Handling Missing Secondary Respondent Data. Journal of Marriage and Family. 2013;75(1):221–234. Available from: https://www.doi.org/10.1111/j.1741-3737.2012.01021.x.

[41] Dohoo IR, Nielsen CR, Emanuelson U. Multiple imputation in veterinary epidemiological studies: A case study and simulation. Preventive Veterinary Medicine. 2016 7;129:35–47. Available from: https://www.sciencedirect.com/science/article/pii/S0167587716301088.

[42] Olsen IC, Kvien TK, Uhlig T. Consequences of handling missing data for treatment response in osteoarthritis: A simulation study. Osteoarthritis and Cartilage. 2012;20(8):822–828. Available from: http://dx.doi.org/10.1016/j.joca.2012.03.005.

[43] Zhu X. Comparison of Four Methods for Handing Missing Data in Longitudinal Data Analysis through a Simulation Study. Open Journal of Statistics. 2014;04(11):933–944. Available from: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/ojs.2014.411088.

[44] Knol MJ, Janssen KJM, Donders ART, Egberts ACG, Heerdink ER, Grobbee DE, et al. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. Journal of Clinical Epidemiology. 2010 7;63(7):728–736. Available from: https://www.sciencedirect.com/science/article/pii/S0895435610000181.

[45] Coquet JB, Tumas N, Osella AR, Tanzi M, Franco I, Diaz MDP. Breast cancer and modifiable lifestyle factors in argentinean women: Addressing missing data in a case-control study. Asian Pacific Journal of Cancer Prevention. 2016;17(10):4567–4575. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27892664.

[46] Ertel KA, Kleinman K, van Rossem L, Sagiv S, Tiemeier H, Hofman A, et al. Maternal perinatal depression is not independently associated with child body mass index in the Generation R Study: methods and missing data matter. Journal of Clinical Epidemiology. 2012 12;65(12):1300–1309. Available from: https://www.sciencedirect.com/science/article/pii/S0895435612001953#bib28.

[47] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification And Regression Trees. vol. 14. Routledge; 2017. Available from: https://www.taylorfrancis.com/books/9781351460491.

[48] Morgan J. Classification and Regression Tree Analysis. Boston University; 2014. Available from: https://www.bu.edu/sph/files/2014/05/MorganCART.pdf.

[49] van Buuren S. mice.impute.cart: Imputation By Classification And Regression Trees;. Available from: https://www.rdocumentation.org/packages/mice/versions/3.8.0/topics/mice.impute.cart.

[50] Loh WY, Eltinge J, Cho M, Li Y. Classification and regression tree methods for incomplete data from sample surveys; 2016. Available from: http://arxiv.org/abs/1603.01631.

[51] Hayes T, McArdle JJ. Should we impute or should we weight? Examining the performance of two CART-based techniques for addressing missing data in small sample research with nonnormal variables. Computational Statistics and Data Analysis. 2017;115:35–52. Available from: http://dx.doi.org/10.1016/j.csda.2017.05.006.

[52] Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology. 2014 3;179(6):764–774. Available from: https://doi.org/10.1093/aje/kwt312.

[53] Tierney NJ, Harden FA, Harden MJ, Mengersen KL. Using decision trees to understand structure in missing data. BMJ Open. 2015 6;5(6):e007450. Available from: http://dx.doi.org/10.1136/bmjopen-2014-007450.

[54] Himes JH, Faricy A. Validity and reliability of self-reported stature and weight of US adolescents. American Journal of Human Biology. 2001 2;13(2):255–260. Available from: http://doi.wiley.com/10.1002/1520-6300%28200102/03%2913%3A2%3C255%3A%3AAID-AJHB1036%3E3.0.CO%3B2-E.

[55] Sherry B, Jefferds ME, Grummer-Strawn LM. Accuracy of adolescent self-report of height and weight in assessing overweight status: A literature review. Archives of Pediatrics and Adolescent Medicine. 2007 12;161(12):1154–1161. Available from: http://archpedi.jamanetwork.com/article.aspx?doi=10.1001/archpedi.161.12.1154.

[56] Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: A systematic analysis for the Global Burden of Disease Study 2013. The Lancet. 2014 8;384(9945):766–781. Available from: https://www.doi.org/10.1016/S0140-6736(14)60460-8.

[57] Government of Canada. Tackling obesity in Canada: Childhood obesity and excess weight rates in Canada; 2018. Available from: https://publications.gc.ca/site/eng/9.848986/publication.html.

[58] Public Health Agency of Canada. Designing Healthy Living. Public Health Agency of Canada; 2017. Available from: https://www.canada.ca/en/public-health/services/publications/chief-public-health-officer-reports-state-public-health-canada/2017-designing-healthy-living.html.

[59] Guh DP, Zhang W, Bansback N, Amarsi Z, Birmingham CL, Anis AH. The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis. BMC Public Health. 2009 12;9(1):88. Available from: https://www.doi.org/10.1186/1471-2458-9-88.

[60] Harriger JA, Thompson JK. Psychological consequences of obesity: Weight bias and body image in overweight and obese youth. International Review of Psychiatry. 2012 6;24(3):247–253. Available from: https://www.doi.org/10.3109/09540261.2012.678817.

[61] Anis AH, Zhang W, Bansback N, Guh DP, Amarsi Z, Birmingham CL. Obesity and overweight in Canada: An updated cost-of-illness study. Obesity Reviews. 2010 1;11(1):31–40. Available from: https://www.doi.org/10.1111/j.1467-789X.2009.00579.x.

[62] Twells LK, Gregory DM, Reddigan J, Midodzi WK. Current and predicted prevalence of obesity in Canada: a trend analysis. CMAJ Open. 2014 3;2(1):E18–E26. Available from: https://www.doi.org/10.9778/cmajo.20130016.

[63] Singh AS, Mulder C, Twisk JWR, Van Mechelen W, Chinapaw MJM. Tracking of childhood overweight into adulthood: A systematic review of the literature. Obesity Reviews. 2008;9(5):474–488. Available from: https://www.doi.org/10.1111/j.1467-789X.2008.00475.x.

[64] Serdula MK, Ivery D, Coates RJ, Freedman DS, Williamson DF, Byers T. Do Obese Children Become Obese Adults? A Review of the Literature. Preventive Medicine. 1993 3;22(2):167–177. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0091743583710145.

[65] Luttikhuis HO, Baur L, Jansen H, Shrewsbury VA, O'Malley C, Stolk RP, et al. Interventions for treating obesity in children. In: Oude Luttikhuis H, editor. Cochrane Database of Systematic Reviews. 1. Chichester, UK: John Wiley & Sons, Ltd; 2009. Available from: http://doi.wiley.com/10.1002/14651858.CD001872.pub2.

[66] Amaro H. The action is upstream: place-based approaches for achieving population health and health equity. American journal of public health. 2014 6;104(6):964. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4062038/.

[67] Sacks G, Swinburn B, Lawrence M. Obesity Policy Action framework and analysis grids for a comprehensive policy approach to reducing obesity. Obesity Reviews. 2009 1;10(1):76–86. Available from: http://doi.wiley.com/10.1111/j.1467-789X.2008.00524.x.

[68] Allender S, Owen B, Kuhlberg J, Lowe J, Nagorcka-Smith P, Whelan J, et al. A community based systems diagram of obesity causes. PLoS ONE. 2015;10(7):1–12. Available from: https://www.doi.org/10.1371/journal.pone.0129683.

[69] Jebb SA, Kopelman P, Butland B. Executive Summary: FORESIGHT Tackling Obesities: Future Choices project. Obesity Reviews. 2007 3;8(s1):vi–ix. Available from: http://doi.wiley.com/10.1111/j.1467-789X.2007.00344.x.

[70] Al-Hazzaa HM, Abahussain NA, Al-Sobayel HI, Qahwaji DM, Musaiger AO. Lifestyle factors associated with overweight and obesity among Saudi adolescents. BMC Public Health. 2012 12;12(1):354. Available from: http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-12-354.

[71] You J, Choo J. Adolescent overweight and obesity: Links to socioeconomic status and fruit and vegetable intakes. International Journal of Environmental Research and Public Health. 2016;13(3). Available from: https://www.doi.org/10.3390/ijerph13030307.

[72] Zakrzewski JK, Gillison FB, Cumming S, Church TS, Katzmarzyk PT, Broyles ST, et al. Associations between breakfast frequency and adiposity indicators in children from 12 countries. International Journal of Obesity Supplements. 2015;5(S2):S80–S88. Available from: https://www.doi.org/10.1038/ijosup.2015.24.

[73] Veltsista A, Laitinen J, Sovio U, Roma E, Järvelin MR, Bakoula C. Relationship between Eating Behavior, Breakfast Consumption, and Obesity Among Finnish and Greek Adolescents. Journal of Nutrition Education and Behavior. 2010;42(6):417–421. Available from: https://www.doi.org/10.1016/j.jneb.2009.12.004.

[74] Zheng M, Rangan A, Olsen NJ, Bo Andersen L, Wedderkopp N, Kristensen P, et al. Sugar-sweetened beverages consumption in relation to changes in body fatness over 6 and 12 years among 9-year-old children: The European Youth Heart Study. European

Journal of Clinical Nutrition. 2014 1;68(1):77–83. Available from: http://www.nature.com/articles/ejcn2013243.

[75] Berkey CS, Rockett HRH, Field AE, Gillman MW, Colditz GA. Sugar-added beverages and adolescent weight change. Obesity Research. 2004;12(5):778–788. Available from: https://www.doi.org/10.1038/oby.2004.94.

[76] Malik VS, Schulze MB, Hu FB. Intake of sugar-sweetened beverages and weight gain: A systematic review. American Journal of Clinical Nutrition. 2006;84(2):274–288. Available from: https://www.doi.org/10.1093/ajcn/84.2.274.

[77] Thompson OM, Ballew C, Resnicow K, Must A, Bandini LG, Cyr H, et al. Food purchased away from home as a predictor of change in BMI z-score among girls. International Journal of Obesity. 2004;28(2):282–289. Available from: https://www.doi.org/10.1038/sj.ijo.0802538.

[78] Taveras EM, Berkey CS, Rifas-Shiman SL, Ludwig DS, Rockett HRH, Field AE, et al. Association of consumption of fried food away from home with body mass index and diet quality in older children and adolescents. Pediatrics. 2005;116(4). Available from: https://www.doi.org/10.1542/peds.2004-2732.

[79] Nardocci M, Leclerc BS, Louzada ML, Monteiro CA, Batal M, Moubarac JC. Consumption of ultra-processed foods and obesity in Canada. Canadian Journal of Public Health. 2019;110(1):4–14. Available from: https://www.doi.org/10.17269/s41997-018-0130-x.

[80] Reedy J, Krebs-Smith SM. Dietary Sources of Energy, Solid Fats, and Added Sugars among Children and Adolescents in the United States. Journal of the American Dietetic Association. 2010;110(10):1477–1484. Available from: http://dx.doi.org/10.1016/j.jada.2010.07.010.

[81] Langlois K, Garriguet D. Sugar consumption among Canadians of all ages. Health Reports. 2011;22(3). Available from: https://pubmed.ncbi.nlm.nih.gov/22106786/.

[82] Peart T, Mondragon HEV, Rohm-Young D, Bronner Y, Hossain MB. Weight status in US youth: The role of activity, diet, and sedentary behaviors. American Journal of Health Behavior. 2011;35(6):756–765. Available from: https://www.doi.org/10.5993/AJHB.35.6.11.

[83] Wethington H, Pan L, Sherry B. The association of screen time, television in the bedroom, and obesity among school-aged youth: 2007 national survey of children's health. Journal of School Health. 2013;83(8):573–581. Available from: https://www.doi.org/10.1111/josh.12067.

[84] Liou YM, Liou TH, Chang LC. Obesity among adolescents: sedentary leisure time and sleeping as determinants. Journal of Advanced Nursing. 2010 4;66(6):1246–1256. Available from: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2648.2010.05293.x.

[85] Kenney EL, Gortmaker SL. United States Adolescents' Television, Computer, Videogame, Smartphone, and Tablet Use: Associations with Sugary Drinks, Sleep, Physical Activity, and Obesity. Journal of Pediatrics. 2017;182:144–149. Available from: http://dx.doi.org/10.1016/j.jpeds.2016.11.015.

[86] Storfer-Isser A, Patel SR, Babineau DC, Redline S. Relation between sleep duration and BMI varies by age and sex in youth age 8-19. Pediatric Obesity. 2012;7(1):53–64. Available from: https://www.doi.org/10.1111/j.2047-6310.2011.00008.x.

[87] Jarrin DC, McGrath JJ, Drake CL. Beyond sleep duration: Distinct sleep dimensions are associated with obesity in children and adolescents. International Journal of Obesity. 2013;37(4):552–558. Available from: https://www.doi.org/10.1038/ijo.2013.4.

[88] Snell EK, Adam EK, Duncan GJ. Sleep and the body mass index and overweight status of children and adolescents. Child Development. 2007;78(1):309–323. Available from: https://www.doi.org/10.1111/j.1467-8624.2007.00999.x.

[89] Farhat T, Iannotti RJ, Simons-Morton BG. Overweight, Obesity, Youth, and Health-Risk Behaviors. American Journal of Preventive Medicine. 2010;38(3):258–267. Available from: http://dx.doi.org/10.1016/j.amepre.2009.10.038.

[90] Pasch KE, Velazquez CE, Cance JD, Moe SG, Lytle LA. Youth Substance Use and Body Composition: Does Risk in One Area Predict Risk in the Other? Journal of Youth and Adolescence. 2012;41(1):14–26. Available from: https://www.doi.org/10.1007/s10964-011-9706-y.

[91] Huang DYC, Lanza HI, Anglin MD. Association Between Adolescent Substance Use And Obesity In Young Adulthood: A Group-Based Dual Trajectory Analysis. Addictive Behaviors. 2013;38(11):2653–2660. Available from: http://dx.doi.org/10.1016/j.addbeh.2013.06.024.

[92] Battista K, Leatherdale ST. Estimating how extra calories from alcohol consumption are likely an overlooked contributor to youth obesity. Health Promotion and Chronic Disease Prevention in Canada. 2017 6;37(6):194–200. Available from: https://www.doi.org/10.24095/hpcdp.37.6.03.

[93] Fox CK, Gross AC, Rudser KD, Foy AMH, Kelly AS. Depression, Anxiety, and Severity of Obesity in Adolescents: Is Emotional Eating the Link? Clinical Pediatrics. 2016;55(12):1120–1125. Available from: https://www.doi.org/10.1177/0009922815615825.

[94] Oddy WH, Allen KL, Trapp GSA, Ambrosini GL, Black LJ, Huang RC, et al. Dietary patterns, body mass index and inflammation: Pathways to depression and mental health problems in adolescents. Brain, Behavior, and Immunity. 2018;69(2018):428–439. Available from: https://doi.org/10.1016/j.bbi.2018.01.002.

[95] Sharafi SE, Garmaroudi G, Ghafouri M, Bafghi SA, Ghafouri M, Tabesh MR, et al. Prevalence of anxiety and depression in patients with overweight and obesity. Obesity Medicine. 2020;17(December 2019). Available from: https://www.doi.org/10.1016/j.obmed.2019.100169.

[96] Anderson SE, Cohen P, Naumova EN, Jacques PF, Must A. Adolescent obesity and risk for subsequent major depressive disorder and anxiety disorder: Prospective evidence. Psychosomatic Medicine. 2007;69(8):740–747. Available from: https://www.doi.org/10.1097/PSY.0b013e31815580b4.

[97] Janssen I, Craig WM, Boyce WF, Pickett W. Associations between Overweight and Obesity with Bullying Behaviors in School-Aged Children. Pediatrics. 2004;113(5 I):1187–1194. Available from: https://www.doi.org/10.1542/peds.113.5.1187.

[98] Queen's University. Student Mental Health and Wellness Framework and Reccomendations for a Comprehensive Strategy; 2012. Available from: https://campusmentalhealth.ca/wp-content/uploads/2018/03/CMHFinalReport.pdf.

[99] Reed DB, Patterson PJ, Wasserman N. Obesity in Rural Youth: Looking Beyond Nutrition and Physical Activity. Journal of Nutrition Education and Behavior. 2011;43(5):401–408. Available from: http://dx.doi.org/10.1016/j.jneb.2010.12.005.

[100] Morgan Hughey S, Kaczynski AT, Child S, Moore JB, Porter D, Hibbert J. Green and lean: Is neighborhood park and playground availability associated with youth obesity? Variations by gender, socioeconomic status, and race/ethnicity. Preventive Medicine. 2017;95:S101–S108. Available from: http://dx.doi.org/10.1016/j.ypmed.2016.11.024.

[101] Carroll-Scott A, Gilstad-Hayden K, Rosenthal L, Peters SM, McCaslin C, Joyce R, et al. Disentangling neighborhood contextual associations with child body mass index, diet, and physical activity: The role of built, socioeconomic, and social environments. Social Science and Medicine. 2013;95(2013):106–114. Available from: http://dx.doi.org/10.1016/j.socscimed.2013.04.003.

[102] Leatherdale ST, Ahmed R. Screen-based sedentary behaviours among a nationally representative sample of youth: Are Canadian kids couch potatoes?; 2011. 4. Available from: https://pubmed.ncbi.nlm.nih.gov/21978636/.

[103] Larsen K, Cook B, Stone MR, Faulkner GEJ. Food access and children's BMI in Toronto, Ontario: assessing how the food environment relates to overweight and obesity. International Journal of Public Health. 2014;60(1):69–77. Available from: https://www.doi.org/10.1007/s00038-014-0620-4.

[104] Aceves-Martins M, Whitehead R, Inchley J, Giralt M, Currie C, Solà R. Self-reported weight and predictors of missing responses in youth. Nutrition. 2018;53:54–58. Available from: https://doi.org/10.1016/j.nut.2018.01.003.

[105] Arbour-Nicitopoulos KP, Faulkner GE, Leatherdale ST. Learning from non-reported data: Interpreting missing body mass index values in young children. Measurement in Physical Education and Exercise Science. 2010;14(4):241–251. Available from: https://www.doi.org/10.1080/1091367X.2010.520243.

[106] Fonseca H, Gaspar de Matos M, Guerra A, Gomes-Pedro J. Emotional, behavioural and social correlates of missing values for BMI. Archives of Disease in Childhood.

2008 9;94(2):104–109. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18801764.

[107] Tiggemann M. Nonreporting of body mass index: A research note on the interpretation of missing data. International Journal of Eating Disorders. 2006 5;39(4):346–349. Available from: http://doi.wiley.com/10.1002/eat.20264.

[108] Himes JH. Challenges of accurately measuring and using BMI and other indicators of obesity in children. Pediatrics. 2009;124(SUPPL. 1):s3–s22. Available from: http://pediatrics.aappublications.org/lookup/doi/10.1542/peds.2008-3586D.

[109] Brewis AA. Stigma and the perpetuation of obesity. Social Science and Medicine. 2014 10;118(C):152–158. Available from: http://dx.doi.org/10.1016/j.socscimed.2014.08.003.

[110] Rosenblum GD, Lewis M. The relations among body image, physical attractiveness, and body mass in adolescence. Child Development. 1999;70(1):50–64. Available from: https://www.doi.org/10.1111/1467-8624.00005.

[111] Ata RN, Ludden AB, Lally MM. The effects of gender and family, friend, and media influences on eating behaviors and body image during adolescence. Journal of Youth and Adolescence. 2007;36(8):1024–1037. Available from: https://www.doi.org/10.1007/s10964-006-9159-x.

[112] Voelker D, Reel J, Greenleaf C. Weight status and body image perceptions in adolescents: current perspectives. Adolescent Health, Medicine and Therapeutics. 2015 8;p. 149. Available from: https://www.doi.org/10.2147/AHMT.S68344.

[113] Marttila-Tornio K, Ruotsalainen H, Miettunen J, Männikkö N, Kääriäinen M. Clusters of health behaviours and their relation to body mass index among adolescents in Northern Finland. Scandinavian Journal of Caring Sciences. 2019;(12). Available from: https://www.doi.org/10.1111/scs.12769.

[114] Lown EA, Lui CK, Karriker-Jaffe K, Mulia N, Williams E, Ye Y, et al. Adverse childhood events and risk of diabetes onset in the 1979 National longitudinal survey of youth cohort. BMC Public Health. 2019;19(1). Available from: https://www.doi.org/10.1186/s12889-019-7337-5.

[115] Di Giacomo D, Ranieri J, Fiasca F, Mattei A. Lifestyle, body mass index and wellness in youth: Strengthens and weakness in Italian youth. Health Psychology Research. 2019;7(1):21–28. Available from: https://www.doi.org/10.4081/hpr.2019.8035.

[116] Verschueren M, Claes L, Palmeroni N, Bogaerts A, Gandhi A, Moons P, et al. Eating Disorder Symptomatology in Adolescent Boys and Girls: Identifying Distinct Developmental Trajectory Classes. Journal of Youth and Adolescence. 2020;49(2):410–426. Available from: http://dx.doi.org/10.1007/s10964-019-01174-0.

[117] Okorodudu DO, Jumean MF, Montori VM, Romero-Corral A, Somers VK, Erwin PJ, et al. Diagnostic performance of body mass index to identify obesity as defined by body adiposity: A systematic review and meta-analysis. International Journal of

Obesity. 2010;34(5):791–799. Available from: http://dx.doi.org/10.1038/ijo.2010.5.

[118] Javed A, Jumean M, Murad MH, Okorodudu D, Kumar S, Somers VK, et al. Diagnostic performance of body mass index to identify obesity as defined by body adiposity in children and adolescents: A systematic review and meta-analysis. Pediatric Obesity. 2015;10(3):234–244. Available from: https://www.doi.org/10.1111/ijpo.242.

[119] Simmonds M, Burch J, Llewellyn A, Griffiths C, Yang H, Owen C, et al. The use of measures of obesity in childhood for predicting obesity and the development of obesity-related diseases in adulthood: A systematic review and meta-analysis. Health Technology Assessment. 2015 6;19(43):1–336. Available from: https://www.journalslibrary.nihr.ac.uk/hta/hta19430/.

[120] Romero-Corral A, Somers VK, Sierra-Johnson J, Korenfeld Y, Boarin S, Korinek J, et al. Normal weight obesity: A risk factor for cardiometabolic dysregulation and cardiovascular mortality. European Heart Journal. 2010 3;31(6):737–746. Available from: https://academic.oup.com/eurheartj/article-lookup/doi/10.1093/eurheartj/ehp487.

[121] Wiklund P, Törmäkangas T, Shi Y, Wu N, Vainionpää A, Alen M, et al. Normal-weight obesity and cardiometabolic risk: A 7-year longitudinal study in girls from prepuberty to early adulthood. Obesity. 2017 6;25(6):1077–1082. Available from: http://doi.wiley.com/10.1002/oby.21838.

[122] Gorber SC, Tremblay M, Moher D, Gorber B. A comparison of direct vs. self-report measures for assessing height, weight and body mass index: A systematic review. Obesity Reviews. 2007 7;8(4):307–326. Available from: http://doi.wiley.com/10.1111/j.1467-789X.2007.00347.x.

[123] Lipsky LM, Haynie DL, Hill C, Nansel TR, Li K, Liu D, et al. Accuracy of Self-Reported Height, Weight, and BMI Over Time in Emerging Adults. American Journal of Preventive Medicine. 2019 6;56(6):860–868. Available from: https://www.doi.org/10.1016/j.amepre.2019.01.004.

[124] Leatherdale ST, Brown KS, Carson V, Childs RA, Dubin JA, Elliott SJ, et al. The COMPASS study: A longitudinal hierarchical research platform for evaluating natural experiments related to changes in school-level programs, policies and built environment resources. BMC Public Health. 2014 12;14(1):331. Available from: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-331.

[125] Patte KA, Bredin C, Henderson J, Elton-Marshall T, Faulkner G, Sabiston C, et al. Development of a mental health module for the compass system: Improving youth mental health trajectories. Part 1: Tool Development and Design; 2017. 2. Available from: https://uwaterloo.ca/compass-system/development-mental-health-module-compass-system-improving.

[126] Patte KA, Bredin C, Henderson J, Elton-Marshall T, Faulkner G, Sabiston C, et al. Development of a mental health module for the COMPASS system: Improving youth mental health trajectories. Part 2: Pi-

lot Test and Focus Group Results. COMPASS Technical Report Series. 2017;4(3). Available from: https://uwaterloo.ca/compass-system/development-mental-health-module-compass-system-improving-0.

[127] Gaylis JB, Levy SS, Hong MY. Relationships between body weight perception, body mass index, physical activity, and food choices in Southern California male and female adolescents. International Journal of Adolescence and Youth. 2020;25(1):264–275. Available from: https://doi.org/10.1080/02673843.2019.1614465.

[128] Sirirassamee T, Phoolsawat S, Limkhunthammo S. Relationship between body weight perception and weight-related behaviours. Journal of International Medical Research. 2018;46(9):3796–3808. Available from: https://www.doi.org/10.1177/0300060518780138.

[129] Health Canada. History of Canada's Food Guides; 2019. Available from: https://www.canada.ca/en/health-canada/services/canada-food-guide/about/history-food-guide.html.

[130] Government of Canada. Canada's food guide; 2020. Available from: https://food-guide.canada.ca/en/.

[131] Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. Applied Psychological Measurement. 1977;1(3):385–401. Available from: https://www.doi.org/10.1177/014662167700100306.

[132] Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: The GAD-7. Archives of Internal Medicine. 2006;166(10):1092–1097. Available from: https://www.doi.org/10.1001/archinte.166.10.1092.

[133] Gratz KL, Roemer L. Multidimensional Assessment of Emotion Regulation and Dysregulation: Development, Factor Structure, and Initial Validation of the Difficulties in Emotion Regulation Scale. Journal of Psychopathology and Behavioral Assessment. 2004;26(1):41–54. Available from: https://www.doi.org/10.1023/B:JOBA.0000007455.08539.94.

[134] Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi Dw, Oishi S, et al. New well-being measures: Short scales to assess flourishing and positive and negative feelings. Social Indicators Research. 2010;97(2):143–156. Available from: https://www.doi.org/10.1007/s11205-009-9493-y.

[135] Marsh HW, Ellis LA, Parada RH, Richards G, Heubeck BG. A short version of the self description questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. Psychological Assessment. 2005;17(1):81–102. Available from: https://www.doi.org/10.1037/1040-3590.17.1.81.

[136] Hammami N. Gender Differences in Chronic Disease Risk Behaviours and their Association with Body Mass Index : Cross-sectional and Longitudinal Multilevel Analyses Among a Large Sample of Youth . University of Waterloo; 2019. Available from: https://uwspace.uwaterloo.ca/handle/10012/14909.

[137] Zuckermann AME, Gohari MR, de Groh M, Jiang Y, Leatherdale ST. Factors associated with cannabis use change in youth: Evidence from the COMPASS study. Addictive Behaviors. 2019 3;90:158–163. Available from: https://www.doi.org/10.1016/j.addbeh.2018.10.048.

[138] Butler A, Romano I, Patte K, Ferro MA, De Groh M, Jiang Y, et al. Psychological correlates and binge drinking behaviours among Canadian youth: A cross-sectional analysis of the mental health pilot data from the COMPASS study. BMJ Open. 2019 6;9(6):e028558. Available from: https://doi.org/10.1136/bmjopen-2018-028558.

[139] Cole AG, Aleyan S, Leatherdale ST. Exploring the association between E-cigarette retailer proximity and density to schools and youth E-cigarette use. Preventive Medicine Reports. 2019 9;15:100912. Available from: https://www.doi.org/10.1016/j.pmedr.2019.100912.

[140] Cole AG, Aleyan S, Qian W, Leatherdale ST. Assessing the strength of secondary school tobacco policies of schools in the COMPASS study and the association to student smoking behaviours. Canadian Journal of Public Health. 2019 4;110(2):236–243. Available from: https://doi.org/10.17269/s41997-019-00178-4.

[141] Ten Eyck P, Cavanaugh JE. An alternate approach to Pseudo-likelihood model selection in the generalized linear mixed modeling framework. Sankhya: The Indian Journal of Statistics. 2018;80B:98–122. Available from: https://www.doi.org/10.1007/s13571-017-0130-5.

[142] Schwarz G. Estimating the Dimension of a Model. The Annals of Statistics. 1978 3;6(2):461–464. Available from: https://www.doi.org/10.1214/aos/1176344136.

[143] Chaurasia A. Model-modified BIC as a competitor of BIC variants for model selection in regression and order selection in time series. Communications in Statistics - Theory and Methods. 2022 6;0(0):1–29. Available from: https://doi.org/10.1080/03610926.2022.2064497.

[144] Chaurasia A, Harel O. Model selection rates of information based criteria. Electronic Journal of Statistics. 2013 1;7(none):2762–2793. Available from: https://www.doi.org/10.1214/13-EJS861.

[145] Boehmke B, Greenwell B. Hands-On Machine Learning with R. CRC Press; 2020.

[146] Therneau T, Atkinson B, Ripley B. Package 'rpart'; 2019. Available from: https://cran.r-project.org/web/packages/rpart/rpart.pdf.

[147] Lin S, Luo W. A New Multilevel CART Algorithm for Multilevel Data with Binary Outcomes. Multivariate Behavioral Research. 2019 7;54(4):578–592. Available from: https://www.tandfonline.com/doi/full/10.1080/00273171.2018.1552555.

[148] van Buuren S. Package 'mice'; 2021. Available from: https://cran.r-project.org/web/packages/mice/mice.pdf.

[149] van Buuren S. Package 'miceadds'; 2022. Available from: https://cran.r-project.org/web/packages/miceadds/miceadds.pdf.

[150] Schafer JL, Zhao Jh. Package ' pan '; 2022. Available from: https://cran.r-project.org/web/packages/pan/pan.pdf.

[151] Adab P, Pallan M, Whincup PH. Is BMI the best measure of obesity? BMJ. 2018 3;360(March):k1274. Available from: http://dx.doi.org/doi:10.1136/bmj.k1274.

[152] Green MA. Do we need to think beyond BMI for estimating population-level health risks?: Table 1. Journal of Public Health. 2016 3;38(1):192–193. Available from: https://www.doi.org/10.1093/pubmed/fdv007.

[153] Must A, Anderson SE. Body mass index in children and adolescents: considerations for population-based applications. International Journal of Obesity. 2006 4;30(4):590–594. Available from: http://www.nature.com/articles/0803300.

[154] Thompson-Haile A, Leatherdale ST. Student-level data collection procedures; 2013. Available from: https://uwaterloo.ca/compass-system/publications/student-level-data-collection-procedures.

[155] Cavanaugh JE, Neath AA. The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements. Wiley Interdisciplinary Reviews: Computational Statistics. 2019;11(3):1–11. Available from: https://www.doi.org/10.1002/wics.1460.

[156] Neath AA, Cavanaugh JE. The Bayesian information criterion: Background, derivation, and applications. Wiley Interdisciplinary Reviews: Computational Statistics. 2012;4(2):199–203. Available from: https://www.doi.org/10.1002/wics.199.

[157] Sykes LL, Walker RL, Ngwakongnwi E, Quan H. A Systematic Literature Review on Response Rates across Racial and Ethnic Populations. Canadian Journal of Public Health. 2010 5;101(3):213–219. Available from: http://link.springer.com/10.1007/BF03404376.

[158] Downing A, West RM, Gilthorpe MS, Lawrence G, Forman D. Using routinely collected health data to investigate the association between ethnicity and breast cancer incidence and survival: what is the impact of missing data and multiple ethnicities? Ethnicity & Health. 2011 6;16(3):201–212. Available from: http://www.tandfonline.com/doi/abs/10.1080/13557858.2011.561301.

[159] Krieger N. Proximal, distal, and the politics of causation: What's level got to do with it? American Journal of Public Health. 2008 2;98(2):221–230. Available from: https://www.doi.org/10.2105/AJPH.2007.111278.

[160] Ma Y, Zhang W, Lyman S, Huang Y. The HCUP SID Imputation Project: Improving Statistical Inferences for Health Disparities Research by Imputing Missing Race Data. Health Services Research. 2018;53(3):1870–1889. Available from: https://www.doi.org/10.1111/1475-6773.12704.

[161] Felderer B, Kirchner A, Kreuter F. The Effect of Survey Mode on Data Quality: Disentangling Nonresponse and Measurement Error Bias. Journal of Official Statistics. 2019;35(1):93–115. Available from: https://www.doi.org/10.2478/jos-2019-0005.

[162] Tourangeau R, Yan T. Sensitive Questions in Surveys. Psychological Bulletin. 2007;133(5):859–883. Available from: https://www.doi.org/10.1037/0033-2909.133.5.859.

[163] Jones DC, Crawford JK. Adolescent boys and body image: Weight and muscularity concerns as dual pathways to body dissatisfaction. Journal of Youth and Adolescence. 2005;34(6):629–636. Available from: https://www.doi.org/10.1007/s10964-005-8951-3.

[164] McCabe MP, Ricciardelli LA. Body image and strategies to lose weight and increase muscle among boys and girls. Health Psychology. 2003;22(1):39–46. Available from: https://www.doi.org/10.1037/0278-6133.22.1.39.

[165] Deierlein AL, Malkan A, Litvak J, Parekh N. Weight perception, weight control intentions, and dietary intakes among adolescents ages 10–15 years in the United States. International Journal of Environmental Research and Public Health. 2019;16(6). Available from: https://www.doi.org/10.3390/ijerph16060990.

[166] Raffoul A, Goodman S, Hammond D, Kirkpatrick SI. Weight Management Efforts, But Not Weight Perceptions, Are Associated with Dietary Quality among Youth and Young Adults in Canada. Journal of the Academy of Nutrition and Dietetics. 2020;p. 1–10. Available from: https://doi.org/10.1016/j.jand.2020.10.011.

[167] Livermore M, Duncan MJ, Leatherdale ST, Patte KA. Are weight status and weight perception associated with academic performance among youth? Journal of Eating Disorders. 2020;8(1):1–10. Available from: https://www.doi.org/10.1186/s40337-020-00329-w.

[168] Shore SM, Sachs ML, Lidicker JR, Brett SN, Wright AR, Libonati JR. Decreased scholastic achievement in overweight middle school students. Obesity. 2008;16(7):1535–1538. Available from: https://www.doi.org/10.1038/oby.2008.254.

[169] Bucchianeri MM, Fernandes N, Loth K, Hannan PJ, Eisenberg ME, Neumark-Sztainer D. Body dissatisfaction: Do associations with disordered eating and psychological well-being differ across race/ethnicity in adolescent girls and boys? Cultural Diversity and Ethnic Minority Psychology. 2016;22(1):137–146. Available from: https://www.doi.org/10.1037/cdp0000036.

[170] Colunga-Rodríguez C, Orozco-Solis MG, Flores-Villavicencio ME, De-la Roca-Chiapas JM, Gómez-Martínez R, Mercado A, et al. Body Image Perception and Internalization Problems Indicators in Mexican Adolescents. Psychology. 2016;07(13):1671–1681. Available from: https://www.doi.org/10.4236/psych.2016.713158.

[171] Rawana JS, Morgan AS. Trajectories of Depressive Symptoms from Adolescence to Young Adulthood: The Role of Self-esteem and Body-Related Predictors. Journal of Youth and Adolescence. 2014;43(4):597–611. Available from: https://www.doi.org/10.1007/s10964-013-9995-4.

[172] Janz T. Current smoking trends; 2012. Available from: https://www150.statcan.gc.ca/n1/en/pub/82-624-x/2012001/article/11676-eng.pdf?st=2Z07PeLv.

[173] Woodgate RL, Busolo DS. A qualitative study on Canadian youth's perspectives of peers who smoke: An opportunity for health promotion. BMC Public Health. 2015;15(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4692065/pdf/12889_2015_Article_2683.pdf.

[174] Bernat D, Gasquet N, Wilson KOD, Porter L, Choi K. Electronic Cigarette Harm and Benefit Perceptions and Use Among Youth. American Journal of Preventive Medicine. 2018;55(3):361–367. Available from: http://dx.doi.org/10.1016/j.amepre.2018.04.043.

[175] Camenga DR, Fiellin LE, Pendergrass T, Miller E, Pentz MA, Hieftje K. Adolescents' perceptions of flavored tobacco products, including E-cigarettes: A qualitative study to inform FDA tobacco education efforts through videogames. Addictive Behaviors. 2018 7;82:189–194. Available from: https://doi.org/10.1016/j.addbeh.2018.03.021.

[176] Romijnders KAGJ, Osch Lv, Vries Hd, Talhout R. Perceptions and reasons regarding e-cigarette use among users and non-users: A narrative literature review. International Journal of Environmental Research and Public Health. 2018 6;15(6):1190. Available from: http://www.mdpi.com/1660-4601/15/6/1190.

[177] Maukonen M, Männistö S, Tolonen H. A comparison of measured versus self-reported anthropometrics for assessing obesity in adults: a literature review. Scandinavian Journal of Public Health. 2018;46(5):565–579. Available from: https://www.doi.org/10.1177/1403494818761971.

[178] Taylor AW, Dal Grande E, Gill TK, Chittleborough CR, Wilson DH, Adams RJ, et al. How valid are self-reported height and weight? A comparison between CATI self-report and clinic measurements using a large cohort study. Australian and New Zealand Journal of Public Health. 2006;30(3):238–246. Available from: https://www.doi.org/10.1111/j.1467-842X.2006.tb00864.x.

[179] Doggett A, Chaurasia A, Chaput JP, Leatherdale ST. Learning from missing data: examining nonreporting patterns of height, weight, and BMI among Canadian youth. International Journal of Obesity. 2022 9;46(9):1598–1607. Available from: https://www.nature.com/articles/s41366-022-01154-8.

[180] Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison with Logistic Regression. Annals of Behavioral Medicine. 2003;26(3):172–181. Available from: https://www.doi.org/10.1207/S15324796ABM2603_02.

[181] Loh WY. Fifty Years of Classification and Regression Trees. International Statistical Review. 2014 12;82(3):329–348. Available from: https://onlinelibrary.wiley.com/doi/10.1111/insr.12016.

[182] Leatherdale ST, Laxer RE. Reliability and validity of the weight status and dietary intake measures in the COMPASS questionnaire: are the self-reported measures of body mass index (BMI) and Canada's food guide servings robust? The international journal of behavioral nutrition and physical activity. 2013 4;10:42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23561578.

[183] Delk J, Creamer MLR, Perry CL, Harrell MB. Weight Status and Cigarette and Electronic Cigarette Use in Adolescents. American Journal of Preventive Medicine. 2018;54(1):e31–e35. Available from: http://dx.doi.org/10.1016/j.amepre.2017.09.007.

[184] Nelson TF, Stovitz SD, Thomas M, LaVoi NM, Bauer KW, Neumark-Sztainer D. Do Youth Sports Prevent Pediatric Obesity? A Systematic Review and Commentary. Current Sports Medicine Reports. 2011 11;10(6):360–370. Available from: http://journals.lww.com/00149619-201111000-00012.

[185] Roberts RE, Duong HT. Do Anxiety Disorders Play a Role in Adolescent Obesity? Annals of Behavioral Medicine. 2016;50(4):613–621. Available from: http://dx.doi.org/10.1007/s12160-016-9786-8.

[186] Wolters Kluwer Health. Epidemiology: Instructions for Authors; 2022. Available from: https://edmgr.ovid.com/epid/accounts/ifauth.htm.

[187] Karahalios A, Baglietto L, Lee KJ, English DR, Carlin JB, Simpson JA. The impact of missing data on analyses of a time-dependent exposure in a longitudinal cohort: A simulation study. Emerging Themes in Epidemiology. 2013;10(1):1–11. Available from: https://www.doi.org/10.1186/1742-7622-10-6.

[188] Masconi KL, Matsha TE, Echouffo-Tcheugui JB, Erasmus RT, Kengne AP. Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus: a systematic review. EPMA Journal. 2015 12;6(1):7. Available from: https://link.springer.com/10.1186/s13167-015-0028-0.

[189] Ho HCH, Maddaloni E, Buzzetti R. Risk factors and predictive biomarkers of early cardiovascular disease in obese youth. Diabetes/Metabolism Research and Reviews. 2019 5;35(4):e3134. Available from: https://onlinelibrary.wiley.com/doi/10.1002/dmrr.3134.

[190] Carrière G. Parent and child factors associated with youth obesity.; 2003. Dec. Available from: https://pubmed.ncbi.nlm.nih.gov/14768292/.

[191] Nonnemaker JM, Morgan-Lopez AA, Pais JM, Finkelstein EA. Youth BMI trajectories: Evidence from the NLSY97. Obesity. 2009;17(6):1274–1280. Available from: https://www.doi.org/10.1038/oby.2009.5.

[192] Davison KK, Birch LL. Child and parent characteristics as predictors of change in girls' body mass index. International Journal of Obesity. 2001;25(12):1834–1842. Available from: https://www.doi.org/10.1038/sj.ijo.0801835.

[193] Walsemann KM, Ailshire JA, Bell BA, Frongillo EA. Body mass index trajectories from adolescence to midlife: Differential effects of parental and respondent education by race/ethnicity and gender. Ethnicity and Health. 2012;17(4):337–362. Available from: https://www.doi.org/10.1080/13557858.2011.635374.

[194] Strobl C, Malley J, Tutz G. An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging, and Random Forests. Psychological Methods. 2009 12;14(4):323–348. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19968396.

[195] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3):1–67. Available from: https://www.doi.org/10.18637/jss.v045.i03.

[196] Gorman B. Decision Trees in R using rpart; 2014. Available from: https://www.gormanalysis.com/blog/decision-trees-in-r-using-rpart/.

[197] Vink G, van Buuren S. mice: Imputing multi-level data;. Available from: https://www.gerkovink.com/miceVignettes/Multi_level/Multi_level_data.html.

APPENDICES

Appendix A

COMPASS Study Funding

The COMPASS study has been supported by a bridge grant from the CIHR Institute of Nutrition, Metabolism and Diabetes (INMD) through the "Obesity – Interventions to Prevent or Treat" priority funding awards (OOP-110788; awarded to SL), an operating grant from the CIHR Institute of Population and Public Health (IPPH) (MOP-114875; awarded to SL), a CIHR project grant (PJT-148562; awarded to SL), a CIHR bridge grant (PJT-149092; awarded to KP/SL), a CIHR project grant (PJT-159693; awarded to KP), and by a research funding arrangement with Health Canada (#1617-HQ-000012; contract awarded to SL), a CIHR-Canadian Centre on Substance Use and Addiction (CCSA) team grant (OF7 B1-PCPEGT 410-10-9633; awarded to SL), a project grant from the CIHR Institute of Population and Public Health (IPPH) (PJT-180262; awarded to SL and KP).

A SickKids Foundation New Investigator Grant, in partnership with CIHR Institute of Human Development, Child and Youth Health (IHDCYH) (Grant No. NI21-1193; awarded to KAP) funds a mixed methods study examining the impact of the COVID-19 pandemic on youth mental health, leveraging COMPASS study data. The COMPASS-Quebec project additionally benefits from funding from the Ministère de la Santé et des Services sociaux of the province of Québec, and the Direction régionale de santé publique du CIUSSS de la Capitale-Nationale.

Appendix B

COMPASS Student Questionnaire

This following pages include the entire COMPASS student questionnaire for the 2018/19 data collection year.

- **This is NOT a test.** All of your answers will be kept **confidential**. No one, not even your parents or teachers, will ever know what you answered. So, please be honest when you answer the questions.

- Mark only **one option per question** unless the instructions tell you to do something else.

- Choose the option that is the **closest** to what you think/feel is true for you.

Please, use a pencil to complete this questionnaire

Please mark all your answers with full, dark marks like this:

START HERE

Please read each sentence below carefully. Write the correct letter, number, or word on the line and then fill in the corresponding circle.

Note: These five questions are only used to link data from one year to the next. They cannot be used to identify participants. Only University of Waterloo researchers have access to the responses, and they never have access to student names or other information. All responses are strictly confidential.

The first letter of your **middle** name (if you have more than one middle name use your first middle name; if you don't have a middle name use "Z"):____	The name of the month in which you were born: ________	The **last** letter of your full **last** name: ___	The **second** letter of your full **first** name: ___	The **first** initial of your **mother's** first name (think about the mother you see the most):___
A J S B K T C L U D M V E N W F O X G P Y H Q Z I R	1 January 2 February 3 March 4 April 5 May 6 June 7 July 8 August 9 September 10 October 11 November 12 December	A J S B K T C L U D M V E N W F O X G P Y H Q Z I R	A J S B K T C L U D M V E N W F O X G P Y H Q Z I R	A J S B K T C L U D M V E N W F O X G P Y H Q Z I R

[serial]

About You

1. What grade are you in?

- (9) Grade 9
- (10) Grade 10
- (11) Grade 11
- (12) Grade 12

Quebec students only

- (1) Secondary I
- (2) Secondary II
- (3) Secondary III
- (4) Secondary IV
- (5) Secondary V
- (6) Other

2. How old are you today?

- (12) 12 years or younger
- (13) 13 years
- (14) 14 years
- (15) 15 years
- (16) 16 years
- (17) 17 years
- (18) 18 years
- (19) 19 years or older

3. Are you female or male?

- (1) Female
- (2) Male

4. How would you describe yourself? *(Mark all that apply)*

- (1) White
- (1) Black
- (1) Asian
- (1) Aboriginal (First Nations, Métis, Inuit)
- (1) Latin American/Hispanic
- (1) Other

5. About how much money do you usually get each week to spend on yourself or to save?
(Remember to include all money from allowances and jobs like baby-sitting, delivering papers, etc.)

- (1) Zero
- (2) $1 to $5
- (3) $6 to $10
- (4) $11 to $20
- (5) $21 to $40
- (6) $41 to $100
- (7) More than $100
- (8) I do not know how much money I get each week

6. How do you <u>usually</u> travel to and from school? *(If you use two or more modes of travel, choose the one that you spend most <u>time</u> doing)*

<u>To school</u>
1. By car (as a passenger)
2. By car (as a driver)
3. By school bus
4. By public bus, subway, or streetcar
5. By walking
6. By bicycling
7. Other

<u>From school</u>
1. By car (as a passenger)
2. By car (as a driver)
3. By school bus
4. By public bus, subway, or streetcar
5. By walking
6. By bicycling
7. Other

7. Did you attend <u>this</u> school last year?
1. Yes, I attended the same school last year
2. No, I was at another school last year

8. How tall are you <u>without</u> your shoes on? (Please write your height in feet and inches **<u>OR</u>** in centimetres, and then fill in the appropriate numbers for your height.)

1. **I do not know how tall I am**

"My height is ____feet, _____inches"

OR

"My height is ______ centimetres"

Height		
Feet	Inches	
0	0	0
1	1	1
2		2
3		3
4		4
5		5
6		6
7		7
		8
		9

OR

Height		
Centimetres		
0	0	0
1	1	1
2	2	2
	3	3
	4	4
	5	5
	6	6
	7	7
	8	8
	9	9

***Example*:**
My height is <u>5 ft 7 in</u>

Height		
Feet	Inches	
0	●	0
1	1	1
2		2
3		3
4		4
●		5
6		6
7		●
		8
		9

9. How much do you weigh <u>without</u> your shoes on? (Please write your weight in pounds **<u>OR</u>** in kilograms, and then fill in the appropriate numbers for your weight.)

1. **I do not know how much I weigh**

"My weight is _______ pounds"

OR

"My weight is _______ kilograms"

Weight		
Pounds		
0	0	0
1	1	1
2	2	2
3	3	3
	4	4
	5	5
	6	6
	7	7
	8	8
	9	9

OR

Weight		
Kilograms		
0	0	0
1	1	1
	2	2
	3	3
	4	4
	5	5
	6	6
	7	7
	8	8
	9	9

***Example*:**
My weight is <u>127 lbs</u>

Weight		
Pounds		
0	0	0
●	1	1
2	●	2
3	3	3
	4	4
	5	5
	6	6
	7	●
	8	8
	9	9

[serial]

10. How do you describe your weight?

1. Very underweight
2. Slightly underweight
3. About the right weight
4. Slightly overweight
5. Very overweight

11. Which of the following are you trying to do about your weight?

1. **Lose** weight
2. **Gain** weight
3. **Stay** the same weight
4. I am **not trying to do anything** about my weight

12. How much time per day do you *usually* spend doing the following activities?

For example: If you spend about 3 hours watching TV each day, you will need to fill in the 3 hour circle, and the 0 minute circle as shown below:

	Hours										Minutes			
a) Watching/streaming TV shows or movies	0	1	2	●	4	5	6	7	8	9	●	15	30	45

	Hours										Minutes			
a) Watching/streaming TV shows or movies	0	1	2	3	4	5	6	7	8	9	0	15	30	45
b) Playing video/computer games	0	1	2	3	4	5	6	7	8	9	0	15	30	45
c) Doing homework	0	1	2	3	4	5	6	7	8	9	0	15	30	45
d) Talking on the phone	0	1	2	3	4	5	6	7	8	9	0	15	30	45
e) Surfing the internet	0	1	2	3	4	5	6	7	8	9	0	15	30	45
f) Texting, messaging, emailing **(note: 50 texts = 30 minutes)**	0	1	2	3	4	5	6	7	8	9	0	15	30	45
g) Sleeping	0	1	2	3	4	5	6	7	8	9	0	15	30	45

13. In the last 30 days, did you gamble online for money?

1. Yes
2. No

Physical Activity

HARD physical activities include jogging, team sports, fast dancing, jump-rope, and any other physical activities that increase your heart rate and make you breathe hard and sweat.

MODERATE physical activities include lower intensity activities such as walking, biking to school, and recreational swimming.

14. Mark how many minutes of HARD physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time.

	Hours					Minutes			
Monday	0	1	2	3	4	0	15	30	45
Tuesday	0	1	2	3	4	0	15	30	45
Wednesday	0	1	2	3	4	0	15	30	45
Thursday	0	1	2	3	4	0	15	30	45
Friday	0	1	2	3	4	0	15	30	45
Saturday	0	1	2	3	4	0	15	30	45
Sunday	0	1	2	3	4	0	15	30	45

For example: If you did 45 minutes of hard physical activity on Monday, you will need to fill in the 0 hour circle and the 45 minute circle, as shown below:

	Hours					Minutes			
Monday	●	1	2	3	4	0	15	30	●

15. Mark how many minutes of MODERATE physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time. Do not include time spent doing hard physical activities.

	Hours					Minutes			
Monday	0	1	2	3	4	0	15	30	45
Tuesday	0	1	2	3	4	0	15	30	45
Wednesday	0	1	2	3	4	0	15	30	45
Thursday	0	1	2	3	4	0	15	30	45
Friday	0	1	2	3	4	0	15	30	45
Saturday	0	1	2	3	4	0	15	30	45
Sunday	0	1	2	3	4	0	15	30	45

For example: If you did 1 hour and 30 minutes of moderate physical activity on Monday, you will need to fill in the 1 hour circle and the 30 minute circle, as shown below:

	Hours					Minutes			
Monday	0	●	2	3	4	0	15	●	45

16. Were the last 7 days a typical week in terms of the amount of physical activity that you usually do?

1. Yes
2. No, I was *more* active in the last 7 days
3. No, I was *less* active in the last 7 days

[serial]

17. Your closest friends are the friends you like to spend the most time with. How many of your closest friends are <u>physically active</u>?

- ⓪ None
- ① 1 friend
- ② 2 friends
- ③ 3 friends
- ④ 4 friends
- ⑤ 5 or more friends

18. Are you taking a physical education class at school this year?

- ① **Yes**, I am taking one **this term**
- ② **Yes,** I will be taking one or have taken one this school year, **but not this term**.
- ③ **No**, I am not taking a physical education class at school this year

19. Do you participate in before-school, noon hour, or after-school physical activities organized by your school? *(e.g., intramurals, non-competitive clubs)*

- ① Yes
- ② No
- ③ None offered at my school

20. Do you participate in competitive school sports teams that compete against other schools? *(e.g., junior varsity or varsity sports)*

- ① Yes
- ② No
- ③ None offered at my school

21. Do you participate in league or team sports outside of school?

- ① Yes
- ② No
- ③ There are none available where I live

22. On how many days in the last 7 days did you do exercises to strengthen or tone your muscles? *(e.g., push-ups, sit-ups, or weight-training)*

- ⓪ 0 days
- ① 1 day
- ② 2 days
- ③ 3 days
- ④ 4 days
- ⑤ 5 days
- ⑥ 6 days
- ⑦ 7 days

Healthy Eating

23. If you do not eat breakfast every day, why do you skip breakfast? *(Mark all that apply)*

- ① I eat breakfast every day

- ① I don't have time for breakfast
- ① The bus comes too early
- ① I sleep in
- ① I'm not hungry in the morning
- ① I feel sick when I eat breakfast
- ① I'm trying to lose weight
- ① There is nothing to eat at home
- ① Other

24. In a *usual* school week *(Monday to Friday)*, on how many days do you do the following?

	None	1 day	2 days	3 days	4 days	5 days
a) Eat breakfast	⓪	①	②	③	④	⑤
b) Eat breakfast provided to you **as part of a school program**	⓪	①	②	③	④	⑤
c) Eat lunch at school - lunch packed and brought from home	⓪	①	②	③	④	⑤
d) Eat lunch at school - lunch purchased in the cafeteria	⓪	①	②	③	④	⑤
e) Eat lunch purchased at a fast food place or restaurant	⓪	①	②	③	④	⑤
f) Eat snacks purchased from a vending machine **in your school**	⓪	①	②	③	④	⑤
g) Eat snacks purchased from a vending machine, corner store, snack bar, or canteen **off school property**	⓪	①	②	③	④	⑤
h) Drink sugar-sweetened beverages (soda pop, Kool-Aid, Gatorade, etc.) Do not include diet/sugar-free drinks	⓪	①	②	③	④	⑤
i) Drink high-energy drinks (Red Bull, Monster, Rock Star, etc.)	⓪	①	②	③	④	⑤
j) Drink coffee or tea **with sugar** (include cappuccino, frappuccino, iced-tea, iced-coffees, etc.)	⓪	①	②	③	④	⑤
k) Drink coffee or tea **without sugar**	⓪	①	②	③	④	⑤

25. On a *usual* weekend *(Saturday and Sunday)*, on how many days do you do the following?

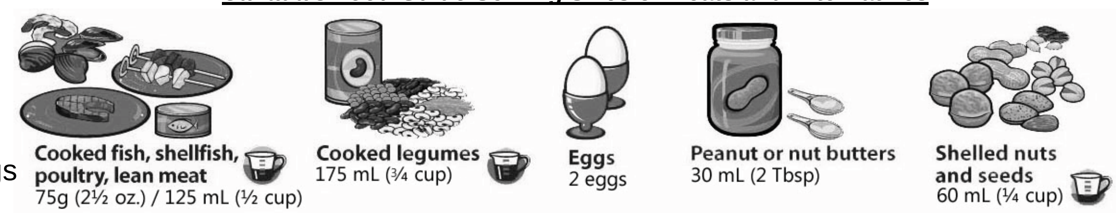
	None	1 day	2 days
a) Eat breakfast	⓪	①	②
b) Eat lunch	⓪	①	②
c) Eat foods purchased at a fast food place or restaurant	⓪	①	②
d) Eat snacks purchased from a vending machine, corner store, snack bar, or canteen	⓪	①	②
e) Drink sugar-sweetened beverages (soda pop, Kool-Aid, Gatorade, etc.) Do not include diet/sugar-free drinks	⓪	①	②
f) Drink high energy drinks (Red Bull, Monster, Rock Star, etc.)	⓪	①	②
g) Drink coffee or tea **with sugar** (include cappuccino, frappuccino, iced-tea, iced-coffees, etc.)	⓪	①	②
h) Drink coffee or tea **without sugar**	⓪	①	②

[serial]

26. YESTERDAY, *from the time you woke up until the time you went to bed*, how many servings of meats and alternatives did you have? *One 'Food Guide' serving of meat and alternatives includes cooked fish, chicken, beef, pork, or game meat, eggs, nuts or seeds, peanut butter or nut butters, legumes (beans), and tofu.*

- (0) None
- (1) 1 serving
- (2) 2 servings
- (3) 3 servings
- (4) 4 servings
- (5) 5 or more servings

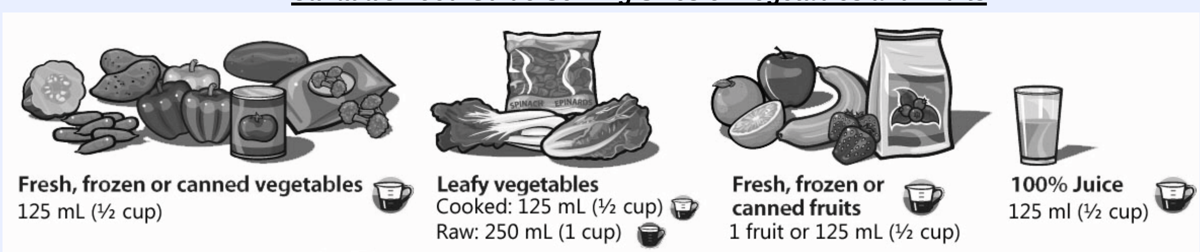
***Canada's Food Guide* Serving Sizes of Meats and Alternatives**



27. YESTERDAY, *from the time you woke up until the time you went to bed*, how many servings of vegetables and fruits did you have? *One 'Food Guide' serving of vegetables and fruit includes pieces of fresh vegetable or fruit, salad or raw leafy greens, cooked leafy green vegetables, dried or canned or frozen fruit, and 100% fruit or vegetable juice.*

- (0) None
- (1) 1 serving
- (2) 2 servings
- (3) 3 servings
- (4) 4 servings
- (5) 5 servings
- (6) 6 servings
- (7) 7 servings
- (8) 8 servings
- (9) 9 or more servings

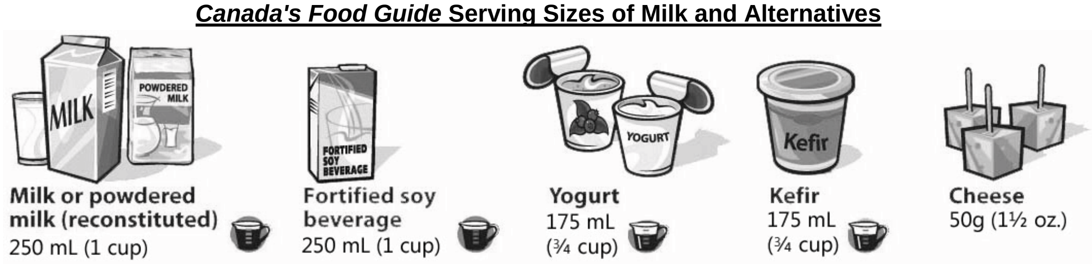
***Canada's Food Guide* Serving Sizes of Vegetables and Fruits**



28. YESTERDAY, *from the time you woke up until the time you went to bed*, how many servings of milk and alternatives did you have? *One 'Food Guide' serving of milk or milk alternatives includes milk, fortified soy beverage, reconstituted powdered milk, canned (evaporated) milk, yogurt or kefir (another type of cultured milk product), and cheese.*

- (0) None
- (1) 1 serving
- (2) 2 servings
- (3) 3 servings
- (4) 4 servings
- (5) 5 servings
- (6) 6 or more servings

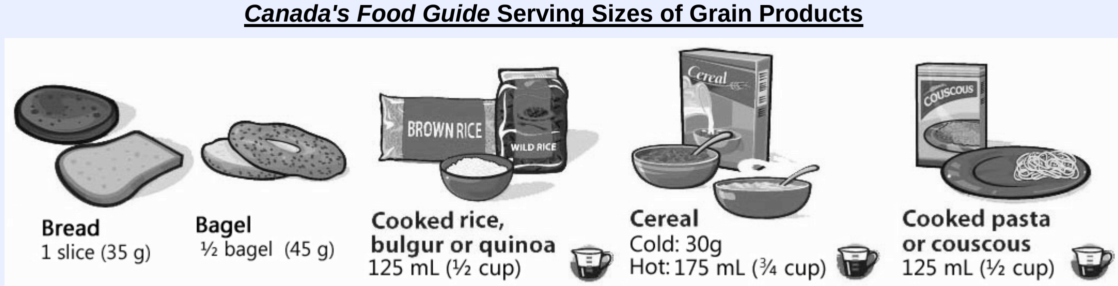
***Canada's Food Guide* Serving Sizes of Milk and Alternatives**



29. YESTERDAY, *from the time you woke up until the time you went to bed*, how many servings of grain products did you have? *One 'Food Guide' serving of grain products includes bread, bagels, flatbread such as tortilla, pita, cooked rice or pasta, and cold cereal.*

- (0) None
- (1) 1 serving
- (2) 2 servings
- (3) 3 servings
- (4) 4 servings
- (5) 5 servings
- (6) 6 servings
- (7) 7 servings
- (8) 8 servings
- (9) 9 or more servings

***Canada's Food Guide* Serving Sizes of Grain Products**



Your Experience with Smoking

30. Have you ever tried cigarette smoking, even just a few puffs?

1. Yes
2. No

31. Do you think in the future you might try smoking cigarettes?

1. Definitely yes
2. Probably yes
3. Probably not
4. Definitely not

32. If one of your best friends were to offer you a cigarette, would you smoke it?

1. Definitely yes
2. Probably yes
3. Probably not
4. Definitely not

33. At any time during the next year do you think you will smoke a cigarette?

1. Definitely yes
2. Probably yes
3. Probably not
4. Definitely not

34. Have you ever smoked 100 or more whole cigarettes in your life?

1. Yes
2. No

35. On how many of the last 30 days did you smoke one or more cigarettes?

1. None
2. 1 day
3. 2 to 3 days
4. 4 to 5 days
5. 6 to 10 days
6. 11 to 20 days
7. 21 to 29 days
8. 30 days *(every day)*

36. Your closest friends are the friends you like to spend the most time with. How many of your closest friends smoke cigarettes?

0. None
1. 1 friend
2. 2 friends
3. 3 friends
4. 4 friends
5. 5 or more friends

37. Have you <u>ever</u> tried to quit smoking cigarettes?

1. I have never smoked
2. I have only smoked a few times
3. I have never tried to quit
4. I have tried to quit once
5. I have tried to quit 2 or 3 times
6. I have tried to quit 4 or 5 times
7. I have tried to quit 6 or more times

38. Have you <u>ever</u> tried an <u>electronic cigarette</u>, also known as an e-cigarette?

1. Yes
2. No

39. Have you used <u>e-cigarettes</u> for any of the following reasons? *(Mark all that apply)*

- ① I have not used e-cigarettes
- ① Curiosity / to try something new
- ① I can use e-cigarettes in places where smoking is not allowed
- ① To smoke fewer cigarettes
- ① To help me quit smoking cigarettes
- ① I have used e-cigarettes for some other reason

40. In the <u>last 30 days</u>, did you use any of the following? *(Mark all that apply)*

- ① Pipe tobacco
- ① Cigarillos or little cigars (*plain or flavoured*)
- ① Cigars (not including cigarillos or little cigars, *plain or flavoured*)
- ① Roll-your-own cigarettes (tobacco only)
- ① Loose tobacco mixed with marijuana
- ① E-cigarettes (electronic cigarettes that \produce vapour instead of smoke, not including Juul)
- ① Juul
- ① Smokeless tobacco (chewing tobacco, pinch, snuff, or snus)
- ① Nicotine patches, nicotine gum, nicotine lozenges, or nicotine inhalers
- ① Hookah (water-pipe) to smoke tobacco
- ① Hookah (water-pipe) to smoke herbal sheesha/shisha
- ① Blunt wraps (a sheet or tube made of tobacco used to roll cigarette tobacco)
- ① I have not used any of these things in the last 30 days

41a. On how many of the <u>last 30 days</u> did you use an <u>e-cigarette</u> (not including Juul)?

- ① None
- ② 1 day
- ③ 2 to 3 days
- ④ 4 to 5 days
- ① 6 to 10 days
- ② 11 to 20 days
- ③ 21 to 29 days
- ④ 30 days *(every day)*

41b. On how many of the <u>last 30 days</u> did you use Juul?

- ① None
- ② 1 day
- ③ 2 to 3 days
- ④ 4 to 5 days
- ① 6 to 10 days
- ② 11 to 20 days
- ③ 21 to 29 days
- ④ 30 days *(every day)*

[serial]

Alcohol and Drug Use

Please remember that we will keep your answers ***completely confidential***.

A DRINK means: 1 regular sized bottle, can, or draft of beer; 1 glass of wine; 1 bottle of cooler; 1 shot of liquor (rum, whisky, etc); or 1 mixed drink (1 shot of liquor with pop, juice, energy drink).

42. In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?

- ① I have never drunk alcohol
- ② I did not drink alcohol in the last 12 months
- ③ I have only had a sip of alcohol
- ④ Less than once a month
- ⑤ Once a month
- ⑥ 2 or 3 times a month
- ⑦ Once a week
- ⑧ 2 or 3 times a week
- ⑨ 4 to 6 times a week
- ◯ Every day

43. How old were you when you first had a drink of alcohol that was more than just a sip?

- ① I have never drunk alcohol
- ② I have only had a sip of alcohol
- ③ I do not know

- ⑧ 8 years or younger
- ⑨ 9 years
- ⑩ 10 years
- ⑪ 11 years
- ⑫ 12 years
- ⑬ 13 years
- ⑭ 14 years
- ⑮ 15 years
- ⑯ 16 years
- ⑰ 17 years
- ⑱ 18 years or older

44. In the last 12 months, how often did you have 5 drinks of alcohol or more on one occasion?

- ① I have never done this
- ② I did not have 5 or more drinks on one occasion in the last 12 months
- ③ Less than once a month
- ④ Once a month
- ⑤ 2 to 3 times a month
- ⑥ Once a week
- ⑦ 2 to 5 times a week
- ⑧ Daily or almost daily

45. In the last 12 months, have you had alcohol mixed or pre-mixed with an energy drink (such as Red Bull, Rock Star, Monster, or another brand)?

- ① I have never done this
- ② I did not do this in the last 12 months
- ③ Yes
- ④ I do not know

46. In the <u>last 12 months</u>, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)

1. I have never used marijuana
2. I have used marijuana but not in the last 12 months
3. Less than once a month
4. Once a month
5. 2 or 3 times a month
6. Once a week
7. 2 or 3 times a week
8. 4 to 6 times a week
9. Every day

47. If you have used marijuana or cannabis in the <u>last 12 months</u>, how did you use it? *(Mark all that apply)*

- ① I have used it by smoking it (e.g., in a joint, a pipe, a bong)
- ① I have used it by vaping it
- ① I have used it by eating or drinking it (e.g., in brownies, cookies, candies, tea)
- ① I have not used marijuana or cannabis in the last 12 months

48. How old were you when you first used marijuana or cannabis?

- ① I have never used marijuana
- ② I do not know

- ⑧ 8 years or younger
- ⑨ 9 years
- ⑩ 10 years
- ⑪ 11 years
- ⑫ 12 years
- ⑬ 13 years
- ⑭ 14 years
- ⑮ 15 years
- ⑯ 16 years
- ⑰ 17 years
- ⑱ 18 years or older

49. Do you think it would be difficult or easy for you to get marijuana if you wanted some?

1. Difficult
2. Easy
3. I do not know

50. Have you used or tried any of the following medications TO GET HIGH?

	No, I have <u>never</u> done this	**Yes**, I have done this in the <u>last 12 months</u>	**Yes**, I have done this, but <u>NOT</u> in the <u>last 12 months</u>
a) Oxycodone (oxy, OC, APO, OxyContin®, percs, roxies, OxyNEO®)	①	②	③
b) Fentanyl (china white, synthetic heroin, china girl)	①	②	③
c) Other prescription pain relievers (codeine, morphine, Tylenol 3)	①	②	③

51. Do you think it would be difficult or easy to get pain relievers (Oxycodone, Fentanyl, codeine, etc.) if you wanted some?

1. Difficult
2. Easy
3. I do not know

[serial]

Mental Health

52. How much do you agree or disagree with the following statements?	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
a) I have a happy home life	1	2	3	4	5
b) My parents/guardians expect too much of me	1	2	3	4	5
c) I can talk about my problems with my family	1	2	3	4	5
d) I can talk about my problems with my friends	1	2	3	4	5

53. How much do you agree or disagree with the following statements?	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
a) I lead a purposeful and meaningful life	1	2	3	4	5
b) My social relationships are supportive and rewarding	1	2	3	4	5
c) I am engaged and interested in my daily activities	1	2	3	4	5
d) I actively contribute to the happiness and well-being of others	1	2	3	4	5
e) I am competent and capable in the activities that are important to me	1	2	3	4	5
f) I am a good person and live a good life	1	2	3	4	5
g) I am optimistic about my future	1	2	3	4	5
h) People respect me	1	2	3	4	5
i) I generally recover from setbacks quickly	1	2	3	4	5

54. Choose the answer that best describes how you feel.	True	Mostly true	Sometimes true, sometimes false	Mostly false	False
a) In general, I like the way I am	1	2	3	4	5
b) Overall, I have a lot to be proud of	1	2	3	4	5
c) A lot of things about me are good	1	2	3	4	5
d) When I do something, I do it well	1	2	3	4	5
e) I like the way I look	1	2	3	4	5

55. If you had concerns regarding your mental health, are there any reasons why you would <u>not</u> talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? *(Mark all that apply)*

1 I would have <u>no</u> problem talking to an adult at school about my mental health

1 Worried about what others would think of me (e.g., I'd be too embarrassed)
1 Lack of trust in these people - word would get out
1 Prefer to handle problems myself
1 Do not think these people would be able to help
1 Would not know who to approach
1 There is no one I feel comfortable talking to

56. Over the last **2 weeks**, how often have you been bothered by the following problems?

	Not at all	Several days	Over half the days	Nearly every day
a) Feeling nervous, anxious, or on edge	1	2	3	4
b) Not being able to stop or control worrying	1	2	3	4
c) Worrying too much about different things	1	2	3	4
d) Trouble relaxing	1	2	3	4
e) Being so restless that it is hard to sit still	1	2	3	4
f) Becoming easily annoyed or irritable	1	2	3	4
g) Feeling afraid as if something awful might happen	1	2	3	4

57. Please indicate how often the following statements apply to you:

	Almost never	Sometimes	About half the time	Most of the time	Almost always
a) I have difficulty making sense out of my feelings	1	2	3	4	5
b) I pay attention to how I feel	1	2	3	4	5
c) When I'm upset, I have difficulty concentrating	1	2	3	4	5
d) When I'm upset, I believe there is nothing I can do to make myself feel better	1	2	3	4	5
e) When I'm upset, I lose control over my behaviour	1	2	3	4	5
f) When I'm upset, I feel ashamed for feeling that way	1	2	3	4	5

58. On how many of the **last 7 days** did you feel the following ways?

	None or less than 1 day	1-2 days	3-4 days	5-7 days
a) I was bothered by things that usually don't bother me	1	2	3	4
b) I had trouble keeping my mind on what I was doing	1	2	3	4
c) I felt depressed	1	2	3	4
d) I felt that everything I did was an effort	1	2	3	4
e) I felt hopeful about the future	1	2	3	4
f) I felt fearful	1	2	3	4
g) My sleep was restless	1	2	3	4
h) I was happy	1	2	3	4
i) I felt lonely	1	2	3	4
j) I could not get "going"	1	2	3	4

59. In general, how would you rate your mental health?

1. Excellent
2. Very good
3. Good
4. Fair
5. Poor

If you are a young person in Canada who needs support, you can reach out to Kids Help Phone's professional counsellors by calling 1-800-668-6868 or visiting kidshelpphone.ca. Their service is free, anonymous, confidential, and available 24/7/365.

Kids Help Phone

1-800-668-6868

[serial]

Your School and You

60. How strongly do you agree or disagree with each of the following statements?

	Strongly agree	Agree	Disagree	Strongly disagree
a) I feel close to people at my school	①	②	③	④
b) I feel I am part of my school	①	②	③	④
c) I am happy to be at my school	①	②	③	④
d) I feel the teachers at my school treat me fairly	①	②	③	④
e) I feel safe in my school	①	②	③	④
f) Getting good grades is important to me	①	②	③	④

61. In the <u>last 30 days</u>, in what ways were you bullied by other students? *(Mark all that apply)*

① I have not been bullied in the last 30 days
① Physical attacks (e.g., getting beaten up, pushed, or kicked)
① Verbal attacks (e.g., getting teased, threatened, or having rumours spread about you)
① Cyber-attacks (e.g., being sent mean text messages or having rumours spread about you on the internet)
① Had someone steal from you or damage your things

62. In the <u>last 30 days</u>, how often have you been bullied by other students?

① I have not been bullied by other students in the last 30 days
② Less than once a week
③ About once a week
④ 2 or 3 times a week
⑤ Daily or almost daily

63. In the <u>last 30 days</u>, in what ways did you bully other students? *(Mark all that apply)*

① I did not bully other students in the last 30 days
① Physical attacks (e.g., beat up, pushed, or kicked them)
① Verbal attacks (e.g., teased, threatened, or spread rumours about them)
① Cyber-attacks (e.g., sent mean text messages or spread rumours about them on the internet)
① Stole from them or damaged their things

64. In the <u>last 30 days</u>, how often have you taken part in bullying other students?

① I did not bully other students in the last 30 days
② Less than once a week
③ About once a week
④ 2 or 3 times a week
⑤ Daily or almost daily

65. How supportive is your school of the following?

	Very supportive	Supportive	Unsupportive	Very unsupportive
a) Making sure there are opportunities for students to be physically active	①	②	③	④
b) Making sure students have access to healthy foods and drinks	①	②	③	④
c) Making sure no one is bullied at school	①	②	③	④
d) Giving students the support they need to resist or quit tobacco	①	②	③	④
e) Giving students the support they need to resist or quit drugs and/or alcohol	①	②	③	④

66. In your current or most recent Math course, what is your approximate overall mark?
(Think about last year if you have not taken math this year)

① 90% - 100%
② 80% - 89%
③ 70% - 79%
④ 60% - 69%
⑤ 55% - 59%
⑥ 50% - 54%
⑦ Less than 50%

67. In your current or most recent English course, what is your approximate overall mark?
(Think about last year if you have not taken English this year)

① 90% - 100%
② 80% - 89%
③ 70% - 79%
④ 60% - 69%
⑤ 55% - 59%
⑥ 50% - 54%
⑦ Less than 50%

68. What is the highest level of education you would like to get? *(Choose only one)*

① Some high school or less
② High school diploma or graduation equivalency
③ College/trade/vocational certificate
④ University Bachelor's degree
⑤ University Master's / PhD / law school / medical school / teachers' college degree
⑥ I don't know

69. What is the highest level of education you think you will get? *(Choose only one)*

① Some high school or less
② High school diploma or graduation equivalency
③ College/trade/vocational certificate
④ University Bachelor's degree
⑤ University Master's / PhD / law school / medical school / teachers' college degree
⑥ I don't know

70. In the last 4 weeks, how many days of school did you miss because of your health?

① 0 days
② 1 or 2 days
③ 3 to 5 days
④ 6 to 10 days
⑤ 11 or more days

71. In the last 4 weeks, how many classes did you skip when you were not supposed to?

① 0 classes
② 1 or 2 classes
③ 3 to 5 classes
④ 6 to 10 classes
⑤ 11 to 20 classes
⑥ More than 20 classes

72. How often do you go to class without your homework complete?

① Never
② Seldom
③ Often
④ Usually

[serial]

Appendix C

CART Details and Diagnostics

This Appendix provides additional methodological details about the CART modelling procedures from Study 2, which could not be covered within the word limits of a submitted manuscript. For the purposes of this Appendix only, the CART modelling performed in Study 2 will be referred to as "S-CART", where the "S" signifies single-level. This is in order to contrast it to M-CART (where the "M" signifies multi-level). R Code corresponding to the CART modelling for the results presented in Study 2 can also be found here.

C.1 Cost-Complexity Pruning

The pruning method used in Study 2 was a post-pruning approach referred to as cost complexity pruning. Cost complexity pruning involves identifying at which point a tree has an acceptable balance between accuracy and tree size [145]. The value of a complexity parameter (CP) is used to determine where the ideal trade-off between accuracy and parsimony exists. Conveniently, R's rpart package used to build the CART models in Study 2 performs 10-fold cross validation by default, simplifying the cost complexity pruning process. In order to find an appropriate CP value at which to prune the tree in Study 2, the 1-SE rule was used. The 1-SE rule refers to the process of finding the least complex tree that lies within 1 standard error of the minimum cross validation error [145]. The overall steps used to conduct the cost complexity pruning and implement the 1-SE rule in Study 2 are written below. For easier comparison with Section C.3, labels produced by R's default cross-validation are given where relevant.

1. Perform 10-fold cross validation
2. Find the tree with the minimum cross-validation error (labelled 'xerror' in R)
3. To this minimum cross-validation error, add the associated standard deviation (labelled 'xstd' in R)
4. Identify the least complex tree (i.e. the tree with the smallest number of splits) that has an cross-validation error which is still lower than the value calculated in step 3
5. Prune the original tree using a CP value that is near the CP value associated with the tree identified in step 4.

Notably, guidelines differ with respect to Step 5, as to whether one should use a higher or lower CP value (i.e. to round up or round down) when implementing the 1-SE rule [145,196]. For the purposes of Study 2 where parsimony was preferred over marginal gains in accuracy, a CP value that was slightly higher than that identified in Step 4 was chosen.

C.2 Multi-level CART

Given the clustered nature of COMPASS data stemming from the school-based nature of data collections, any analysis performed should ideally account for school-level clustering. Multi-level approaches for CART were not available until quite recently. For regression trees, there now exists an R package (REEMTree) to build multi-level decisions trees where the outcome of interest is continuous. For classification trees, which is the method used in Study 2, no such R-package is available. However, Lin & Luo recently published an approach for multi-level classification trees [147]. Given the clustered nature of COMPASS data, best attempts were made to apply this multi-level CART (M-CART) approach. The algorithm was obtained from author Luo, and was edited slightly to comply with the COMPASS data. The algorithm successfully ran with COMPASS data, and did converge. However, problems arose at the cross-validation stage. To illustrate the problem, Figure C.1 shows the cross-validation plot for an S-CART model that models BMI missingness in the COMPASS 2018/19 data.

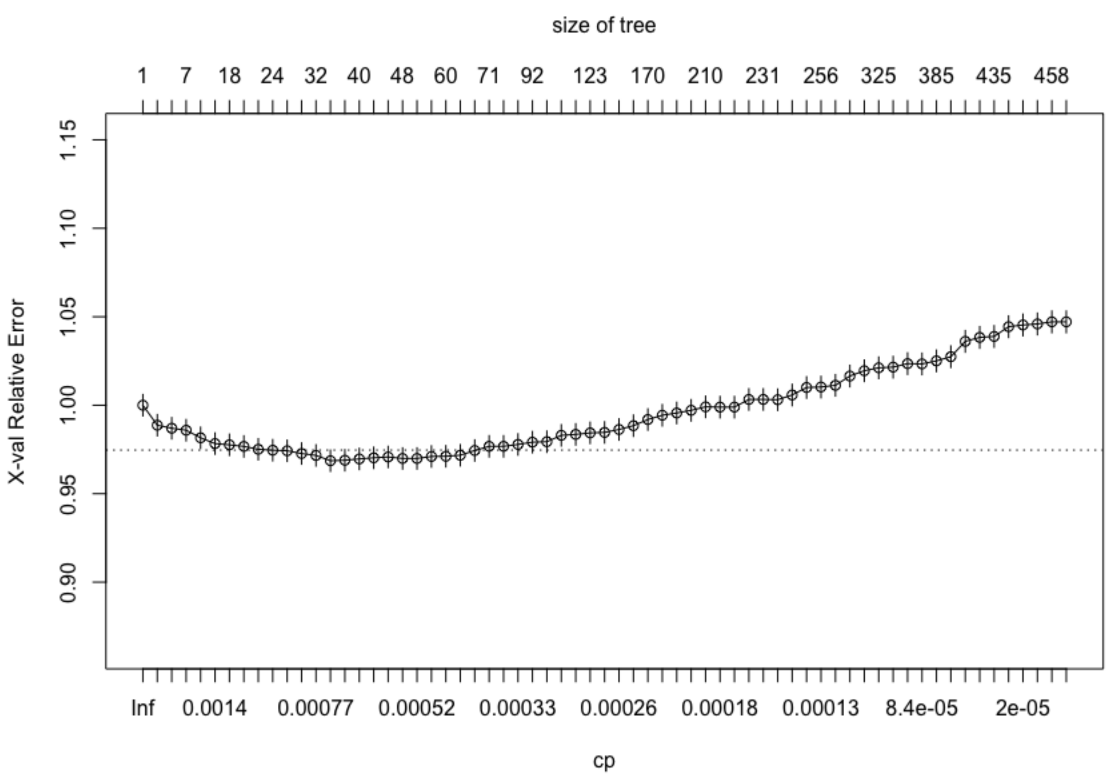


Figure C.1: Cross-validation plot for BMI Missingness using S-CART (COMPASS 2018/19)

The slightly u-shaped plot in Figure C.1 demonstrates that relative error decreases as tree size increases, but only up a certain point, after which there are diminishing returns and eventually a highly complex tree exceeds the error of smaller trees. The 1-SE rule described in Section C.1 is used to identify this point of diminishing returns. Unfortunately, this point of diminishing return could not be identified using the M-CART approach. Figure C.2 presents the cross-validation plot for the equivalent M-CART model which models BMI missingness in COMPASS 2018/19 data.

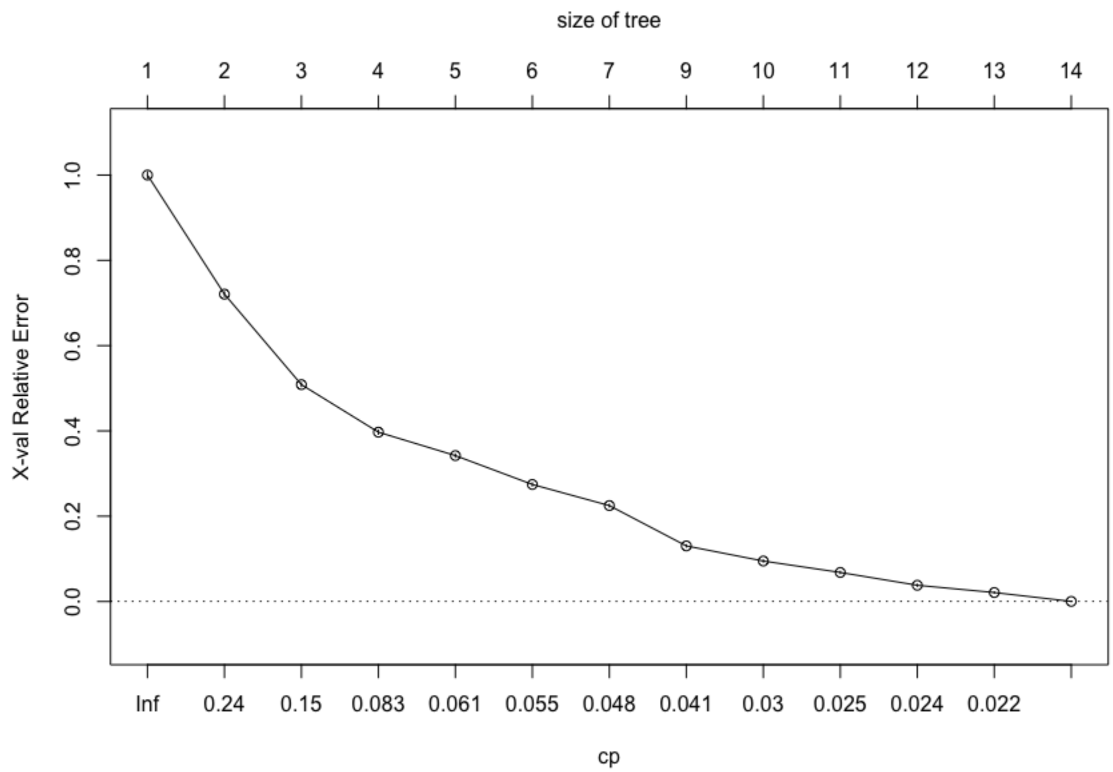


Figure C.2: Cross-validation plot for BMI Missingness using M-CART (COMPASS 2018/19)

Figure C.1 plot suggests that a more complex model is needed, as there are still gains in error as the size of the tree increases. However, M-CART models would not converge when a lower (i.e., more complex) CP value was implemented. Unlike the S-CART models, which were run with an unrestricted CP (an unrestricted CP is the lowest possible CP), M-CART models did not converge with a CP lower than 0.01. Notably, the fact that S-CART and M-CART models had to be run with different CP values is also the reason that the cross-validation plots in Figures C.1 and C.2 appear on different scales. The reason why M-CART implied that a highly complex tree was needed is unclear, however it is possible that the clustering effect of school did not have a sufficiently significant impact on missing BMI in this sample to warrant a multi-level model. Trying to force an overly complex model onto data that do not measure up to the complexity could be causing the issues described.

C.3 R Code for CART Implementation

Code for the construction of the female and male BMI missingness trees is included below. Code for the construction of the height and weight CART models is not included as it is largely similar to the code for the BMI CART models.

```
library(dplyr)
library(rpart)
library(rattle)

####BMI TREE - FEMALES####
data_Female <- data_Full[which(data_Full$SEX=="Female"),]
rm(list=ls()[ls()!="data_Female"])

##Split data into training and testing datasets (80/20)##
set.seed(145) #Ensure that the set.seed is run *with*(i.e., in
↪  conjunction, not in pieces) anything with random components
ind_BMI_F<- rbinom(nrow(data_Female), size=1, prob=0.8)

train_BMIID_F <- (1:nrow(data_Female))[as.logical(ind_BMI_F)]
test_BMIID_F <- (1:nrow(data_Female))[!as.logical(ind_BMI_F)]

train_BMI_Female <- data_Female[train_BMIID_F,]
test_BMI_Female <- data_Female[test_BMIID_F,]


##Build the initial tree##
tree_BMI_Females <- rpart(
  BMI_MISS ~ .,
  data = train_BMI_Female,
  method = "class",
  minbucket=14, #min number of female students in the smallest school
  cp=-1,
)


##Perform cross-validation##
set.seed(145) #NB: cross-validation has random components, run these fxs
↪  with seed for reproducability
plotcp(tree_BMI_Females)

#creation of a 'pick' column in cptable() fx to find lowest xerror more
↪  easily
min.val<- min(tree_BMI_Females$cptable[,'xerror'])
tree_BMI_Females$cptable[tree_BMI_Females$cptable[,'xerror'] == min.val,]
pick<- (tree_BMI_Females$cptable[,'xerror'] == min.val)
```

```
(    tree_BMI_Females$cptable<- cbind(tree_BMI_Females$cptable,
 ↪  pick=ifelse(pick,1,0))    )


##Prune the inital tree based on cross validation results##
tree_BMI_Females_pruned <- prune(tree_BMI_Females, cp =0.0015)
fancyRpartPlot(tree_BMI_Females_pruned)


##Compare accuracies between inital and pruned trees##
test_BMI_Female$pred <- predict(tree_BMI_Females,test_BMI_Female, type =
 ↪  "class")
base_accuracy_BMI_Females <- mean(test_BMI_Female$pred ==
 ↪  test_BMI_Female$BMI_MISS)

test_BMI_Female$pred <- predict(tree_BMI_Females_pruned,test_BMI_Female,
 ↪  type = "class")
pruned_accuracy_BMI_Females <- mean(test_BMI_Female$pred ==
 ↪  test_BMI_Female$BMI_MISS)

base_accuracy_BMI_Females
pruned_accuracy_BMI_Females

table(test_BMI_Female$BMI_MISS,test_BMI_Female$pred, dnn=c("observed",
 ↪  "predicted") ) #full confusion matrix

####BMI TREE - MALES####
data_Male <- data_Full[which(data_Full$SEX=="Male"),]
rm(list=ls()[ls()!="data_Male"])


##Split data into training and testing datasets (80/20)##
set.seed(145)
ind_BMI_M<- rbinom(nrow(data_Male), size=1, prob=0.8)

train_BMIID_M <- (1:nrow(data_Male))[as.logical(ind_BMI_M)]
test_BMIID_M <- (1:nrow(data_Male))[!as.logical(ind_BMI_M)]

train_BMI_Male <- data_Male[train_BMIID_M,]
test_BMI_Male <- data_Male[test_BMIID_M,]


##Build the initial tree##
tree_BMI_Males <- rpart(
  BMI_MISS ~ .,
  data = train_BMI_Male,
  method = "class",
```

```
  minbucket=16, #min number of male students in the smallest school
  cp=-1,
)


##Perform cross-validation##
set.seed(145)
plotcp(tree_BMI_Males)

min.val<- min(tree_BMI_Males$cptable[,'xerror'])
tree_BMI_Males$cptable[tree_BMI_Males$cptable[,'xerror'] == min.val,]
pick<- (tree_BMI_Males$cptable[,'xerror'] == min.val)
(    tree_BMI_Males$cptable<- cbind(tree_BMI_Males$cptable,
↪  pick=ifelse(pick,1,0))    )


##Prune the inital tree based on cross validation results##
tree_BMI_Males_pruned <- prune(tree_BMI_Males, cp =0.0017)
fancyRpartPlot(tree_BMI_Males_pruned)


##Compare accuracies between inital and pruned trees##
test_BMI_Male$pred <- predict(tree_BMI_Males,test_BMI_Male, type =
↪  "class")
base_accuracy_BMI_Males <- mean(test_BMI_Male$pred ==
↪  test_BMI_Male$BMI_MISS)

test_BMI_Male$pred <- predict(tree_BMI_Males_pruned,test_BMI_Male, type =
↪  "class")
pruned_accuracy_BMI_Males <- mean(test_BMI_Male$pred ==
↪  test_BMI_Male$BMI_MISS)

base_accuracy_BMI_Males
pruned_accuracy_BMI_Males

table(test_BMI_Male$BMI_MISS,test_BMI_Male$pred, dnn=c("observed",
↪  "predicted") ) #full confusion matrix
```

Appendix D

Imputation Details and Diagnostics

This Appendix provides additional methodological details about the multiple imputation procedure from Study 3, which could not be covered within the word limits of a submitted manuscript. R Code to demonstrate the multiple imputation procedure for the results presented in Study 3 can also be found here.

D.1 Predictor Matrices

A component of the Multiple Imputation by Chained Equations (MICE) procedure is the creation of the predictor matrix, which identifies the variables to be used in the imputation procedure. Auxiliary variable selection, which identifies the variables to be used in the imputation model, involved directly using the variables identified in the CART models from Study 2. That is, any variable from the pruned CART models for BMI, height, or weight were considered to be useful auxiliary variables in their corresponding female or male imputation model. However, there are many more decisions related to the predictor matrix that need to be considered. By default, each variable is involved in the imputation of every other variable. However there are situations where this needed to be overridden; for example, it would not make sense to use an individual's scan id (their unique identification key) to impute any variables. Moreover, in this study multi-level imputation was used, which requires alteration of the of matrix values.

Tables D.1 and D.2 show the predictor matrices used for the female and male imputation models, respectively. The rows of the predictor matrix indicate the variables to *be imputed*, while the columns indicate which variables should be *used for imputation*. In single level imputation, only 0s and 1s are allowed in the matrix. A zero indicates no effect, while a 1 indicates an effect. Multi-level imputation adds 2s and -2s to the matrix, where -2 is used to identify the cluster variable (in this study, school), and a 2 indicates a random effect. Recommendations when implementing multi-level MICE are that random effects are specified for all variables [197], however that was computationally not feasible in this sample, so the where necessary several variables were reduced to only fixed effects.

Table D.1: Predictor Matrix for Multiple Imputation Procedure (COMPASS 2018/19 Female Sample

	scanID	SCHOOLID	AGE	ETHNICITY	BINGE	SMOKING	ECIG	CANNABIS	FAST_FOOD	STRENGTH	BREAKFAST	FRIEND_PA	WEIGHT_PERCEP	SPORTS	PA_HOURS	CONCEPT	SELF_MH	ENGLISH	STSB	WELLBEING	SLEEP	ANXIETY	DEPRESSION	HEIGHT	WEIGHT	BMI
scanID	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCHOOLID	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGE	0	-2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ETHNICITY	0	-2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BINGE	0	-2	1	1	0	2	2	2	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
SMOKING	0	-2	1	1	1	0	1	1	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
ECIG	0	-2	1	1	2	2	0	2	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
CANNABIS	0	-2	1	1	1	1	1	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1
FAST_FOOD	0	-2	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
STRENGTH	0	-2	1	1	0	0	0	0	1	0	1	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1
BREAKFAST	0	-2	1	1	0	0	0	0	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
FRIEND_PA	0	-2	1	1	0	0	0	0	1	2	1	0	2	2	2	1	1	1	1	1	1	1	1	1	1	1
WEIGHT_PERCEP	0	-2	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
SPORTS	0	-2	1	1	1	1	1	1	1	1	1	2	1	0	1	1	1	1	1	1	1	1	1	1	1	1
PA_HOURS	0	-2	1	1	1	1	1	1	1	2	1	2	1	2	0	1	1	1	1	1	1	1	1	1	1	1
CONCEPT	0	-2	1	1	2	2	2	2	0	1	0	1	2	2	1	0	1	1	1	1	1	1	1	1	1	1
SELF_MH	0	-2	1	2	2	2	2	2	0	1	0	1	2	2	1	1	0	1	1	1	1	1	1	1	1	1
ENGLISH	0	-2	1	1	1	1	1	1	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1
STSB	0	-2	1	1	2	2	2	2	0	2	0	2	2	2	2	2	2	2	0	2	2	2	2	2	2	2
WELLBEING	0	-2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2	2	2
SLEEP	0	-2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2	2
ANXIETY	0	-2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2
DEPRESSION	0	-2	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2
HEIGHT	0	-2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	0
WEIGHT	0	-2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0
BMI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table D.2: Predictor Matrix for Multiple Imputation Procedure (COMPASS 2018/19 Male Sample

	scanID	SCHOOLID	AGE	BINGE	ETHNICITY	ECIG	WEIGHT_PERCEP	CANNABIS	FAST_FOOD	BREAKFAST	PA_HOURS	SPORTS	MILK	GRAIN	FRUIT_VEG	MEAT	CONCEPT	STSB	ANXIETY	SLEEP	DEPRESSION	HEIGHT	WEIGHT	BMI
scanID	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCHOOLID	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGE	0	-2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
BINGE	0	-2	1	0	1	2	0	2	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1
ETHNICITY	0	-2	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ECIG	0	-2	1	2	1	0	0	2	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1
WEIGHT_PERCEP	0	-2	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
CANNABIS	0	-2	1	1	1	1	0	0	0	0	0	1	0	0	0	0	1	1	1	1	1	1	1	1
FAST_FOOD	0	-2	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1
BREAKFAST	0	-2	1	0	1	0	1	0	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1
PA_HOURS	0	-2	1	1	1	1	1	1	1	1	0	2	1	1	1	1	1	1	1	1	1	1	1	1
SPORTS	0	-2	1	1	1	1	1	1	1	1	2	0	1	1	1	1	1	1	1	1	1	1	1	1
MILK	0	-2	1	0	1	0	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1
GRAIN	0	-2	1	0	1	0	1	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
FRUIT_VEG	0	-2	1	0	1	0	1	0	1	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1
MEAT	0	-2	1	0	1	0	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1
CONCEPT	0	-2	1	2	1	2	2	2	0	0	1	2	1	1	1	1	0	1	1	1	1	1	1	1
STSB	0	-2	1	2	1	2	2	2	0	0	2	2	0	0	0	0	2	0	2	2	2	2	2	2
ANXIETY	0	-2	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2	2
SLEEP	0	-2	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2	2
DEPRESSION	0	-2	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	2	2
HEIGHT	0	-2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	2	0
WEIGHT	0	-2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	0	0
BMI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

D.2 Imputation Model Diagnostics

Section 7.4.2 of Chapter 7 highlights that a disadvantage to imputation models is that there is no way to concretely establish the validity of an imputation model. As such, performing thorough diagnostics on an imputation model is essential. The imputation model checking procedures used in Study 3 are described in this section.

D.2.1 Convergence

Examining convergence is necessary to gauge whether an appropriate number of iterations have been used to create the multiply imputed datasets. Convergence plots should sufficiently intermingle and not follow any particular pattern by the later iterations. Examination of convergence plots resulted in the iterations being increased to 10 (from the default of 5) in final models. Convergence plots for the imputation models used in Study 3 are available in Figures D.1 through D.8 for females, and Figures D.9 through D.16 for males.

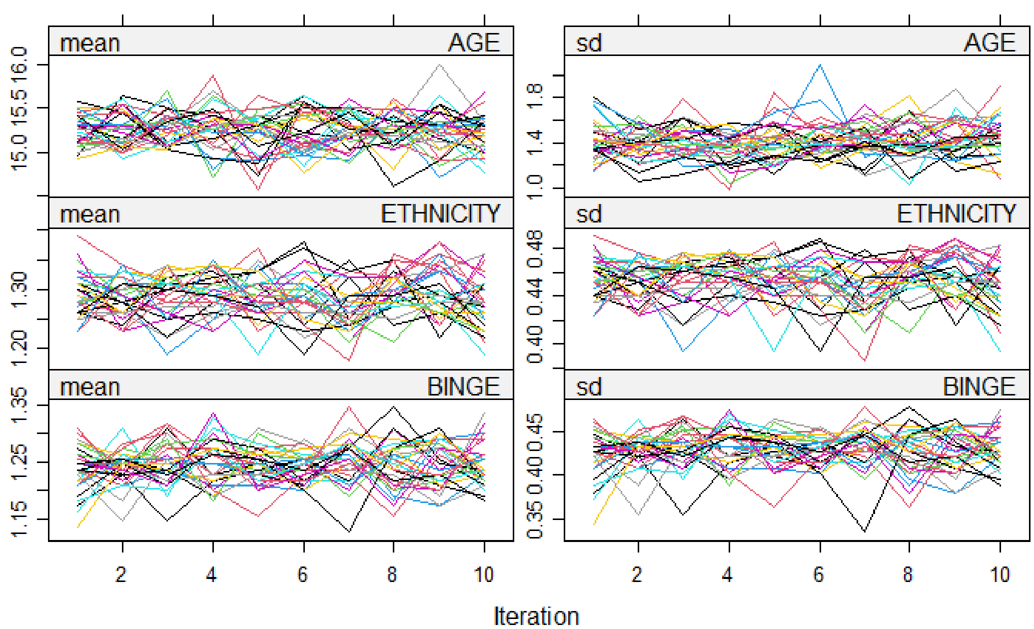


Figure D.1: Imputation Convergence Plots for Females (1 of 8)

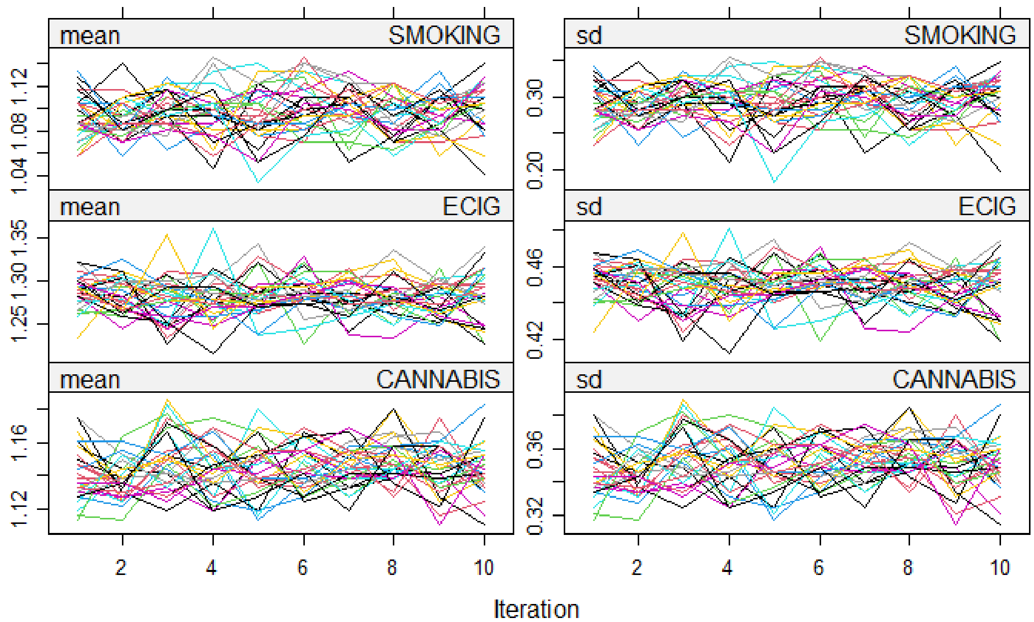


Figure D.2: Imputation Convergence Plots for Females (2 of 8)

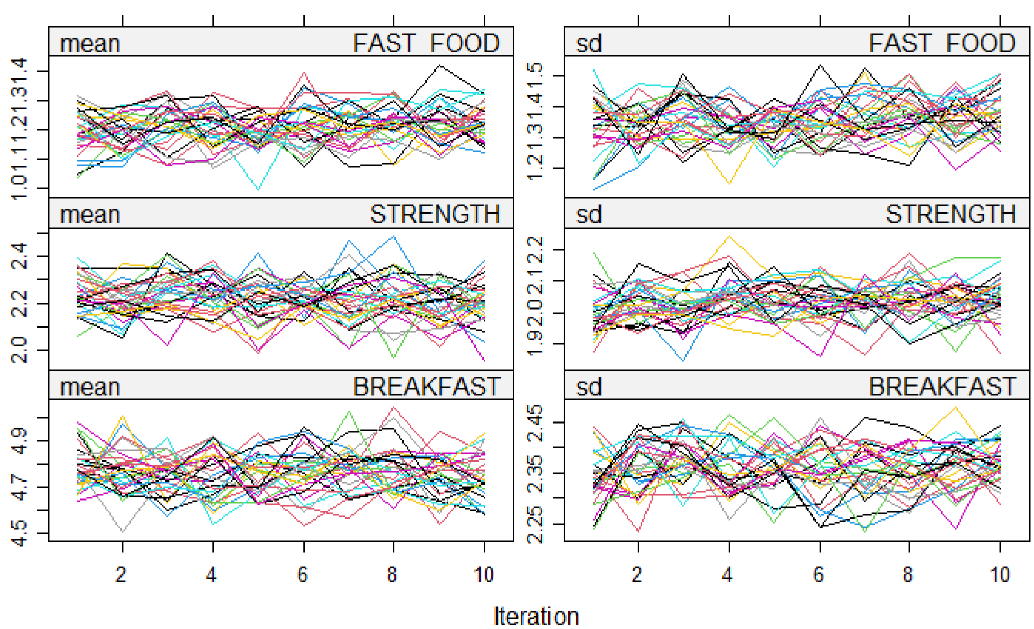


Figure D.3: Imputation Convergence Plots for Females (3 of 8)

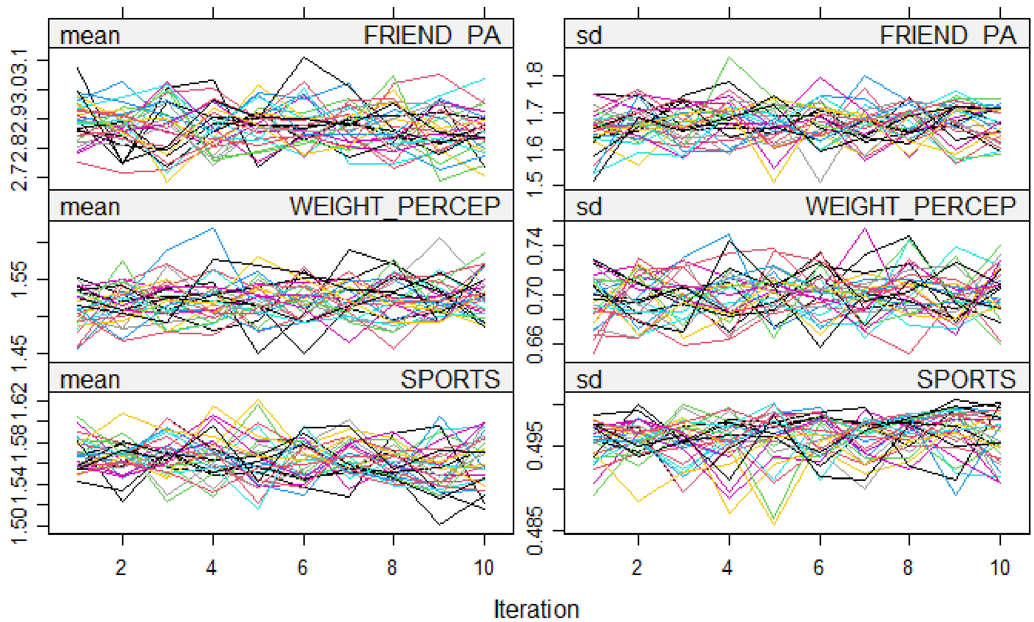


Figure D.4: Imputation Convergence Plots for Females (4 of 8)

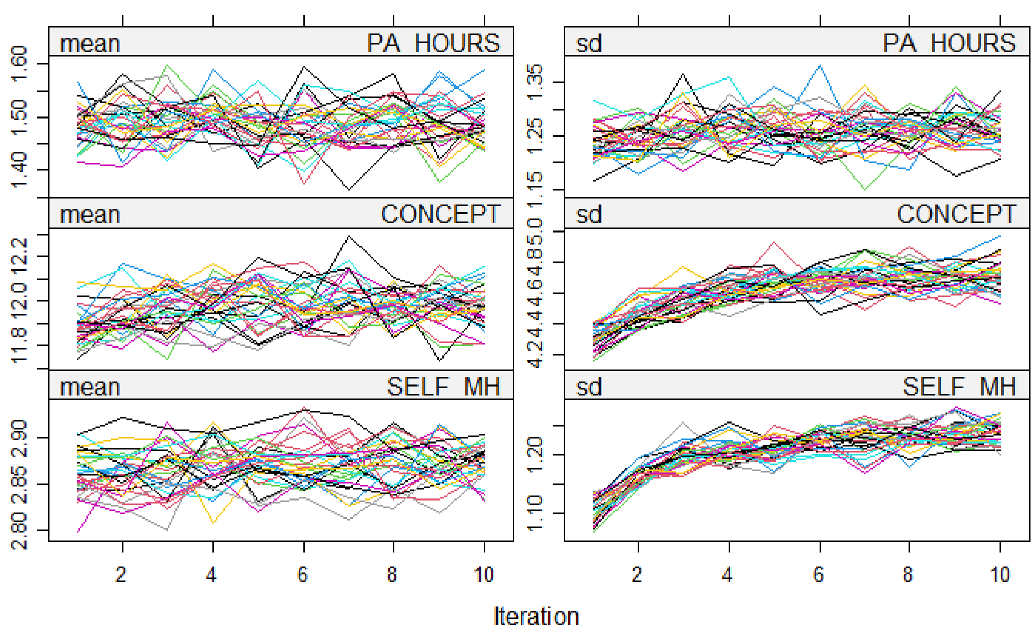


Figure D.5: Imputation Convergence Plots for Females (5 of 8)

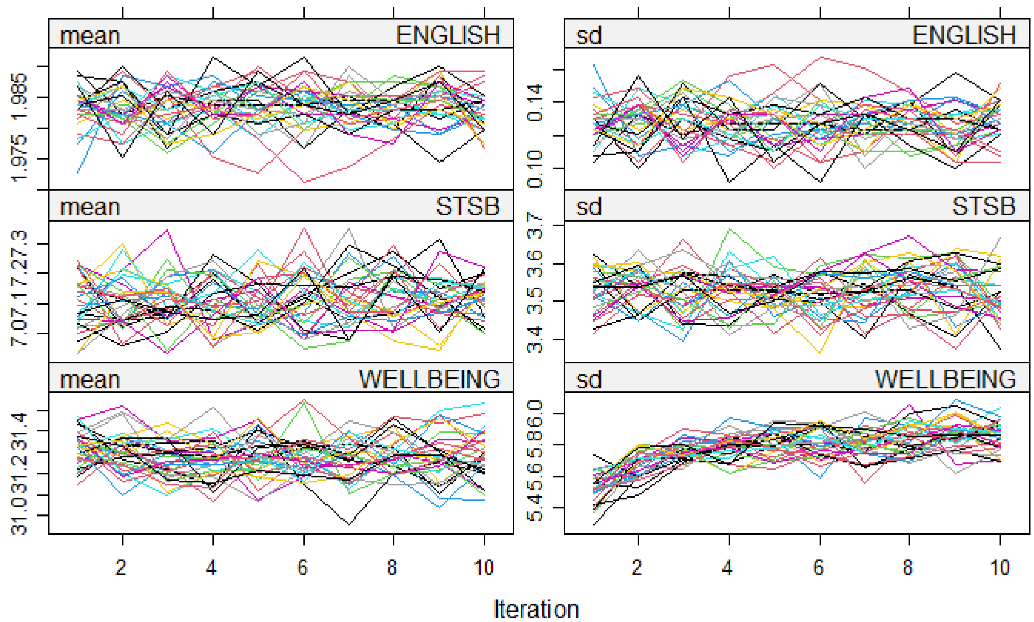


Figure D.6: Imputation Convergence Plots for Females (6 of 8)

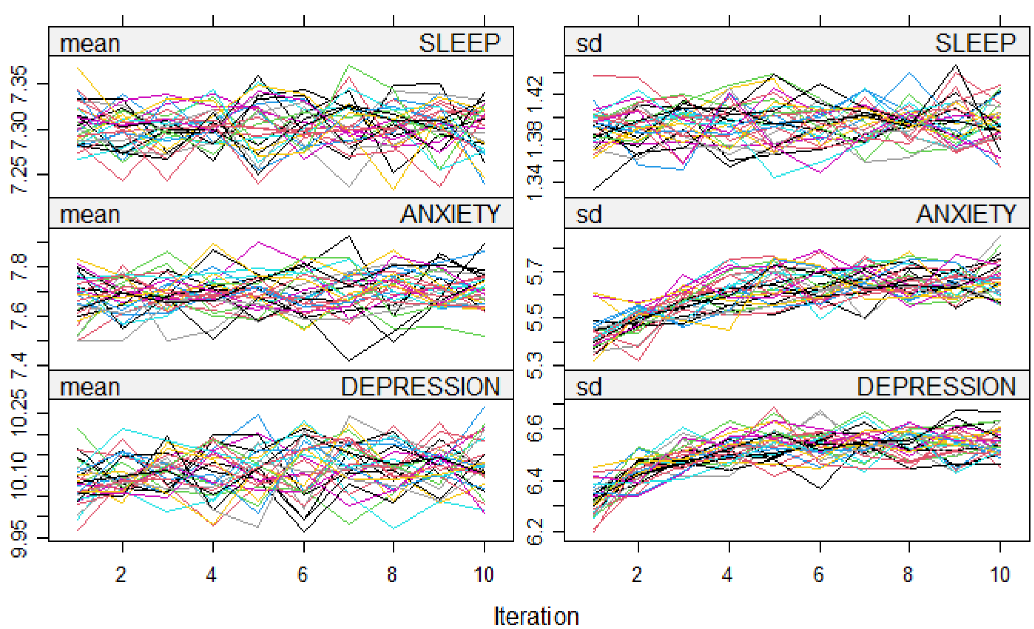


Figure D.7: Imputation Convergence Plots for Females (7 of 8)

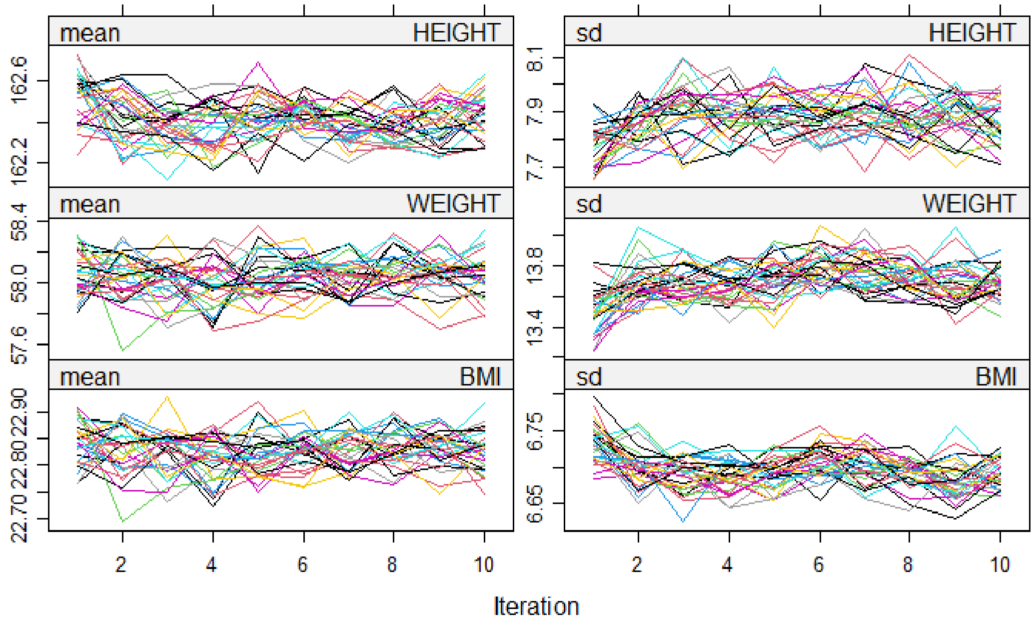


Figure D.8: Imputation Convergence Plots for Females (8 of 8)

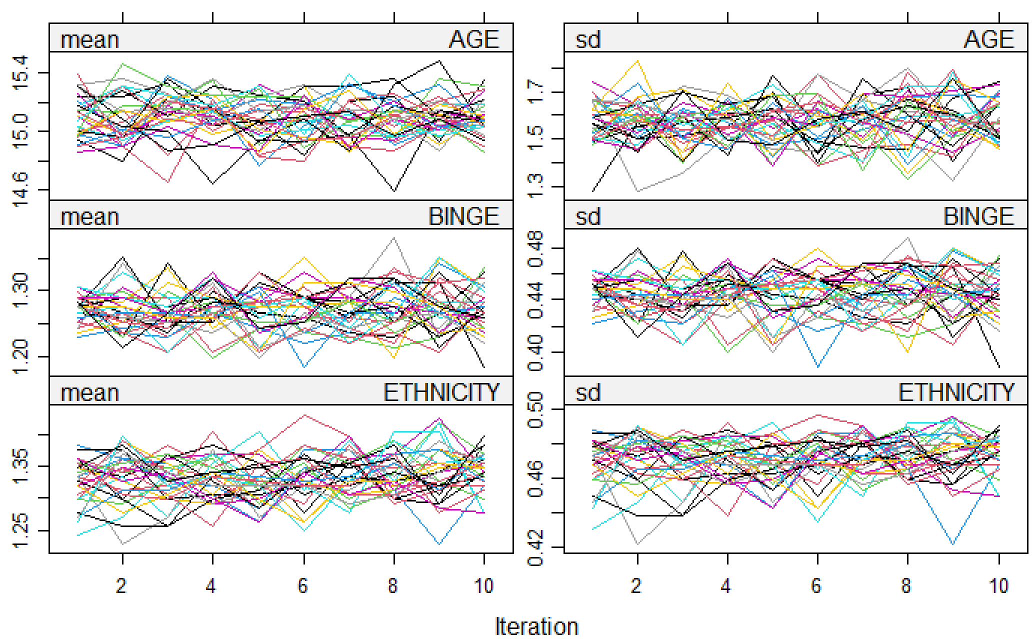


Figure D.9: Imputation Convergence Plots for Males (1 of 8)

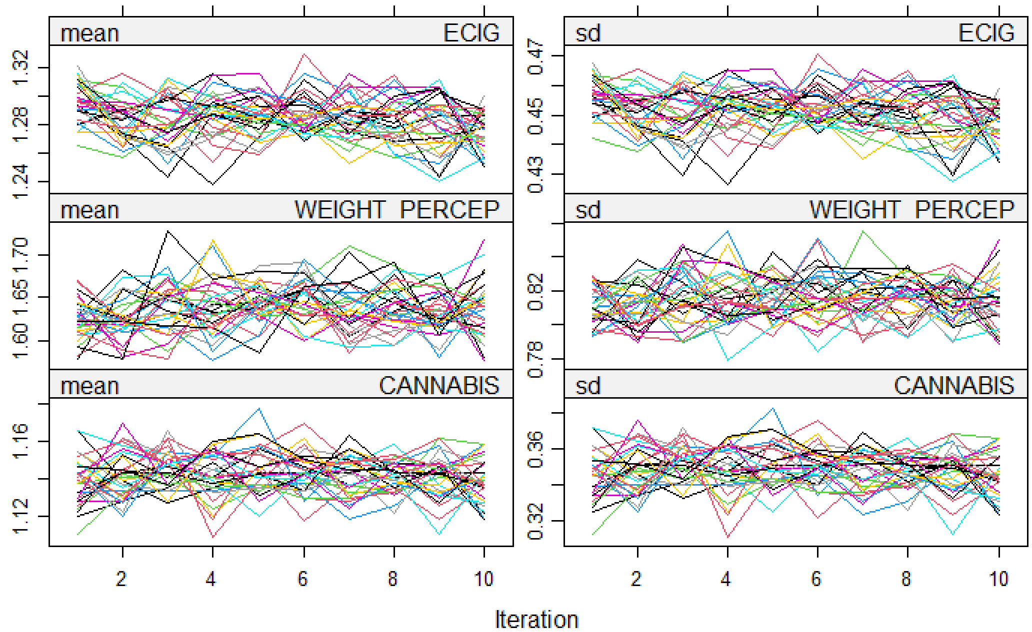


Figure D.10: Imputation Convergence Plots for Males (2 of 8)

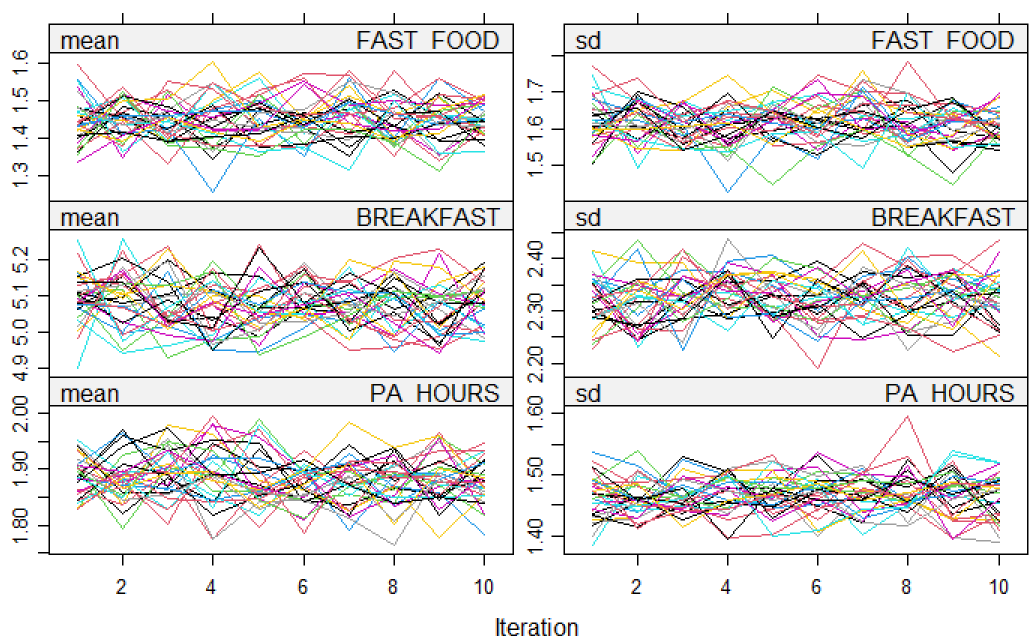


Figure D.11: Imputation Convergence Plots for Males (3 of 8)

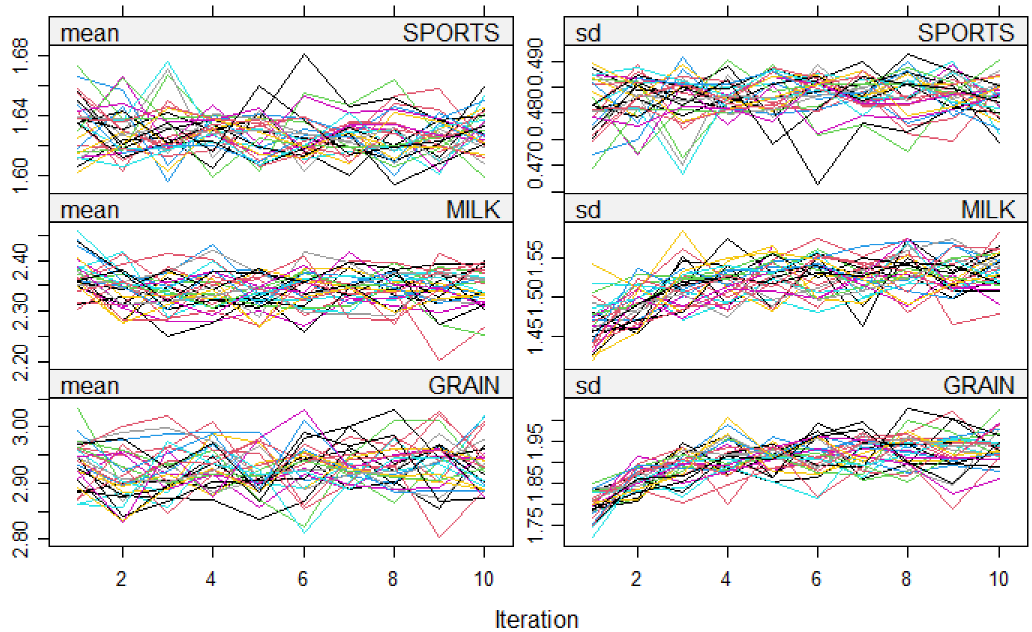


Figure D.12: Imputation Convergence Plots for Males (4 of 8)

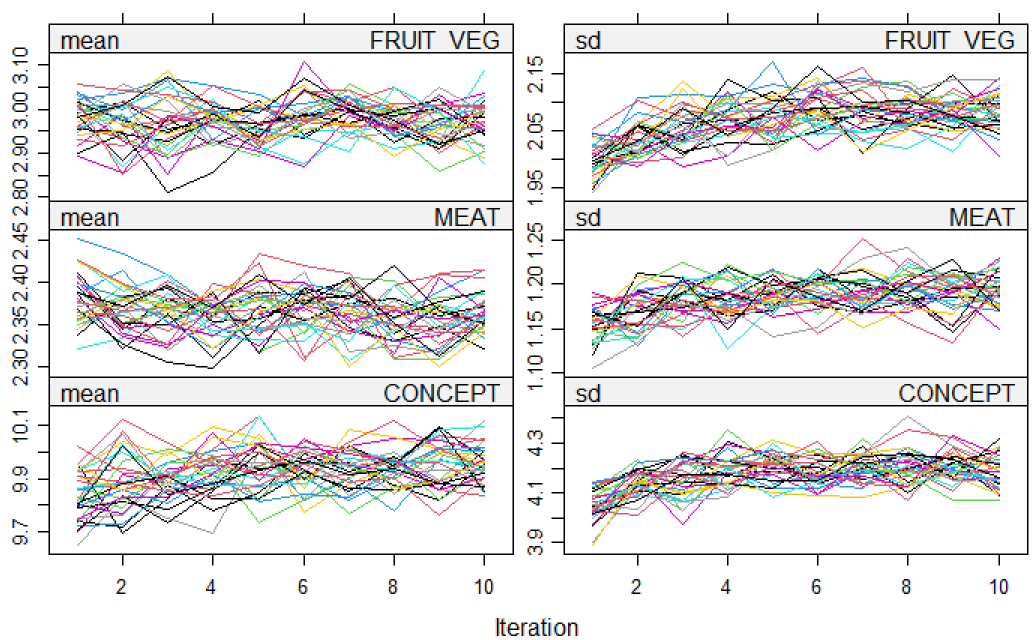


Figure D.13: Imputation Convergence Plots for Males (5 of 8)

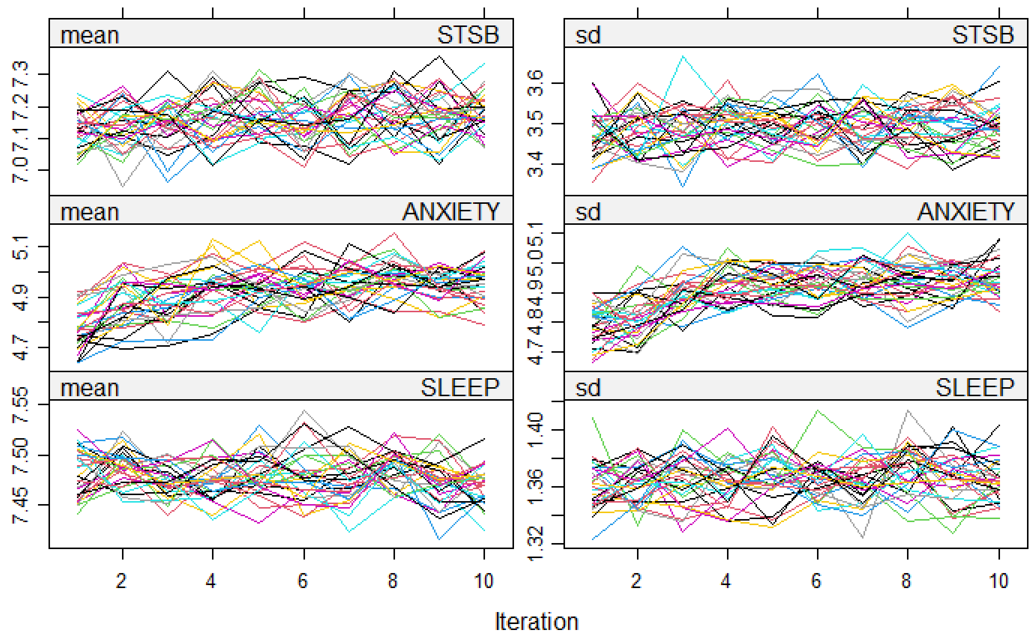


Figure D.14: Imputation Convergence Plots for Males (6 of 8)

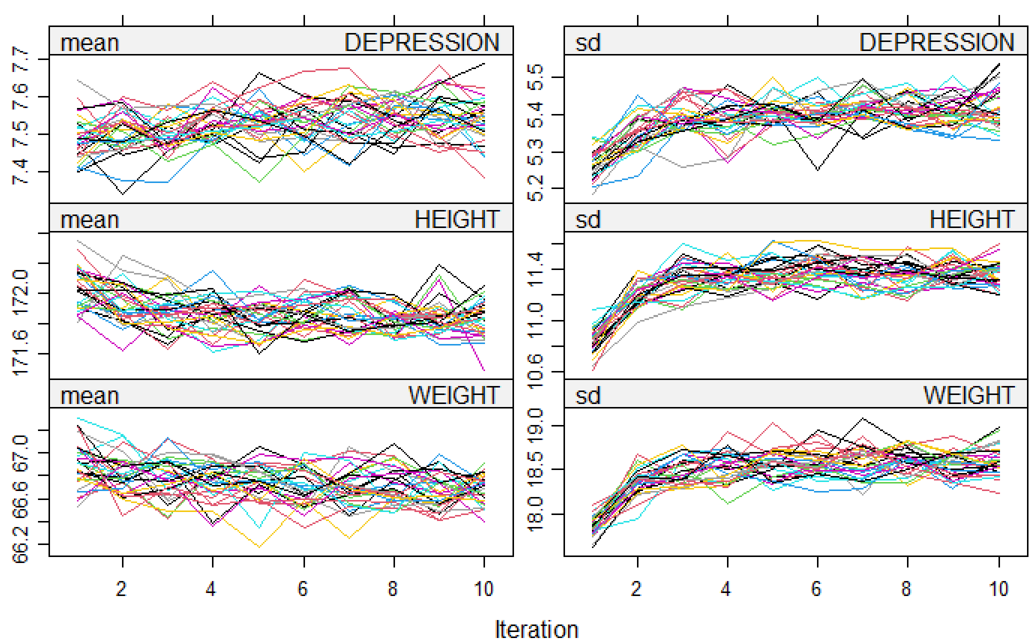


Figure D.15: Imputation Convergence Plots for Males (7 of 8)

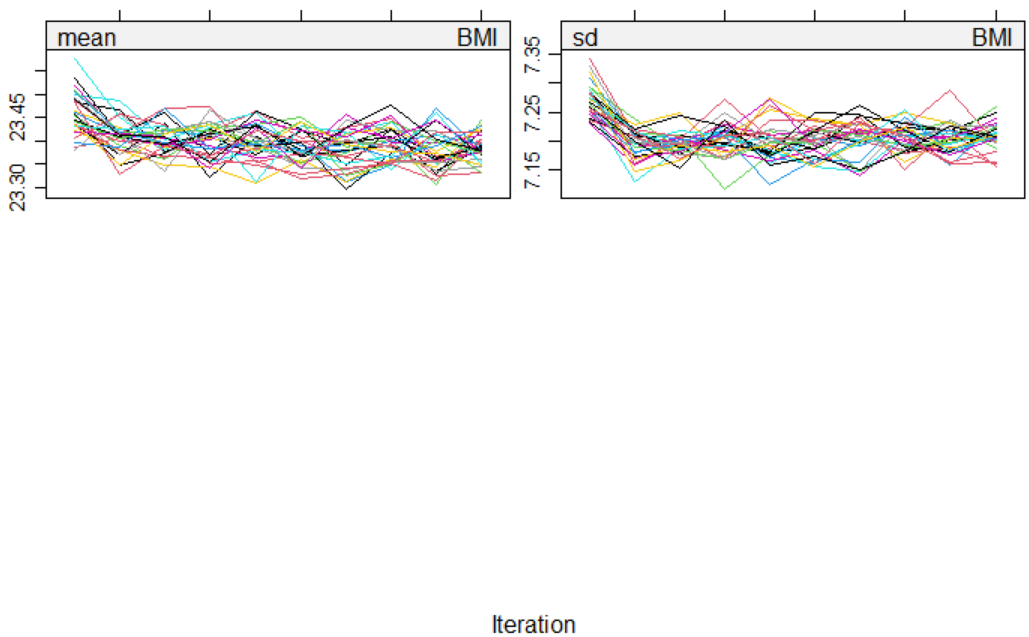


Figure D.16: Imputation Convergence Plots for Males (8 of 8)

D.2.2 Comparing Observed and Imputed Values

In additional to examining convergence, imputation diagnostics focus on comparing imputed values to real values. Of course, comparing *individual* imputations to real values isn't necessarily appropriate, since an individual imputation can be unreasonable. However, the *distribution* of observed and imputed variables should be similar in a properly constructed MI model if data are missing at random. In Study 3, imputed values were compared to real values through the use of density plots and summary statistics for all variables; the imputations were considered acceptable if they followed the same approximate distribution of observed values. BMI was the exception to this, as distributions showed that MI skewed the data towards higher values, however this was to be expected considering that BMI was thought to likely follow an NMAR pattern of missingness.

D.3 R Code for Imputation Procedure

Provided in this section is the R code used to implement the imputation procedure from Study 3. Only the code for the imputation of the female model is provided as the male imputation code is highly similar. In practice, implementing an MI procedure is highly specific to the data, so I don't recommend naively copy-pasting this code. However, this dissertation advocates for better missing data handling in research, so not providing resources that could potentially be useful to others would be mismatched with the recommendations I made throughout this document. My recommendation to applied researchers who wish to implement a multiple imputation procedure in R is to start by reviewing key resources. I personally recommend Craig Enders' *Applied Missing Data Analysis* [15] and Paul Allison's *Missing Data* [13] to be very useful and accessible resources to understand missing data. In terms of implementation, I found that Stef van Buuren's *Flexible Imputation of Missing Data* [8], and Vink & van Buuren's vignettes on MICE [197] to be extremely valuable.

```
library(dplyr)
####Clean data for imputation####
#Remove unneeded variables from dataset
dropMost <- names(impDataFemale) %in% c("HEIGHT", "WEIGHT","BMI", "AGE",
↪  "scanID", "SCHOOLID", "ETHNICITY",
                                        "SPORTS", "PA_HOURS", "STSB",
                                        ↪  "SLEEP",
                                        "BREAKFAST", "FAST_FOOD",
                                        "ANXIETY", "DEPRESSION",
                                        ↪  "CONCEPT",
                                        "BINGE", "CANNABIS", "ECIG",

                                        "WEIGHT_PERCEP", "STRENGTH",
                                        ↪  "FRIEND_PA", "ENGLISH",
```

```
                                              "WELLBEING", "SELF_MH",
                                              ↪  "SMOKING")
impDataFemaleUnordered <- impDataFemale[dropMost]
rm(list=ls()[ls()!="impDataFemaleUnordered"])

#Specify the visit sequence by rearranging the data - recall that
↪  default is that FCS proceeds left to right
##Reorder actual data instead of using visitSequence command because
↪  visitSequence does not reorder the matrix to be exported##
col_order <- c("scanID","SCHOOLID","AGE", "ETHNICITY", "BINGE",
                "SMOKING", "ECIG",
                "CANNABIS", "FAST_FOOD", "STRENGTH", "BREAKFAST",
                "FRIEND_PA", "WEIGHT_PERCEP", "SPORTS",
                "PA_HOURS", "CONCEPT", "SELF_MH",

                "ENGLISH", "STSB", "WELLBEING", "SLEEP",
                "ANXIETY", "DEPRESSION", "HEIGHT", "WEIGHT", "BMI")

impDataFemale <- impDataFemaleUnordered[, col_order]
rm(col_order)
rm(impDataFemaleUnordered)


#### Imputation Procedure ####
library(mice)
library(pan)
library(miceadds)
library(lme4)
library(broom.mixed)

#Run an inital imputation with no iterations to get matrix and methods
↪  objects
impIni <- mice(impDataFemale,print=FALSE, seed =145, maxit=0 )

##Only need to run this commented chunk initially, afterwards it is
↪  easier to edit and important from excel##
# updateMatrix <- impIni$predictorMatrix
#
# #based on how passive imputation works, this is redundant but good
↪  practice
# updateMatrix["BMI", ] <- 0
#
# #scanID shouldnt be imputed nor be used to impute anything
# updateMatrix["scanID", ] <- 0
# updateMatrix[, "scanID"] <- 0
#
# #SCHOOLID shouldnt be imputed
```

```
# updateMatrix["SCHOOLID", ] <- 0
#
# #don't want(well, actually cannot) to use BMI to impute height or
↪   weight
# updateMatrix[c("WEIGHT", "HEIGHT"),"BMI" ] <- 0

#Export predictor matrix to excel for easier editing
#write.csv(updateMatrix, file="filepath/iniMatrix.csv")

#Import matrix back into R
impMatrixRandom <- read.csv("filepath/matrix.csv", row.names =1 )
impMatrixRandom <- data.matrix(impMatrixRandom)
impMatrixRandom <- matrix(as.numeric(impMatrixRandom),
↪   ncol=ncol(impMatrixRandom), dimnames=(dimnames(impMatrixRandom)))

## Updating the Imputation Methods for each variable ##
updateMethod <- impIni$method
updateMethod["BMI"] <- "~I(WEIGHT/(HEIGHT/100)^2)"
updateMethod["WEIGHT"] <- "2l.pan"
updateMethod["HEIGHT"] <- "2l.pan"

updateMethod["AGE"] <- "pmm"
updateMethod["ETHNICITY"] <- "pmm"

updateMethod["SPORTS"] <- "2l.pmm"
updateMethod["PA_HOURS"] <- "2l.pan"
updateMethod["STSB"] <- "2l.pan"
updateMethod["SLEEP"] <- "2l.pan"

updateMethod["BREAKFAST"] <- "2l.pmm"
updateMethod["FAST_FOOD"] <- "2l.pmm"

updateMethod["ANXIETY"] <- "2l.pan"
updateMethod["DEPRESSION"] <- "2l.pan"
updateMethod["CONCEPT"] <- "2l.pan"

updateMethod["BINGE"] <- "2l.pmm"
updateMethod["CANNABIS"] <- "2l.pmm"
updateMethod["ECIG"] <- "2l.pmm"

updateMethod["WEIGHT_PERCEP"] <- "2l.pmm"
updateMethod["STRENGTH"] <- "2l.pan"
updateMethod["FRIEND_PA"] <- "2l.pan"
updateMethod["ENGLISH"] <- "2l.pmm"

updateMethod["WELLBEING"] <- "2l.pan"
updateMethod["SELF_MH"] <- "2l.pan"
```

```
updateMethod["SMOKING"] <- "2l.pmm"

#Convert cluster (documentation calls this 'class') variable to an
↪  integer (required for multi-level mice fx)
impDataFemale$SCHOOLID <- as.integer(impDataFemale$SCHOOLID)

## Run imputation ##
#NB: if wanting to replicate via seed, it is insufficient to only
↪  specify seed argument in mice() fx...its unclear why this does not
↪  work
#instead need to globally specify set.seed() before run of the
↪  imputation
set.seed(145)
imp.object30it10 <- mice(impDataFemale,m=30,
↪  maxit=10,method=updateMethod, predictorMatrix = impMatrixRandom, seed
↪  = 145)


#### Imputation Model Checking ####
#Check to see that height, weight, and BMI correspond appropriately
head(complete(imp.object30it10)[is.na(impData$BMI), ], 3)

#Check the convergence plots
plot(imp.object30it10)

#Check the distribution of real vs. imputed data
densityplot(imp.object30it10)

#Compare the distributions of observed and imputed data conditional on
↪  propensity score
fit <- with(imp.object30it10, glm(ici(imp.object30it10) ~ AGE + BMI +
↪  WEIGHT + HEIGHT,
                             family = binomial))
ps <- rep(rowMeans(sapply(fit$analyses, fitted.values)),
          imp.object30it10$m + 1)

xyplot(imp.object30it10, BMI ~ ps | as.factor(.imp),
       xlab = "Probability that record is incomplete",
       ylab = "BMI", pch = c(1, 19), col = mdc(1:2))


#Filter mids object by parameters (e.g. age) to examine data subsets
midsAge12 <- filter(mi.res, AGE==12)

#stripplot() and bwplot() fxs are also useful, but less so for large
↪  datasets
```

```
####Export imputations to SAS if needed####
# library(foreign)
# impExport <- complete(imp.object30it10, "long")
# write.foreign(impDataExportJuly27,datafile="filepath/impExport.sas",
#               codefile="filepath/impExportCODE.sas", package="SAS")

####Analyses - Imputed Data####
#Fit a random intercept LMM
fitModel <- with(imp.object30it10, lmer(BMI ~ AGE + ETHNICITY +
                                  SPORTS + PA_HOURS + STSB + SLEEP +
                                  FAST_FOOD + BREAKFAST+
                                  ANXIETY + DEPRESSION + CONCEPT +
                                  BINGE + CANNABIS + ECIG + (1 |
                                  ↪  SCHOOLID)))

#Pool estimates (pool() default is RR combination)
pooledModel <- pool(fitModel)

#Sensitivity analysis with GEE - are SEs better?
fitModelGEE <- with(imp.object30it10, geeglm(BMI ~ AGE + ETHNICITY +
                                           SPORTS + PA_HOURS + STSB +
                                           ↪  SLEEP +
                                           FAST_FOOD + BREAKFAST +
                                           ANXIETY + DEPRESSION +
                                           ↪  CONCEPT +
                                           BINGE + CANNABIS + ECIG ,
                                           ↪  family=gaussian,
                                           ↪  id=SCHOOLID, corstr =
                                           ↪  "exchangeable"))



####Analysis - CCA ####
#Remove unneeded variables (i.e. remove auxilliary vars)
dropAnalysis <- names(impDataFemale) %in% c("HEIGHT", "WEIGHT","BMI",
↪  "AGE", "scanID", "SCHOOLID", "ETHNICITY",
                                        "SPORTS", "PA_HOURS", "STSB",
                                        ↪  "SLEEP",
                                        "FAST_FOOD","BREAKFAST",
                                        "ANXIETY", "DEPRESSION",
                                        ↪  "CONCEPT",
                                        "BINGE", "CANNABIS", "ECIG")
analysisSubset <- impDataFemale[dropAnalysis]

#remove all missingness to create complete case sample
ccaData <- na.omit(analysisSubset)

#Fit a random intercept LMM
```

```r
ccaModel <- lmer(BMI ~ AGE + ETHNICITY +
                 SPORTS + PA_HOURS + STSB + SLEEP +
                 FAST_FOOD + BREAKFAST+
                 ANXIETY + DEPRESSION + CONCEPT +
                 BINGE + CANNABIS + ECIG + (1 | SCHOOLID), data=ccaData)

#Sensitivity analysis with GEE - are SEs better?
ccaModelGEE <- geeglm(BMI ~ AGE + ETHNICITY +
                        SPORTS + PA_HOURS + STSB + SLEEP +
                        FAST_FOOD + BREAKFAST +
                        ANXIETY + DEPRESSION + CONCEPT +
                        BINGE + CANNABIS + ECIG , family=gaussian,
                        ↪  id=SCHOOLID, corstr = "exchangeable",
                        ↪  data=ccaData)
```