https://doi.org/10.24095/hpcdp.43.5.03

# Original quantitative research

# Using classification and regression trees to model missingness in youth BMI, height and body mass data

*Amanda Doggett, PhD (1); Ashok Chaurasia, PhD (1); Jean-Philippe Chaput, PhD (2,3); Scott T. Leatherdale, PhD (1)*

This article has been peer reviewed.



## Abstract

**Introduction:** Research suggests that there is often a high degree of missingness in youth body mass index (BMI) data derived from self-reported measures, which may have a large effect on research findings. The first step in handling missing data is to examine the levels and patterns of missingness. However, previous studies examining youth BMI missingness used logistic regression, which is limited in its ability to discern subgroups or identify a hierarchy of importance for variables, aspects that may go a long way in helping understand missing data patterns.

**Methods:** This study used sex-stratified classification and regression tree (CART) models to examine missingness in height, body mass and BMI data among 74 501 youth participating in the 2018/19 COMPASS study (a prospective cohort study examining health behaviours among Canadian youth), where 31% of BMI data were missing. Diet, movement, academic, mental health and substance use variables were examined for associations with missingness in height, body mass and BMI.

**Results:** CART models indicated that the combination of being younger, having a self-perception of being overweight, being less physically active and having poorer mental health yielded female and male subgroups highly likely to be missing BMI values. Survey respondents who did not perceive themselves as overweight and who were older were unlikely to be missing BMI values.

**Conclusion:** The subgroups identified by the CART models indicate that a sample that deletes cases with missing BMI would be biased towards physically, emotionally and mentally healthier youth. Given the ability of CART models to identify these subgroups and a hierarchy of variable importance, they are an invaluable tool for examining missing data patterns and appropriate handling of missing data.

***Keywords:*** *missing data, decision trees, overweight, obesity, adolescents*

### *Highlights*

- Almost one-third (31%) of the 74 501 youth participating in the COMPASS study in 2018/19 were missing body mass index (BMI) values.
- Missing weight values were more prevalent among female youth than among male youth.
- Social desirability likely plays a large role in youth not reporting their height and weight.
- Classification and regression tree models are useful in identifying important subgroups with missing data.

## Introduction

### *Missing data in overweight and obesity literature*

As one of the strongest predictors of chronic diseases,[1] overweight and obesity (OWOB) remains one of the top health concerns globally. Many studies that examine OWOB use body mass index (BMI) derived from self-reported measures of height and body mass to provide a proxy measure of body adiposity. Self-reported measures are usually less accurate than direct anthropomorphic measurements—individuals tend to underreport their body mass and overreport their height[2-5]—but self-reporting is generally more feasible (logistically and financially) than other approaches to population surveillance,[3-5] and these measures are useful 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, large proportions (sometimes over 50%) of self-reported height and body mass data tend to be missing.[6,7] If data are missing completely at random (MCAR), the probability of missingness depends neither on the hypothetical true value of the missing variable (i.e. what the value would be if it *was* reported), nor on any observed covariates. But if data are missing at random (MAR) or not missing at random (NMAR), the probability of missingness depends on observed covariates (for missing at random) and/or the hypothetical true value of the missing variable

**Author references:**

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

**Correspondence:** Amanda Doggett, School of Public Health Sciences, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1; Tel: 519-888-4567; Email: adoggett@uwaterloo.ca

(for not missing at random). Deleting these missing cases (a method called complete case analysis) is a problematic approach, particularly for the last two mechanisms, as it leads to statistical bias.[8] For example, if data are missing at random because younger youth are more likely to neglect reporting their weight, the sample is biased towards older individuals (and then, logically, heavier ones, given child growth patterns).

This introduction of statistical bias as a result of deleting cases has also been proven through numerous simulation studies; it is particularly prominent when there is a large proportion of non-random missingness.[8,9] Despite this, complete case analysis remains the most common approach in epidemiological literature.[10,11] The high degree of missingness in youth self-reported height and body mass data raises concerns about how methods take into account missing data and how mishandling of missing data affects research findings as well as concomitant policy and programming recommendations.

Statistical approaches are often required to deal with missing data; while researchers should follow best practices in survey design, in many cases there may be little they can do to improve reporting patterns.[12,13] Although sophisticated statistical approaches to handling large proportions of non-random missingness are available, they generally require more time and expertise, which may be a barrier to their overall use. 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 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 that missingness is generally unique to studies, there is no clear framework for the process for identifying sources or mechanisms of missingness.

### *Regression approaches*

Research examining BMI or body mass missingness has used regression approaches[6,7,14] where the outcome of a logistic regression is missing versus not missing, and other variables are examined for their potential association with the likelihood of missingness. However, regression approaches may not be ideal in this situation 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. When building a regression model, an initial step to selecting variables might be to review the literature for similar analyses, but the literature in the context of examining BMI missingness is scarce.

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 the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), can be used to select variables, but these procedures can be challenging in practice: we previously examined BMI, height and body mass missingness using model selection procedures for generalized linear mixed models,[15] but this required many additional modelling decisions and a customized algorithm suitable for pseudo-likelihood methods.[16]

Lastly, 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 a hierarchy of the importance of variables would be beneficial: it may help with parsimony and clearer interpretation, and it may be a necessary step to employ certain missing data approaches like multiple imputation. Although our previous study added to the literature on missingness in youth BMI, we were unable to identify which variables were most important or identify which combinations of factors were most likely to lead to nonreporting.[15] The limitations associated with a regression approach to examining missing data may be addressed by using a different methodological approach.

### *Decision trees*

Decision trees are a type of machine-learning approach that has been leveraged in applied research, including in public health.[17,18] Decision trees are useful for analyzing primary data and for examining missing data; they 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 some inference on variable importance. Moreover, decision trees allow important and highly specific subgroups to be identified beyond what would be feasible using interaction terms in a regression model.

In addition, unlike regression, the entire decision tree model can be easily visualized, which may help interpretation. In 2015, Tierney et al.[19] published work demonstrating the utility of using decision trees to examine missing data, but to our knowledge no published studies have leveraged this approach.

The purpose of this study is (1) to add to the limited literature on missing data in youth self-reported height and body mass; (2) to identify potential areas of bias stemming from nonreporting in the youth OWOB domain; and (3) to demonstrate the use of decision trees to model missing data, which builds on the work of Tierney et al.,[19] who first identified the utility of this approach.

## Methods

### *Sample*

This study uses a large cross-sectional dataset from the 2018/19 wave of the COMPASS (Cannabis, Obesity, Mental health, Physical activity, Alcohol, Smoking, Sedentary behaviour) study, a prospective cohort study that collects data on a variety of different health behaviours among youth. The 2018/19 COMPASS wave collected data from 74 501 youth, representing an 84.3% participation rate. COMPASS uses an active-information, passive-consent protocol that yields high participation rates, and non-participation is usually due to absence from school on the data collection day or being in a scheduled spare during the data collection time.

### *Variables*

This study focusses on missingness in BMI values 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 recorded based on responses to the question asked of 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 recorded in response to the question, "How tall are you without your shoes on? (Please write your height in feet and inches OR in centimetres, and then fill in the appropriate numbers for your height)." BMI is derived by dividing body mass (kg) by height squared ($m^2$).

A benefit of decision tree approaches is the feasibility of including many variables. In this study, we included a variety of diet, movement, academic, mental health and substance use variables. Diet-related variables included number of servings of fruits and vegetables, grain products, meat and alternatives, and milk and alternatives as well as number of days per week when breakfast, energy drinks and fast foods were consumed. Movement-related variables included moderate-to-vigorous physical activity, sports participation (inside or outside of school), strength training, physically active peers, screen time sedentary behaviour (STSB) and sleep.

Academic-related variables included English grade (or French grade, for French language schools), Math grade and truancy. Mental health variables included clinically relevant symptoms of depression (CESD-R-10 scale[20]), anxiety (GAD-7 scale[21]), socio-emotional skills (DERS scale[22]), self-reported well-being (Flourishing scale[23]), self-concept (Self Description Questionnaire II short form[24]), 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 use of alcohol mixed with energy drinks. Although all these variables were input into analyses, only a subset of variables appeared in the final models.

### Outliers

In some cases, missingness was imposed onto the data. We used the 1.5 × interquartile range (IQR) method to identify statistical outliers, and these cut-offs were considered alongside biological plausibility in order to determine how to handle these cases. We marked as missing weights less than 45 lbs (20 kg) or greater than 390 lbs (177 kg) and height less than 4' (1.22 m) or greater than 6'11" (2.11 m). Sleep and STSB were two variables that had a number of unfeasible outliers in the dataset. Youth who reported regularly sleeping less than 4 hours a night or having a collective STSB greater than 16.25 hours per day were marked as missing. Notably, missingness was only imposed for that particular variable; for example, those who reported less than 4 hours of sleep had their sleep value marked as missing, but all other reported variables remained the same.

### Analysis

We used classification and regression trees (CART) as the approach for this study where the outcome was binary (i.e. missing vs. not missing). All models were stratified by self-reported sex (female, male). Consistent with decision tree approaches,[25] 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.

We used cost complexity pruning alongside the one standard error (1-SE) rule[25] to help correct for overfitting and yield a more parsimonious final tree. Decision tree analyses were conducted in R (R Foundation for Statistical Computing, Vienna, AT) 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; this was 14 for females and 16 for males. Models included individuals with missing covariate data, as CART conveniently handles this by surrogate splitting; if a covariate value is missing, an observed variable with the most similar predictive capacity is used instead.

## Results

### Descriptive statistics

Table 1 shows stratified descriptive statistics for any variable that appeared in at least one of the CART models. Of the whole sample (n = 74 501), 31% were missing BMI values. Height missingness was slightly more prevalent among males (19%) than among females (15%), whereas body mass missingness was slightly more prevalent among females (22%) than among males (20%).

### Interpreting the CART models

Sex-stratified results of the CART models are shown in Figures 1 to 3. Figure 1 presents results for BMI missingness, Figure 2 for body mass missingness and Figure 3 for height missingness. 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 the situation that is more probable for data 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 the tree (i.e. closer to node 1) and those that appear more often can be considered more relevant criteria than variables that only appear once further down the tree.

For example, in the female BMI missingness CART model (Figure 1), the data are first split by weight perception. If individuals in this sample perceived their weight to be "about right" or underweight, they are in node 2. Node 2 contains 74% of the sample, and in this node the probability of missing BMI values is 0.27. If individuals perceived themselves to be overweight (i.e. the other remaining category for this variable), they are in node 3, which contains 26% of the data and where the probability of missing BMI values is 0.38. Similarly, for continuous variables, cut-offs are identified by the CART models. For example, in the female BMI missingness model the second node indicates that the model determined that 15 years of age was the cut-off that most differentiated the following sub-nodes.

### CART model accuracy

Accuracy testing using the test partition of the dataset showed that all models

**TABLE 1**
**Descriptive statistics of COMPASS study sample, 2018/19 (n = 74 501)**

| Variables[a] | Females (n = 36 546) | Males (n = 37 126) | Total[b] (n = 74 501) |
|---|---|---|---|
| **BMI variables** | | | |
| Mean BMI[c], score (SD) | 20.98 (3.02) | 21.21 (3.24) | 21.10 (3.14) |
| Missing scores, % (n) | 30.35 (11 093) | 31.22 (11 591) | 31.31 (23 329) |
| Mean height, m (SD) | 163.4 (7.50) | 174.2 (10.24) | 168.7 (10.47) |
| Missing, % (n) | 14.88 (5439) | 19.04 (7067) | 17.52 (13 049) |
| Mean body mass, kg (SD) | 57.42 (13.13) | 66.59 (17.74) | 62.16 (16.44) |
| Missing, % (n) | 21.75 (7948) | 19.79 (7348) | 21.33 (15 894) |
| **Age** | | | |
| Mean age, years (SD) | 15.14 (1.50) | 15.18 (1.51) | 15.16 (1.51) |
| Missing, % (n) | 0.08 (31) | 0.19 (69) | 0.73 (541) |
| **Ethnicity[d]** | | | |
| Racialized, % (n) | 69.45 (25 383) | 68.62 (25 477) | 68.48 (51 017) |
| Non-racialized, % (n) | 30.27 (11 063) | 30.99 (11 505) | 30.63 (22 822) |
| Missing, % (n) | 0.27 (100) | 0.39 (144) | 0.89 (662) |
| **Weight perception** | | | |
| Underweight, % (n) | 11.47 (4190) | 21.00 (7795) | 16.30 (12 140) |
| Overweight, % (n) | 25.85 (9448) | 19.93 (7398) | 22.87 (17 038) |
| About right, % (n) | 61.14 (22 343) | 57.19 (21 233) | 58.92 (43 893) |
| Missing, % (n) | 1.55 (565) | 1.89 (700) | 1.92 (1430) |
| **Diet-related variables** | | | |
| Fruit/vegetable consumption (24-hour recall) | | | |
| Mean number of servings, n (SD) | 2.89 (1.89) | 3.06 (2.11) | 2.98 (2.01) |
| Missing, % (n) | 2.44 (890) | 4.74 (1759) | 3.79 (2822) |
| Meat/meat alternatives consumption (24-hour recall) | | | |
| Mean number of servings, n (SD) | 1.88 (1.03) | 2.41 (1.20) | 2.15 (1.15) |
| Missing, % (n) | 2.45 (896) | 4.76 (1766) | 3.80 (2833) |
| Breakfast consumption | | | |
| Mean days per week, n (SD) | 4.67 (2.37) | 5.05 (2.33) | 4.85 (2.36) |
| Missing, % (n) | 1.31 (479) | 2.30 (855) | 1.99 (1484) |
| Grain consumption (24-hour recall) | | | |
| Mean number of servings, n (SD) | 2.41 (1.52) | 2.98 (1.93) | 2.69 (1.77) |
| Missing, % (n) | 2.33 (851) | 4.61 (1711) | 3.67 (2737) |
| Milk/alternatives consumption (24-hour recall) | | | |
| Mean number of servings, n (SD) | 1.77 (1.32) | 2.39 (1.54) | 2.08 (1.47) |
| Missing, % (n) | 2.33 (853) | 4.57 (1697) | 3.66 (2724) |
| Fast-food consumption | | | |
| Mean number of days per week, n (SD) | 1.19 (1.34) | 1.43 (1.61) | 1.31 (1.49) |
| Missing, % (n) | 1.03 (380) | 2.16 (801) | 1.81 (1345) |
| **Movement-related variables** | | | |
| Sports participation | | | |
| Participated in sports, % (n) | 56.70 (20 720) | 62.05 (23 036) | 59.24 (44 135) |
| Did not participate in sports, % (n) | 41.70 (15 241) | 35.25 (13 088) | 38.41 (28 618) |
| Missing, % (n) | 1.60 (585) | 2.70 (1002) | 2.35 (1748) |
| Strength training | | | |
| Mean number of days per week, n (SD) | 2.24 (2.02) | 2.77 (2.27) | 2.51 (2.16) |
| Missing, % (n) | 1.29 (473) | 1.93 (717) | 1.80 (1344) |

**Continued on the following page**

**TABLE 1 (continued)**
**Descriptive statistics of COMPASS study sample, 2018/19 (n = 74 501)**

| Variables[a] | Females (n = 36 546) | Males (n = 37 126) | Total[b] (n = 74 501) |
|---|---|---|---|
| Physically active friends | | | |
| Mean number, n (SD) | 3.03 (1.68) | 3.52 (1.69) | 3.28 (1.71) |
| Missing, % (n) | 1.35 (494) | 2.13 (789) | 1.92 (1430) |
| Screen time sedentary behaviour | | | |
| Mean hours per day, n (SD) | 5.92 (3.35) | 6.37 (3.37) | 6.15 (3.37) |
| Missing, % (n) | 4.41 (1613) | 5.94 (2206) | 5.44 (4056) |
| Moderate-to-vigorous physical activity | | | |
| Mean hours per day, n (SD) | 1.60 (1.23) | 2.00 (1.47) | 1.80 (1.38) |
| Missing, % (n) | 1.87 (683) | 2.56 (949) | 2.39 (1777) |
| Sleep | | | |
| Mean hours per night, n (SD) | 7.47 (1.30) | 7.60 (1.28) | 7.54 (1.29) |
| Missing, % (n) | 7.33 (2679) | 8.92 (3310) | 8.38 (6241) |
| **Academic variables** | | | |
| English grade (or French grade, in the case of French-language schools) | | | |
| Grade < 50%, % (n) | 1.09 (399) | 2.44 (907) | 1.83 (1362) |
| Grade ≥ 50%, % (n) | 95.39 (34 862) | 91.92 (34 128) | 93.41 (69 590) |
| Missing, % (n) | 3.52 (1285) | 5.63 (2091) | 4.76 (3549) |
| **Mental health–related variables** | | | |
| Self-rated mental health | | | |
| Mean score (SD) | 2.76 (1.21) | 2.21 (1.15) | 2.49 (1.21) |
| Missing, % (n) | 3.37 (1230) | 6.05 (2245) | 4.93 (3670) |
| Well-being[e] | | | |
| Mean score (SD) | 31.78 (5.75) | 32.64 (5.60) | 32.19 (5.72) |
| Missing, % (n) | 4.84 (1770) | 6.78 (2518) | 6.02 (4486) |
| Self-concept[f] | | | |
| Mean score (SD) | 11.79 (4.69) | 9.76 (4.19) | 10.79 (4.58) |
| Missing, % (n) | 3.34 (1221) | 5.51 (2045) | 4.64 (3455) |
| **Substance use variables** | | | |
| Smoking | | | |
| In the last 30 days, % (n) | 6.64 (2425) | 8.00 (2969) | 7.43 (5532) |
| Not in the last 30 days, % (n) | 92.89 (33 949) | 91.01 (33 790) | 91.70 (68 320) |
| Missing, % (n) | 0.47 (172) | 0.99 (367) | 0.87 (649) |
| E-cigarette use | | | |
| In the last 30 days, % (n) | 25.48 (9312) | 30.34 (11 264) | 27.99 (20 852) |
| Not in the last 30 days, % (n) | 73.75 (26 951) | 67.98 (25 237) | 70.62 (52 614) |
| Missing, % (n) | 0.77 (172) | 1.68 (625) | 1.39 (1035) |
| Cannabis use | | | |
| In the last 30 days, % (n) | 10.95 (4001) | 14.70 (5458) | 12.97 (9662) |
| Not in the last 30 days, % (n) | 88.06 (32 183) | 83.36 (30 950) | 85.42 (63 637) |
| Missing, % (n) | 1.00 (362) | 2.32 (718) | 1.61 (1202) |

**Abbreviations:** BMI, body mass index; SD, standard deviation.

[a] Only those variables present in at least one of the final classification and regression tree (CART) models.

[b] Includes respondents who did not report sex, so sex-stratified counts may not add up to total counts.

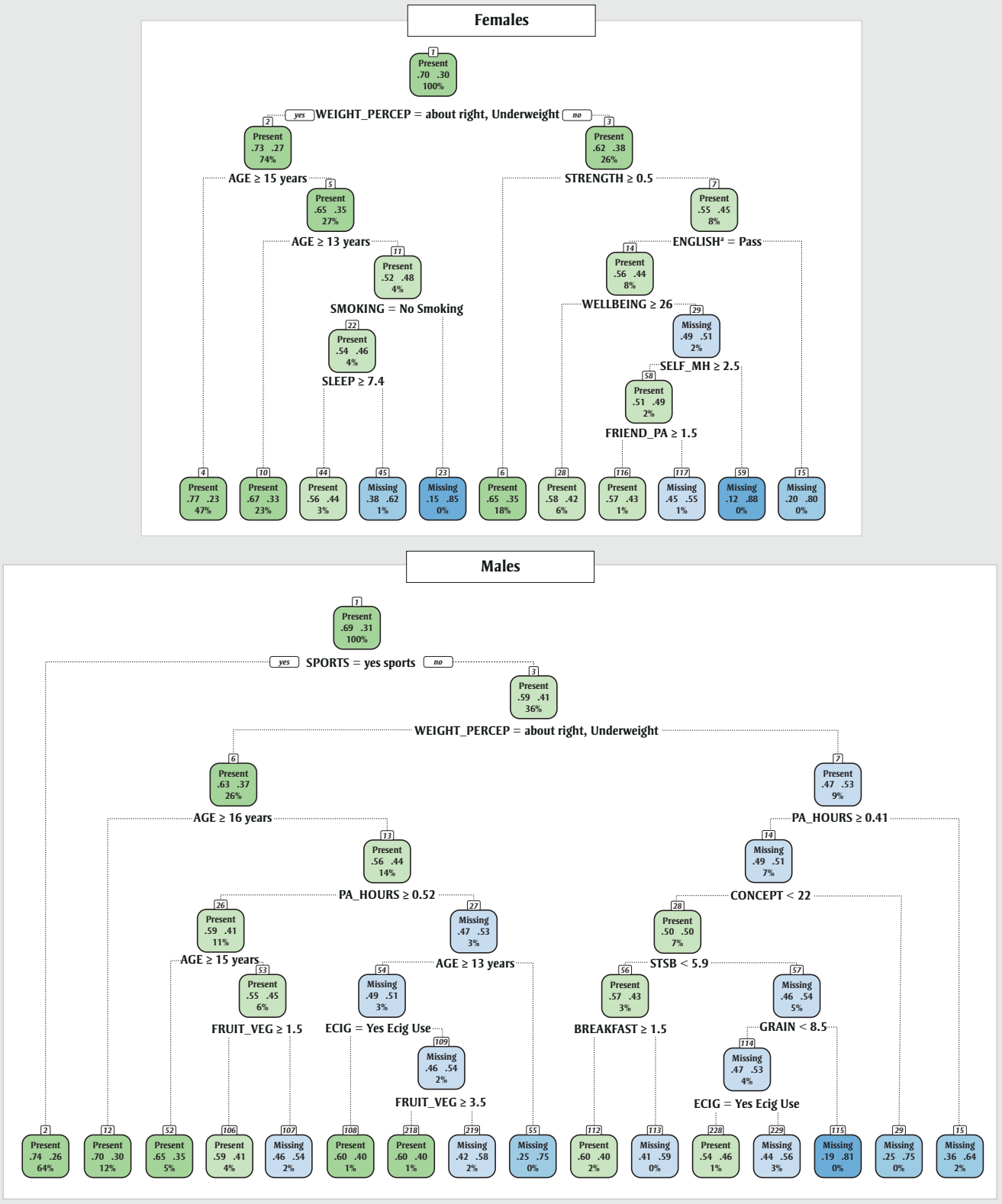
[c] Derived by dividing body mass (kg) by height squared ($m^2$).

[d] The survey question was "How would you describe yourself?" (Select all that apply), with the following response options: White, Black, Asian, Aboriginal (First Nations, Métis, Inuit), Latin American/Hispanic, Other. Respondents who selected "White" were classified as non-racialized, while respondents who selected any other option (including the selection of multiple options) were classified as racialized.

[e] Based on the Flourishing scale.[23]

[f] Based on the Self Description Questionnaire II short form.[24]

**FIGURE 1**
**BMI missingness CART models, for females (n = 36 546) and males (n = 37 126), COMPASS 2018/19**



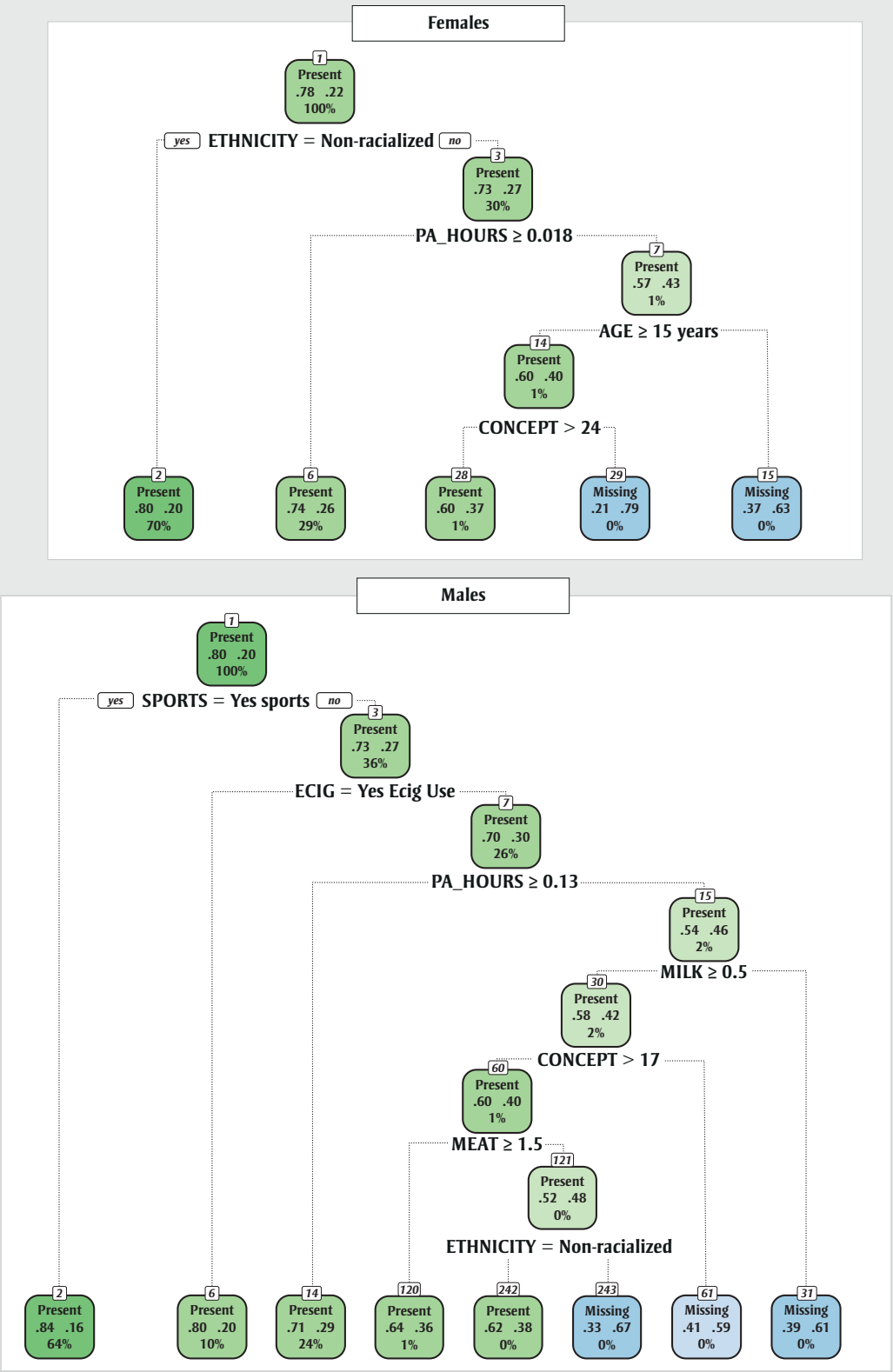

**Abbreviations:** BMI, body mass index; BREAKFAST, breakfast consumption; CART, classification and regression tree; CONCEPT, self-concept (based on the Self Description Questionnaire II short form[24]); ECIG, e-cigarette use; FRIEND_PA, physically active friends; FRUIT_VEG, fruit/vegetable consumption; GRAIN, grain consumption; PA_HOURS, moderate-to-vigorous physical activity; PERCEP, perception; SELF_MH, self-rated mental health; SPORTS, sports participation; STSB, screen time sedentary behaviour.

**Notes:** The label and colour of each node, "present" (green) or "missing" (blue), represents the situation that is more probable for data in that node; darker colours indicate higher probability. The left side of each node shows the probability of being present, and the right side shows the probability of being missing.

% indicates percentage of the sample in that node.

[a] In the case of French-language schools, this is the French grade.

**FIGURE 2**
**Body mass missingness CART models, for females (n = 36 546) and males (n = 37 126), COMPASS 2018/19**



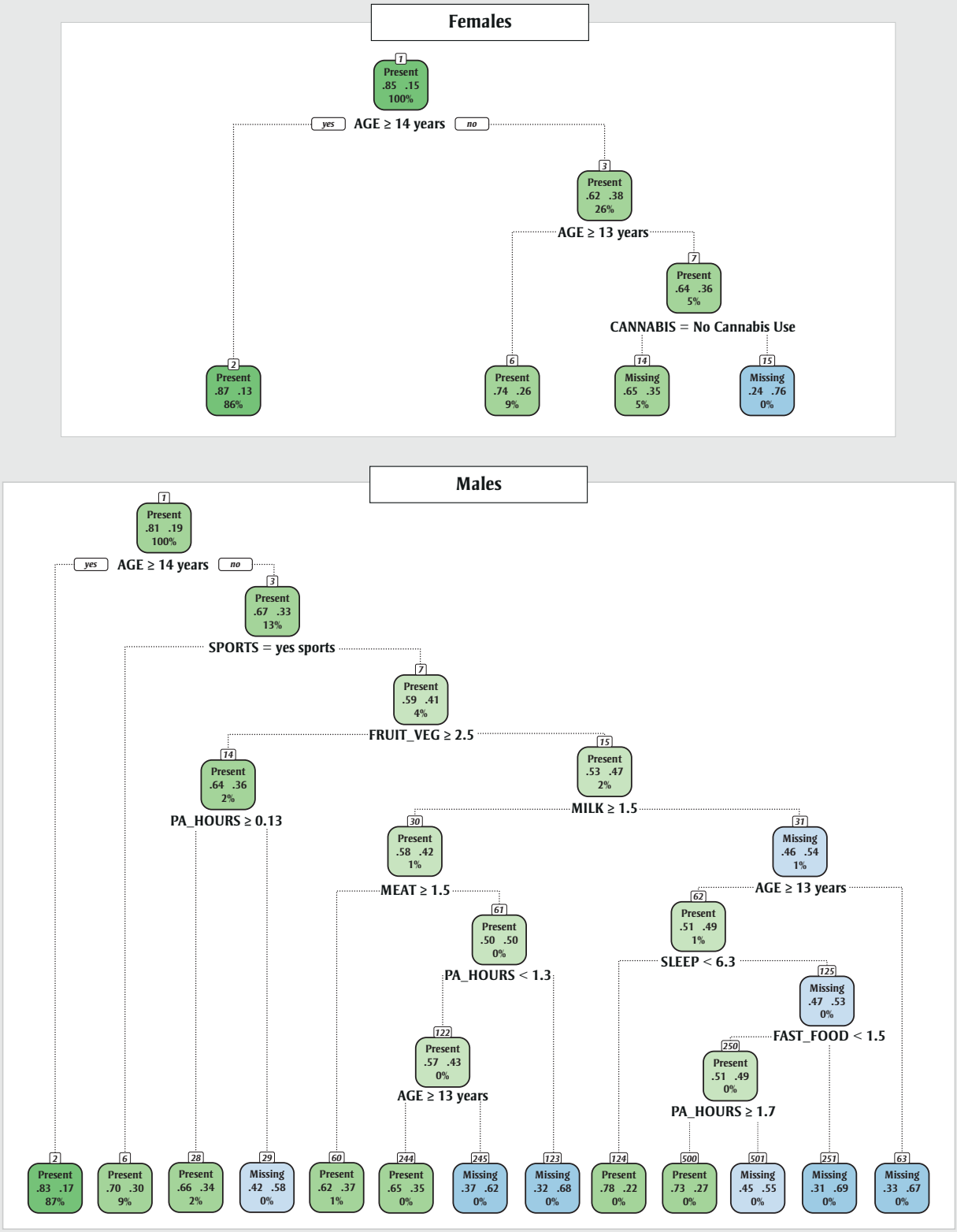

**Abbreviations:** CART, classification and regression tree; CONCEPT, self-concept (based on the Self Description Questionnaire II short form[24]); ECIG, e-cigarette use; MEAT, meat/meat alternatives consumption; MILK, milk/alternatives consumption; PA_HOURS, moderate-to-vigorous physical activity.

**Notes:** The label and colour of each node, "present" (green) or "missing" (blue), represents the situation that is more probable for data in that node; darker colours indicate higher probability. The left side of each node shows the probability of being present, and the right side shows the probability of being missing.

% indicates percentage of the sample in that node.

The survey question for the "Ethnicity" variable was "How would you describe yourself?" (Select all that apply), with the following response options: White, Black, Asian, Aboriginal (First Nations, Métis, Inuit), Latin American/Hispanic, Other. Respondents who selected "White" were classified as non-racialized, while respondents who selected any other option (including the selection of multiple options) were classified as racialized.

**FIGURE 3**
**Height missingness CART models for females (n = 36 546) and males (n = 37 126), COMPASS 2018/19**




**Abbreviations:** BMI, body mass index; CANNABIS, cannabis use; CART, classification and regression tree; FAST_FOOD, fast-food consumption; FRUIT_VEG, fruit/vegetable consumption; MEAT, meat/meat alternatives consumption; MILK, milk/alternatives consumption; PA_HOURS, moderate-to-vigorous physical activity; SLEEP, sleep duration; SPORTS, sports participation.

**Notes:** The label and colour of each node, "present" (green) or "missing" (blue), represents the situation that is more probable for data in that node; darker colours indicate higher probability. The left side of each node shows the probability of being present, and the right side shows the probability of being missing.

% indicates percentage of the sample in that node.

became more accurate after pruning. Pruned accuracy of CART BMI models was 69% for females and 70% for males, of CART body mass models was 78% for females and 80% for males and of CART height models was 85% for females and 81% for males.

## 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 can be missing in large proportions and published examinations of this missingness are lacking. 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.

When we previously examined missing BMI, height and body mass data in this sample using a regression approach,[15] we found that more information was needed on the structure of missingness and hierarchy of the importance of variables. The decision tree approach used in this study yielded insights into the mechanisms of missingness in this sample that can inform future studies on youth OWOB.

### *Mechanisms of BMI, height and body mass missingness*

In the BMI missingness CART models we developed, age and weight perception were among the first few primary splits for both males and females. Previous research has suggested that individuals who are younger are more likely to be missing BMI values because they don't know their own height and body mass;[26] this is consistent with 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 from their "about right" and underweight counterparts, leading to a higher likelihood of missingness in the group who perceived themselves as overweight. Previous studies examining BMI missingness mechanisms did not include a measure of weight perception, but two studies have found that poorer body satisfaction was associated with greater likelihood of missing BMI values.[14,27]

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 individuals who did not do any strength training were more likely on average to be missing BMI values. A similar mechanism was observed for males, but with sports and hours of physical activity; not playing sports or being, on average, less physically active each day led to splits where the likelihood of missing BMI values was greater. This is consistent with previous research that included some measures of physical activity.[6,7,14]

Mental health–related variables also appeared in both male and female models. For females, well-being and self-rated mental health were used for splitting, and for males, self-concept was used. For all these mental health–related variables, lower scores (i.e. scores indicating poorer mental health) were associated with a greater likelihood of missing BMI values.

The consistent splitting of individuals who perceived themselves as overweight into a separate group more likely to be missing BMI values suggests that those with a higher BMI were more likely to be nonreporters. Notably, weight perception cannot be assumed as a direct proxy for BMI or body mass because youth may miscategorize themselves;[28-30] however, weight perception may be considered alongside other factors to determine which missingness pattern is most probable.

Findings related to physical activity support the idea that individuals missing BMI values are more likely to have a higher BMI, as those who are less physically active were also split into groups more likely to be missing BMI values, and inverse associations between physical activity and BMI are well-established.[31,32] These findings, along with what we know about heightened body image concerns during adolescence,[33] demonstrate that social desirability may be playing a role in youth nonreporting of height and body mass in this sample.

Height and body mass missingness CART models had some split criteria similar to those of the BMI missingness models, with age a common partitioning variable and physical activity, diet, mental health and substance use variables also observed. One finding exclusive to the body mass missingness models was ethnicity: the model indicated that racialized individuals 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.

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

The decision tree approach used in this study to examine missingness appears to have several advantages over traditional regression approaches. The visual nature of decision tree models makes them particularly useful for understanding *how* missingness might be influenced by other variables. For example, the inclusion and directions of splitting related to weight perception, physical activity and mental health in the CART models suggest that the missingness in BMI may be not missing at random because missing data appear more likely among those with a higher BMI. While not missing at random is not a provable phenomenon, the CART models provide evidence against a missing-completely-at-random mechanism, as several subgroups who are highly likely to be missing BMI were identified based on observed covariates.[34] Future OWOB research should consider the mechanisms and degree of missingness in BMI, and where examinations indicate that data may be missing at random or not missing at random, certain statistical approaches (e.g. complete case analysis) may not be ideal because of the risk of bias.[9]

While a regression model could similarly highlight the associations between predictor variables and BMI missingness (e.g. observing a positive odds ratio for self-perception as overweight), one advantage of the CART models is the easily observed hierarchy of the importance of variables. 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. We previously examined BMI missingness using regression;[15] while weight perception was significantly associated with missingness, it was only one of many significant variables

and relative importance couldn't be empirically discerned.

Another advantage of CART models is that one can follow through a decision tree order to identify 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 than not to be missing BMI values. Moreover, following subgroups to the bottom of the trees reveals that, overall, individuals 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 values. In other words, CART models identified that those in the complete sample (i.e. those not missing BMI data) were physically, emotionally and mentally healthier than their counterparts with missing data. As such, a complete case analysis approach on these data would certainly be biased, potentially leading to incorrect research conclusions and inappropriate related policy and programming recommendations.

Examining missing data is often the first step in certain statistical approaches, such as multiple imputation. Although such examinations are needed to identify auxiliary variables that can inform reasonable imputed values, selecting these variables can be difficult if there are many variables related to missingness. This was the case with our 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 have been appropriate as these variables use different scales.[15]

The hierarchical nature of CART models makes the process of selecting auxiliary variables more systematic. For example, CART models can parse out redundant variables; while previous regression work identified weight goal as significantly related to BMI missingness,[15] the CART models in this study did not perform any splits based on this variable, possibly because BMI missingness is sufficiently explained by the weight perception variable alone.

In this study we demonstrated the utility of using CART models to examine missingness in youth height, body mass and BMI. However, missingness is pervasive, 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 make this approach reasonably accessible and feasible for applied researchers.

## Conclusion

This study adds to the limited existing research examining missingness in youth BMI, height and body mass data. CART models demonstrated that age, self-perception as overweight, lower physical activity and poorer mental health identified the subgroups most likely to be missing BMI values. The direction of model partitioning for these variables suggests that youth with higher BMI may be more likely to be missing BMI values and that deleting missing cases in an analysis would likely lead to biased findings.

Future research using youth self-reported data may find that CART models are a particularly useful tool for examining missingness and help select a statistical approach appropriate for handling missing data.

## Acknowledgements

The COMPASS study has been supported by a bridge grant from the Canadian Institutes of Health Research (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); a research funding arrangement with Health Canada (#1617-HQ-000012; contract awarded to SL); and a CIHR–Canadian Centre on Substance Abuse and Addiction (CCSA) team grant (OF7 B1-PCPEGT 410-10-9633; awarded to SL).

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 also benefits from funding from the Ministère de la Santé et des Services sociaux of the province of Quebec, and the Direction régionale de santé publique du CIUSSS de la Capitale-Nationale.

## Conflicts of interest

The authors report no conflicts of interest.

## Authors' contributions and statement

AD, AC, JPC and SL – Conceptualization; AD and AC – Methodology; AD – Formal analysis, Writing – Original draft; SL – Funding acquisition, Resources, Supervision; AD, AC, JPC and SL – Writing – Review and editing. All authors read and agreed upon the published version of the manuscript.

The content and views expressed in this article are those of the authors and do not necessarily reflect those of the Government of Canada.

## References

1. 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;9(1):88. https://doi.org/10.1186/1471-2458-9-88

2. Maukonen M, Männistö S, Tolonen H. A comparison of measured versus self-reported anthropometrics for assessing obesity in adults: a literature review. Scand J Public Health. 2018;46(5):565-79. https://doi.org/10.1177/1403494818761971

3. Sherry B, Jefferds ME, Grummer-Strawn LM. Accuracy of adolescent self-report of height and weight in assessing overweight status: a literature review. Arch Pediatr Adolesc Med. 2007;161(12):1154-61. https://doi.org/10.1001/archpedi.161.12.1154

4. Lipsky LM, Haynie DL, Hill C, et al. Accuracy of self-reported height, weight, and BMI over time in emerging adults. Am J Prev Med. 2019;56(6):860-8. https://doi.org/10.1016/j.amepre.2019.01.004

5. Taylor AW, Dal Grande E, Gill TK, et al. How valid are self-reported height and weight? A comparison between CATI self-report and clinic measurements using a large cohort study. Aust N Z J Public Health. 2006;30(3):238-46. https://doi.org/10.1111/j.1467-842X.2006.tb00864.x

6. 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;53:54-8. https://doi.org/10.1016/j.nut.2018.01.003

7. Arbour-Nicitopoulos KP, Faulkner GE, Leatherdale ST. Learning from non-reported data: interpreting missing body mass index values in young children. Meas Phys Educ Exerc Sci. 2010;14(4):241-51. https://doi.org/10.1080/1091367X.2010.520243

8. Hallgren KA, Witkiewitz K. Missing data in alcohol clinical trials: a comparison of methods. Alcohol Clin Exp Res. 2013;37(12):2152-60. https://doi.org/10.1111/acer.12205

9. White IR, Carlin JB. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med. 2010;29(28):2920-31. http://doi.org/10.1002/sim.3944

10. 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 Med Res Methodol. 2012;12(1):96. https://doi.org/10.1186/1471-2288-12-96

11. Eekhout I, de Boer RM, Twisk JW, de Vet HC, Heymans MW. Missing data: a systematic review of how they are reported and handled. Epidemiology. 2012;23(5):729-32. https://doi.org/10.1097/EDE.0b013e3182576cdb

12. Lin JY, Lu Y, Tu X. How to avoid missing data and the problems they pose: design considerations. Shanghai Arch Psychiatry. 2012;24(3):181-4. https://doi.org/10.3969/j.issn.1002-0829.2012.03.010

13. Van Dyke N, Drinkwater EJ, Rachele JN. Improving the accuracy of self-reported height and weight in surveys: an experimental study. BMC Med Res Methodol. 2022;22(1):241. https://doi.org/10.1186/s12874-022-01690-x

14. Fonseca H, de Matos MG, Guerra A, Gomes-Pedro J. Emotional, behavioural and social correlates of missing values for BMI. Arch Dis Child. 2008;94(2):104-9. https://doi.org/10.1136/adc.2008.139915

15. Doggett A, Chaurasia A, Chaput JP, Leatherdale ST. Learning from missing data: examining nonreporting patterns of height, weight, and BMI among Canadian youth. Int J Obes. 2022;46(9):1598-607. https://doi.org/10.1038/s41366-022-01154-8

16. Ten Eyck P, Cavanaugh JE. An alternate approach to pseudo-likelihood model selection in the generalized linear mixed modeling framework. Sankhya B. 2018;80:98-122. https://doi.org/10.1007/s13571-017-0130-5

17. Loh WY. Fifty years of classification and regression trees. Int Stat Rev. 2014;82(3):329-48. https://doi.org/10.1111/insr.12016

18. 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. Ann Behav Med. 2003;26(3):172-81. https://doi.org/10.1207/S15324796ABM2603_02

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

20. Radloff LS. The CES-D Scale: a self-report depression scale for research in the general population. Appl Psychol Meas. 1977;1(3):385-401. https://doi.org/10.1177/014662167700100306

21. Spitzer RL, Kroenke K, Williams JBW, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Arch Intern Med. 2006;166(10):1092-7. https://doi.org/10.1001/archinte.166.10.1092

22. Gratz KL, Roemer L. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. J Psychopathol Behav Assess. 2004;26(1):41-54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94

23. Diener E, Wirtz D, Tov W, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res. 2010;97(2):143-56. https://doi.org/10.1007/s11205-009-9493-y

24. 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. Psychol Assess. 2005;17(1):81-102. https://doi.org/10.1037/1040-3590.17.1.81

25. Boehmke B, Greenwell B. Hands-on machine learning with R. Boca Raton (FL): CRC Press; 2020.

26. Himes JH. Challenges of accurately measuring and using BMI and other indicators of obesity in children. Pediatrics. 2009;124 Suppl 1:S3-22. https://doi.org/10.1542/peds.2008-3586D

27. Tiggemann M. Nonreporting of body mass index: a research note on the interpretation of missing data. Int J Eat Disord. 2006;39(4):346-9. http://doi.org/10.1002/eat.20264

28. Sirirassamee T, Phoolsawat S, Limkhunthammo S. Relationship between body weight perception and weight-related behaviours. J Int Med Res. 2018;46(9):3796-808. https://doi.org/10.1177/0300060518780138

29. 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. Int J Adolesc Youth. 2020;25(1):264-75. https://doi.org/10.1080/02673843.2019.1614465

30. Wang Y, Liu H, Wu F, et al. The association between BMI and body weight perception among children and adolescents in Jilin City, China. PLoS One. 2018;13(3):e0194237. https://doi.org/10.1371/journal.pone.0194237

31. Peart T, Velasco Mondragon HE, Rohm-Young D, Bronner Y, Hossain MB. Weight status in US youth: the role of activity, diet, and sedentary behaviors. Am J Health Behav. 2011;35(6):756-65. https://doi.org/10.5993/AJHB.35.6.11

32. 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(1):354. https://doi.org/10.1186/1471-2458-12-354

33. Voelker DK, Reel JJ, Greenleaf C. Weight status and body image perceptions in adolescents: current perspectives. Adolesc Health Med Ther. 2015;6:149-58. https://doi.org/10.2147/AHMT.S68344

34. Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581-92. https://doi.org/10.1093/biomet/63.3.581