4-21-2020 1:00 PM

Evaluating quantitative methods for intercategorical-intersectionality research: a simulation study

Mayuri Mahendran, *The University of Western Ontario*

Supervisor: Bauer, Greta R, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Epidemiology and Biostatistics

Abstract

This study evaluated eight quantitative methods for their predictive accuracy for intersectionally-defined subgroups, via a simulation study. The methods included two forms of single-level regression with interaction terms, cross-classification, multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA), and four decision tree methods: classification and regression trees (CART), conditional inference trees, chi-square automatic interaction detector, and random forest. The simulated datasets varied by outcome variable type, input variable types, sample size, and size and direction of the effects. Predictive accuracy improved with increasing sample size for all methods except CART. At small sample sizes, random forest and MAIHDA generally created the most precise predictions. While performing well for prediction, variable selection by random forest and confidence interval coverage and power of MAIHDA main effects coefficients were suboptimal. We have identified differences in methods ideal for intersectional prediction versus variable identification, highlighting that different objectives and data scenarios require different methods.

Key words

Intersectionality, prediction, quantitative methods, multilevel analysis, health inequalities, decision trees

Summary for Lay Audience

Intersectionality acknowledges that an individual's multiple social positions or identities (e.g. gender, ethnicity) can interact to affect health-related outcomes in unique ways. Calculating health outcomes for intersectional groups (defined by a combination of positions), rather than by each position separately, can create more accurate outcome estimates. Since it is unclear which methods do this best, this study evaluated eight methods in terms of their predictive performance for intersectional groupings, using simulated data with known true values. The methods included single-level and multilevel regression, cross-classification, and four machine learning methods (classification and regression trees (CART), conditional inference trees, chi-square automatic interaction detector, and random forest). The accuracy of predictions created by all methods generally improved with increasing sample size, except for the CART method. Generally, random forest and the multilevel method created the most precise predictions compared to the other methods, especially for small sample sizes. However, they did not always correctly identify variables which were significantly associated with outcome. Random forest sometimes incorrectly suggested that a variable that had no true effect on the outcome was important, and MAIHDA created estimates for the effects of individual variables that were not reflective of the expected values. This shows that while some methods are reliable to predict the outcome for intersectionally defined groups, they are not ideal to identify the effects or importance of individual variables that make up those groups (e.g. the specific effect of being in a high income group, or being male). Results from this work will improve the application of quantitative methods for accurately estimating outcomes for population subgroups. Correctly estimating outcomes for these groups is an important step in understanding existing health inequities. The goal of this work is to produce a guide for researchers who are interested in the applications of quantitative intersectionality approaches.

Acknowledgements

I have many people to thank for supporting and encouraging me throughout this project. First, I would like to thank my supervisor Dr. Greta Bauer, who was given me guidance, opportunity, and foremost confidence in my own abilities. She has been a great mentor who has known just how to push me further and taught me to trust myself as a researcher. Dr. Bauer has truly helped me make this degree a fulfilling learning experience. I would also like to thank the members of my advisory committee Dr. Dan Lizotte and Dr. Yayuan Zhu, who enabled me to venture farther into statistics than I imagined I would. I thank them for their time, valuable insights, and ongoing encouragements.

I would also like to acknowledge the Department of Epidemiology and Biostatistics at Western. Having both completed both my undergraduate and graduate degrees here, I am grateful for the support and opportunities that many staff and faculty here have provided me. They have seen me grow a little older and I hope a little wiser.

To all the current and past members of the Health Equity in Epidemiology Research group, thank you for all the kind words of advice and support. A special thank you to the members of the team who patiently waited while I exhausted their desktops running simulations. I am grateful to have worked with such a wonderful group of individuals.

Finally, I would like to extend my utmost thanks to my loved ones, family and friends. To my family, who supported and always encouraged me to pursue and value my education. And to all the friends I have made along the way, who have celebrated my successes like their own, and made sure I remembered to enjoy the process. I am fortunate to have such wonderful people in my life.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations and Symbols

CART: Classification and regression trees

CHAID: Chi-square Automatic Interaction Detector

CTree: Conditional inference trees

ICC: Intra-class correlation coefficient

IRR: Incidence rate ratio

MAIHDA: Multilevel analysis of individual heterogeneity and discriminatory accuracy

MAPE: Mean absolute percentage error

MSE: Mean squared error

OLS: Ordinary least squares

OR: Odds ratio

RR: Relative risk

SE: Standard error

SES: Socioeconomic status

Chapter 1

1 Introduction and objectives

1.1 Health equity and heterogeneity of effects

Health equity research aims to identify and reduce the modifiable differences in health between groups defined by social, economic, or geographic means. [1] Link and Phelan [2] argued that social conditions, such as socioeconomic status or race and ethnicity, are "fundamental causes" of diseases. Similarly, Geoffrey Rose [3] stated that, ''The primary determinants of disease are mainly economic and social, and therefore its remedies must also be economic and social''. These fundamental causes are connected to disease because they determine resource accessibility and availability, and likelihood of exposure to risk factors for and protective factors against disease. Along with acknowledging the existence of social determinants, it is also important to consider heterogeneity of effects. From a public health perspective, interaction and effect measure modification among social determinants should be recognized as a possibility when performing subgroup identification for targeted interventions. [4] As stated by Greenland [4] "In the absence of bias, departures from risk additivity imply that some subgroups would obtain a greater absolute risk reduction from the intervention than others would.". Departures from the additive scale can occur if the excess risk is beyond the additive ("super additivity"), or if the outcome occurs only when certain factors coincide ("synergism"). The identification of "super additivity" can indicate that groups may benefit from intervention more than expected, and synergism is seen as an indicator that only one factor need be addressed by interventions to affect the outcome. Research in health equity should incorporate the possibility of heterogeneity, and intersectionality theory can function as a research framework to address that fundamental causes or social determinants of health may have heterogeneous effects.

1.2 Intersectionality theory

Intersectionality theory acknowledges that an individual occupies multiple social categories or identities such as gender, race, and class, which overlap and can interact to create unique positions of systemic privilege and oppression. [5, 6] The term intersectionality first came to use by Black feminist legal scholar Kimberlé Crenshaw, to describe the position of Black women and their exclusion from both racial and gender discourse. This theory has since been extended to social positions and identities beyond gender and race, such as income, age, sexuality, and disability status, and to disciplines such as sociology [7], psychology [8], and education [9].

Intersectionality has applications to public and population health research [10, 11]. Bowleg [10] suggests that intersectionality can contribute to public health research not as a testable theory to be proven or disproven, but rather as a guiding perspective or framework, that acknowledges that individuals occupy multiple social identities and positions that can interact together and with the surrounding socio-structural factors (e.g. racism, sexism) to affect health outcomes. Intersectionality encourages research to make space for individuals who occupy multiple disadvantaged positions, as well as those who occupy a mix of advantaged and disadvantaged positions. [11] No one position or identity has presumed importance over the other. [12] This framework encourages the study of health the way it is actually experienced in society, as a result of complex interactions. Multiple micro- and macro-level factors can be incorporated, which aligns with addressing “fundamental causes” for inequalities (e.g. discrimination and poverty). An intersectional approach encourages targeted health promotion and policy, rather than assuming homogeneity across single factors, which can result in policies that are ineffective or harmful for oppressed or marginalized groups. [10]

McCall [13] describes three approaches to how intersectionality is incorporated into research. The first is the anticategorical approach which acknowledges that categories are not set truths, because they over-simplify the complexity of actual experiences, which are fluid and dynamic. The second is the intracategorical approach, which focusses on experiences within a particular group or intersection, which usually experience some

level of marginalization. This approach requires some stability in the definition of belonging to these groups, but allows the researcher to delve into the complexity and variety of the experiences of different group members. The third is the intercategorical approach, which uses multiple defined categories to compare outcomes between intersectionally defined groups. This final approach is most readily applied by quantitative research. [14] Hancock [12] describes how intersectionality is distinguished from the "multiple approach". The multiple approach allows for several positions (e.g. gender and race) to be relevant to an outcome, but views them as separate effects that do not overlap. The underlying assumption is that these separate effects can be added together to predict the outcome. This is analogous to fitting regression models with main effects for gender and sex without interaction terms. The intersectional approach acknowledges that these positions cannot be simply added together, they exist in ways that cannot be separated. To move beyond the additive model, intercategorical intersectionality research is commonly applied by the inclusion of interaction terms or cross-classified groups. To clarify, the "multiple approach" as referenced by Hancock is what other studies mentioned below reference as the "additive model", because it assumes effects are additive. The term "multiplicative model" is sometimes used for what Hancock references as the "intersectional approach". Additionally, the language around additive and multiplicative models in intersectionality theory is not related to the statistical terminology for additive and multiplicative scales. [11] For example, the multiplicative approach can be applied on the additive scale by using a linear regression with interaction terms, or on the multiplicative scale by using a logistic regression with interaction terms. Similarly, the additive model can be applied on either the additive or multiplicative scale, depending on the type of regression, by the inclusion of only main effects.

1.3 Intersectionality theory for health equity stratification: application and issues

The current discussion is limited to descriptive intercategorical intersectionality and health equity stratification, which does not aim to prove causality, but rather describes the

differences and inequities between groups. This is a steppingstone for further qualitative or quantitative analytic intersectionality research.

When observing inequalities in self-rated health by race, sex, class, and sexual orientation in Canada using data from the Canadian Community Health Survey, Veenstra [15] demonstrated that the multiplicative model leads to different outcome predictions than the purely additive model (with no interaction terms). This was done by comparing a logistic regression model with no interaction terms with one including all two- and three-way interaction terms. Use of the multiplicative model also changed the interpretation of the inequities. For example, from the additive model Asian respondents in the lowest income group had a 32.6% probability of reporting fair or poor health, compared to 28.3% of white respondents in the lowest income group. However, when using an intersectional model with interaction terms, Asian respondents in the lowest income group actually fared better than their white counterparts, with a 17.4% probability of reporting fair or poor health, compared to 30.2%. These results show that assuming that social determinants function completely independently can affect conclusions regarding which groups face greater inequities. The authors note that not all intersections experienced “multiple jeopardy”, where those at the most marginalized groups were expected to experience the worst outcomes. This is similar to what Greenland [4] referred to as “super additivity”. But as conceptualized by Bright et. al. [16] “switch intersectionality” is a possibility that researchers should be mindful of, where the effects of a variable can actually be in the opposite direction than expected or completely unique to a particular intersection, because a causal process is only activated when individuals occupy certain intersectional positions. This is similar to the “synergism” mentioned by Greenland (4) when discussing heterogeneity of effects.

Other authors have attempted to further break down the meaning of differences between intersectionally defined groups. Jackson et. al. [17] looked at the intersection of race (non-Hispanic Black versus non-Hispanic white) and early life socioeconomic status (SES - low versus high), for differences in unemployment, wages, and incarceration. They separated the total difference between groups (joint disparity), as the sum of the referent and excess intersectional disparity. The joint disparity for example could be the

difference in the outcomes between a low SES Black male respondent, and a high SES white male respondent. The referent disparity can be seen as the "additive effects": the effects of being Black compared to white among those who are high SES, and the effects of being low compared to high SES among white males. The intersectional disparity is the remaining joint disparity that remains unaccounted for by the referent disparity, indicating a departure from solely additive effects. They found in some cases that the intersectional disparity was significant. Importantly, the authors remarked that in cases where the intersectional disparity is not significant, the joint disparity for multiply marginalized groups can still be quite large, and they still may experience the greatest inequities. Intersectional groups that don't have statistically significant intersectional effects may still be the most important targets for intervention or policy. These comments outline the importance of not focussing on intersectionality as a "testable explanation" [18], but rather as a research framework.

Quantitatively applying descriptive intersectionality into population health research faces challenges that have been outlined by several authors. Some specific issues include that although regression is a common analytic method, the use of regression often requires underlying assumptions regarding the relationship between variables, such as the linearity of main effects and interactions, which may not hold and generally go against the expectations of intersectionality. [10] Low sample sizes make it difficult to study every intersection, or to include the number of intersectional positions that would be of interest. For example, to use regression methods to study a larger number of intersectional groups necessitates the inclusion of multiple higher-order interaction terms, which require large sample sizes for sufficient statistical power. [4] Therefore, often only certain intersections, usually the most marginalized groups, are prioritized for study. [11] However, positions with a mix of both privilege and marginalization should also be considered in research, given that unknown intersectional effects could exist in these groups. Especially with the availability of larger datasets, "intersectional mapping" or "socio-demographic mapping" can be a way to describe outcomes across a large number of intersectional groups and identify intersections for further study. [11]

1.4 Review of quantitative intersectionality methods

A recent unpublished systematic review [19] assessed the state of the published quantitative intersectionality research through mid-2017, identifying quantitative intersectionality papers across multiple disciplines including epidemiology, psychology, political sciences, social sciences, and education. The result was a total of 319 studies published between 1989, when the term was first coined by Kimberlé Crenshaw, to May 2017. Of the 303 applied intersectionality studies identified by this review, 34.3% had a health-related outcome. The review found that the most applied methods were regression models, including Ordinary Least Squares (OLS), logistic, Poisson, and negative binomial. This includes regression models with main effects and either cross-classification or stratification (27.4%), or interaction terms (24.8%). Additionally, 6.6% of papers used main effects regression models as the only form of "intersectional" analysis. 18.5% of studies only used univariate or bivariate measures. Other applied methods included: multilevel modelling, MANOVA, structural equation modelling/path analysis, growth curve analysis, cluster analysis, multi-group segregation indices, latent class analysis, meta-regression, classification and regression trees (CART), intersectional decomposition, canonical correspondence analysis, Chi-square Automatic Interaction Detector (CHAID), and factor analysis. An example of the typical application of an intersectional regression model is a study by Cummings et. al. [20] looking at self-rated health along the intersection of gender, race, and SES. The regression included cross-classified variables by having separate dummy variables for white women, Black women and Black males. Interaction terms were also included to represent the intersection of all three positions, by including a separate interaction term between each of the three cross-classified variables and family income. Applications of regression may also be stratified by having separate regression analyses for each category of a social position (e.g. stratifying by gender by having separate regression models for male and female). [21] An example application of purely descriptive analysis is also found in the study by Cummings et. al. [20] where average self-rated health was tabulated by twelve categories created by the combination of gender (male and female), race (white and Black) and SES (low, middle, and high income). This simple descriptive method is described as cross-classification for the duration of this thesis.

According to the systematic review, [19] the social positions or identities most commonly included in intersectional research were sex/gender and race/ethnicity, in 76% and 73% of studies respectively. Other common intersectional positions were: SES (22%), sexual orientation (18%), immigration/nativity (13%), education (13%), age (10%), income (8%), and geography (6%). Figure 2.1 presents the number of intersectional positions and identities included in each study. Most studies included only 2 to 3 intersectional variables, reflective of the limitations of the most commonly used methods, regression and uni-/bi-variate analysis.

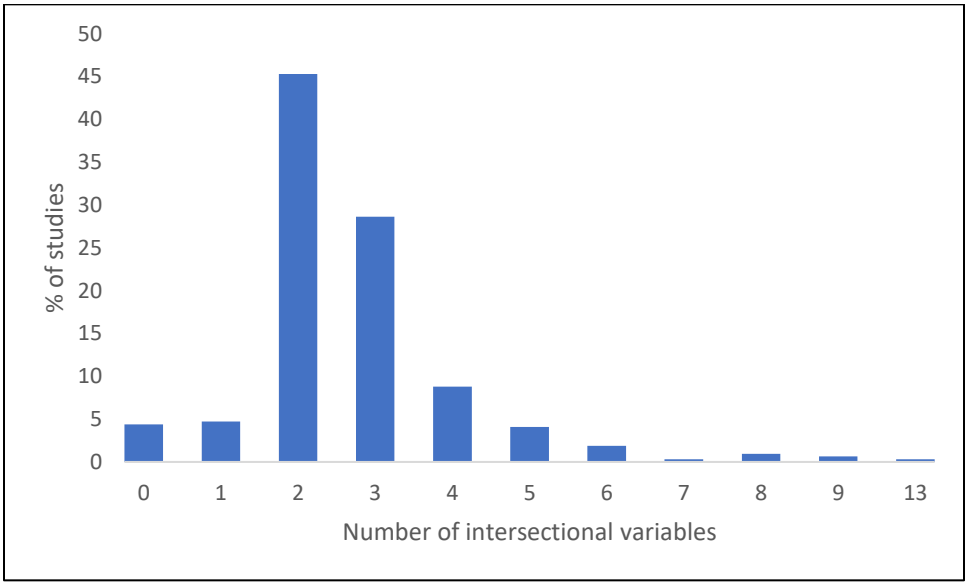

Figure 1.1: Number of intersectional variables included in analyses of intersectional studies. Data used with permission from Churchill SM.

Overall, the methods identified by the systematic review were applied to only a small number of intersections. The study did identify a few novel methods of interest, especially for the purposes of intersectional mapping: CART, CHAID, and the multi-level method MAIHDA (Multilevel analysis of individual heterogeneity and discriminatory accuracy).

1.5 Thesis objectives

It is currently unclear how to best incorporate the intercategorical intersectional perspective into descriptive health research, specifically for the purposes of intersectional mapping. This thesis will address gaps in the literature regarding which methodologies researchers may use, primarily when studying a larger number of intersecting positions. We compared the conventional intersectionality methods of regression and univariate cross-classification, the novel methods CART, CHAID, and MAIHDA, identified by the literature review detailed in section 1.4, as well as two additional methods identified by further review of the current literature (see Chapter 2), random forest and conditional inference trees (CTree).

The primary objective was to formally evaluate the predictive performance of eight methods, via a simulation study. This was achieved by answering:

1) Which methods have the lowest predictive error, when predicting outcomes for intersectionally-defined population-level subgroups?

The secondary objectives were to evaluate performance measures specific to the different methodologies. These were achieved by answering:

1) Regression:
    - **a.** How well do regression methods identify significant main effects and interactions?
    - **b.** What is the validity of the estimates for main effects and interaction terms?

2) MAIHDA:
    - a. How well does MAIHDA identify which variables are significant to the outcome?
    - b. What is the validity of the main effect estimates?

3) Decision Trees: The decision tree methods included in this study were CHAID, CART, random forest, and CTree.
    a. How well does each decision tree method identify variables relevant to the outcome?
    b. How many unique subgroups does each method identify?

Differences in each method's performance was assessed across a number of dataset parameters: sample size, variable input types, and outcome type. These parameters were selected with particular focus on dataset qualities and outcomes typical of and relevant for intersectional research and the social determinants of health, and were informed by the systematic review referenced in section 1.4 and the literature review detailed in Chapter 2.

Chapter 2

2 Literature Review

Based on the existing intersectionality literature, it is fairly well understood how regression with interaction terms and cross-classification are applied to intersectionality research and correspond to intersectionality theory. Simple descriptive studies use cross-classification by summarizing outcomes averages or prevalences across intersections, without any further statistical adjustment. Studies using regression most often include main effects and interaction terms, and interaction terms are interpreted as intersectional effects. However, it is unclear how novel methods for quantitative intersectionality research are being applied and interpreted. Therefore, a literature search was conducted of intersectionality studies using decision tree methods and intersectionality studies using MAIHDA. The following chapter explores what kinds of data scenarios are used with these methods, how the analyses are conducted, and how the results from these analyses are interpreted in relation to intersectionality theory. Given the limited variety of decision trees used in intersectionality research, further applications and discussions of decision trees in epidemiology were also explored. For both MAIHDA and decision trees, the current state of the literature regarding quantitative assessment of these methods was considered.

2.1 Decision Trees

2.1.1 What are decision trees

Decision trees fall under the category of supervised machine learning techniques, where an algorithm is given a set of potential input variables and a defined outcome variable. [22] In decision trees, data is partitioned according to a set of decision rules, resulting in groups defined based on a set of predictors or input variables. [23] The final end nodes are called “leaves”, or “terminal nodes” and are the final subgroupings identified by the tree. Decision trees can perform either classification analyses (for categorical outcomes) or regression analyses (for continuous outcomes). The terminal nodes or “leaves” of a classification tree depict what percentage of respondents from each node report the

outcome. The leaves of regression trees report the mean of the outcome. Often decision trees can be visualized as a tree diagram or flowchart, where the path from the initial "root" to final "leaf" is the set of decision rules. The general algorithm of a decision tree begins with the initial parent node, a group containing all data points, which is subsequently split into child nodes (or subgroups), using one of the given input variables (e.g. gender, or age). The criteria to identify a splitting variable can vary but the overall goal is to create groups based on covariates, that are similar to one another in regard to the outcome. Child nodes are then split repeatedly until a stopping criterion is reached. This is thus called recursive partitioning. Decision trees have been generally cited as beneficial for their ability to create accurate prediction models, consider a large number of variables, and as a non-parametric method can easily incorporate interactions and effects that are linear and non-linear. [24] Some of the negatives are that it can be prone to over-fitting the data [23], continuous variables with a true linear effect on the outcome require a great deal of splits to create predictions, and methods such as CART have been found to be biased to split on continuous variables over categorical [25].

Figure 2.1 is a figure published in a study looking at self-reported past year pap-tests among sexual minority women, and shows a visual example of a CART decision tree. [26] While 25 potential covariates were used as input variables, only 6 were actually identified as relevant and used in the tree building process. The final tree had 7 terminal nodes (those presented in colour for Figure 2.1), which are described by their decision rules in Table 2.1. Here we can see for example that certain variables like health insurance only have an effect on the outcome after a certain cut-off for age. Because decision trees are non-parametric, they are not required to consider effects as linear, and therefore are inherently able to account for effects such as this without further specification by the user.

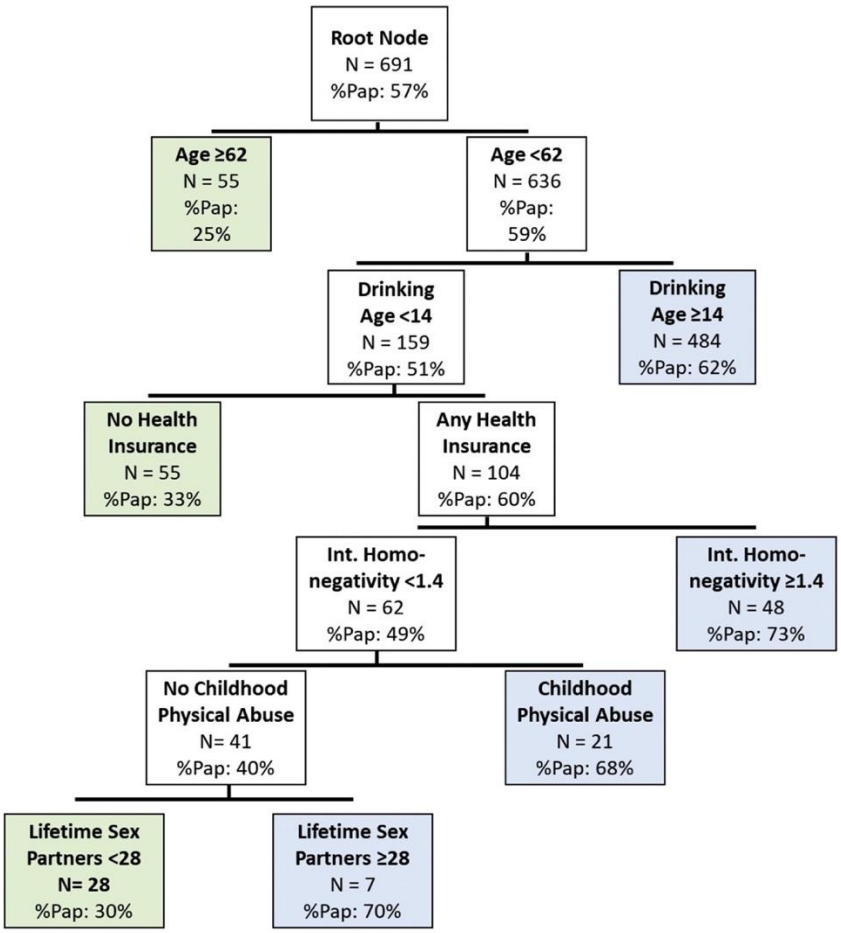

Figure 2.1: CART model from Greene et al. [26] © (https://doi.org/10.1016/j.pmedr.2018.11.007). Figure re-used under the Creative Commons Noncommercial-No Derivatives license (https://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 2.1: Subgroups characteristics from Greene et. al. [26] decision tree, predicting past year pap-tests

Leaves	Past year pap-test	Characteristics
1	25%	Age ≥ 62
2	30%	Age < 62, Drinking age < 14, Has health insurance, Internalized Homonegativity scale < 1.4, No childhood physical abuse, Lifetime sex partners < 28
3	33%	Age < 62, Drinking age < 14, No health insurance
4	62%	Age < 62, Drinking age ≥ 14
5	68%	Age < 62, Drinking age < 14, Has health insurance, Internalized Homonegativity scale < 1.4, Childhood physical abuse

6	70%	Age < 62, Drinking age < 14, Has health insurance, Internalized Homonegativity scale < 1.4, No childhood physical abuse, Lifetime sex partners ≥ 28
7	73%	Age < 62, Drinking age < 14, Has health insurance, Internalized Homonegativity scale ≥ 1.4

2.1.2 Literature search - use of decision trees in intersectionality

The application of decision trees has expanded into intersectionality research, where the resulting "leaves" represent intersectional groupings. Notably, a decision tree may not identify all intersectional groups possible from a theoretical perspective, but rather from a data-driven perspective will use given input variables to create enough intersectional groups to predict the outcome. A literature search was conducted of intersectionality studies using decision tree methods, and studies which reference intersectionality are presented in Table 2.2. The search yielded seven studies using two decision tree methods, CART and CHAID. Both these methods function by building single decision trees, and they are distinguished by their splitting criteria used to build the trees. CART is able to incorporate both continuous and categorical data as potential splitting variables and outcomes, whereas CHAID can only use categorical variables. Another distinction between the two is that CHAID allows for multiway splits (a parent node can split into more than two child nodes), while CART only performs binary splits.

Table 2.2: Intersectionality studies using decision trees

Study	**Year published**	**Outcome type (prevalence for binary outcomes)**	**Decision tree method**	**Sample size**
Shaw et. al. [27]	2012	Binary (common: 12%)	CHAID	211,736
Cairney et. al. [28]	2014	Binary (common: 24%)	CART	1,213
Zufferey [29]	2016	Binary (rare: 0.25%)	CART	775,000
Dey et. al. [30]	2018	Binary (common: 86%, 45%, 69%)	CART	5,565
Sridharan et. al. [31]	2018	Binary (common: 51%)	CART	5,666
Villanti et. al. [32]	2018	Binary (common: 27%)	CART	9,110 6,338
Greene et. al. [26]	2019	Binary (common: 57%)	CART	691

The earliest work used exhaustive CHAID analysis to identify combinations at the intersection of gender, race, age, and disability type, that best predict reporting harassment as a form of discrimination. [27] The authors describe this as a data mining approach, that can account for interactions between variables and create the best predictions. The sample size was 211,736, and 34 subgroups were identified, varying in sample size from 285 to 26,840. The authors display the risk of the outcome as a percentage, for groups 1 to 34. They describe in detail what characteristics make up the 5 highest and lowest risk groups, described by CHAID as "end groups". They describe these end groups as potential targets for further qualitative work, to identify further details on experiences and processes. To assess the model created, the authors state the percent risk of false classification (12%) and risk for cross-classification (12%), and they use this to suggest that results may be replicable in other samples.

Another study used CART analysis to assess the social determinants of accessing mental health service among those with mood or anxiety disorders, using linked Canadian Community Health Survey data. [28] This study used eight input variables, and with a sample size of 1213 participants, 6 terminal nodes were identified. The authors interpreted the model by walking through the splitting criteria. They report overall fit of the model by its sensitivity and specificity. The authors pair their CART analysis with a main effects logistic regression, including non-linear variables for age. The CART analysis identified complex interactions that were not visible from their regression analysis. The authors state their perceived benefits for using CART specifically for intersectionality research include that it doesn't make any assumptions about the distributions of variables or their relationships (e.g., not all interactions are linear), and can identify "complex or unsuspected interactions". They describe CART as a tool of interest for policy and care providers, to identify groups that are most at risk or under-served. They also acknowledge that there are some limitations, primarily the selection of cut points for continuous variables, which may or may not be relevant to actual policy or practice. Additionally, they state that it is more of an exploratory technique, because it is not capable of hypothesis testing.

Another study used Swiss National Cohort data, with approximately 775,000 lines of data, to assess mortality among migrant populations. [29] They conducted an analysis with CART, and then a "confirmatory analysis" using regression models with some interaction terms. They describe this as an inductive method and found that the confirmatory analysis supported the results found by CART. Fifteen categorical variables were inputted into the model, and the resulting tree presented 47 terminal nodes. They highlight intersectional effects, where splits create unique groupings. Similar to other studies, they state that the advantage of CART is the detection of interactions specific to particular groups. Furthermore, they clarify that this method is an exploratory analysis that requires further statistical analysis, such as regression modelling, to confirm the identified patterns.

A recent study used CART to understand the interaction of social determinants for maternal healthcare utilization, within a rural area of India. [30] With a sample size of 5,565, they created 3 different CART models using different binary outcomes (pregnancy registration, antenatal care in third trimester, and institutional delivery), and six different input variables. The three models produced four to six terminal nodes. For interpretation, the authors walked through the tree structure, and identified interactions visually. They identify the strength of CART as the ability to identify at-risk subgroups in the population, and the identification of specific interactions that can be used to guide policy and address inequities. As well, CART will consider multiple memberships or "inequities" at the same time. Their stated limitation is that there aren't estimates of the strength of the determinants or interactions. As well, a large enough sample size is needed to identify sub-groups. They warn that if no stopping criterion is used, CART may continue splitting groups until they are too small and not relevant. They also identify that because they predetermined the categories for certain variables (e.g. creating a binary variable from a continuous measure), this adds an "analytic bias", and if categories had been created differently, this may have affected splitting. A similar study using the same data set was conducted for solely the binary outcome for if the women had received any antenatal care. [31] A key difference in this article was that the CART regression was then paired with a multilevel model, of individual and district-level effects. Based on the

results of the CART model, cross-level interaction terms were included in a second set of models. The authors once again state that the tree method is exploratory, and not for causal inference.

Similar to the stratified regression analyses that have been used in other intersectional works (for example separate regression analyses for male and female), a study looking at cigarette and menthol cigarette smoking in American young adults conducted separate CART analyses for two different age groups (18 to 24 years and 25 to 34 years). [32] The authors viewed this as incorporating intersectionality by allowing potential predictors to differ between the age groups. This was paired with stratified main effects logistic regression analyses including the same predictors. When comparing results between the logistic regression models and corresponding decision trees, the CART models would create splits only on variables that were identified as significant by the regression, but did not always use all the significant variables. For example, for menthol cigarette smoking in the younger age group, all variables identified as significant from the regression (sex, race, education, and region) were used in the tree, while for the older age group only two variables (race and education) were used to build the tree, but sex and region were still significant in the adjusted logistic regression model. While the effect of sex was smaller in the older age group than the younger age group (OR of 1.56 versus OR of 1.69), the effect of region was actually greater (OR's of 0.69, 0.73, and 0.44 versus OR's of 0.81, 0.91, and 0.60), therefore splitting variables that were significant in the main effects logistic regression but not included the CART models weren't necessarily excluded simply because of a required main effect size threshold. The authors made no comment on the difference in results between the two types of methods, but stated that the CART analysis is a good way to identify "risk profiles" that can be used to guide policy.

Finally, the previously mentioned study from Figure 2.1 looking at the probability of cervical cancer screening among sexual minority women used CART analysis. [26] The authors used intersectionality theory to select the variables to be inputted into the model, including race/ethnicity, income, employment status, and experiences of discrimination. The authors interpreted the fit of the model by reporting accuracy of the model and

comparing it to the root node error. They found that the accuracy was 64.8%, which was an improvement over the root node accuracy of 56.7%. They interpreted this as a moderate accuracy and concluded that their included variables do not completely account for differences in cervical cancer screening between groups. The sensitivity, specificity, positive predictive value, and negative predictive value were also reported. They state that CART can be applied to see how multiple factors intersect to affect risk, but that once again the method does not admit causal interpretation, and can rather be used for hypothesis generation.

2.1.3 Use of decision trees in epidemiology

Because other works may have similar goals to intersectionality within health research, it is also important to look at the use of decision trees in works surrounding health and interacting social determinants of health. Firstly identified was Conditional inference trees (CTree) as a method of interest that has not yet been explicitly applied to an intersectionality study, but could be a potential methodological option. Conditional inference trees are similar to CART in that they can handle both continuous and categorical variables, however are distinguished by incorporating statistical hypothesis testing into building decision trees, and splits are given p-values. [33] Wolfson and Venkatasubramaniam [24] suggest that “the simplicity and inferential focus of conditional inference tress make them an appealing option for epidemiologists”. Compared to other decision tree methods, the inclusion of statistical inference has been suggested as a way to possibly minimize the issue of over-fitting. As well, the selection bias of CART to split on continuous variables is potentially minimized for CTree by a two-stage splitting process, which separates the identification of variables significant to the outcome from identification of the splitting point for each variable. This minimizes the bias created when continuous variables have more opportunities to provide splits than categorical variables. [33] One example study looked at the risk of intimate partner violence amongst 268 men and 299 women, by constructing two separate conditional inference trees. [34] The authors’ stated advantages over regression models included no assumption of linearity of effects, and less potential overfitting. The authors used only

one predictor variable, baseline physical aggression, to predict physical aggression at follow up. This allowed for the establishment of cut-points in the baseline measure to define risk groups. They found that among women, three terminal nodes were identified using the predictor, which they labelled as low, moderate, and high risk. Among men, two terminal nodes were identified (low and high risk). The decision tree was assessed using sensitivity, specificity, negative predictive value and positive predictive value. This was explained as a way to assess the relevance of cut-offs identified from a data-driven approach, to clinical practice. The authors found that their results could suggest clinically-significant cut-offs to use in clinical practice.

A second popular decision tree method in epidemiology that has yet to be applied to intersectionality research is random forests. Random forest models are created by fitting multiple decision trees from bootstrapped subsamples of the data and combining results from multiple trees together. [35] This method aims to address issues of over- or under-fitting in other decision tree methods. Because multiple trees are combined together to create a random forest model, unlike CART, CHAID or CTree, there is no single tree that can be observed and used to identify splitting variables or final subgroups. Instead, to identify if a variable is relevant to the outcome, the "variable importance measure" assesses the average performance of a variable across the multiple trees. There is more than one way to calculate variable importance, but the basic construct is that variables with high variable importance improve the fit of the decision tree, for example by contributing to the accuracy of the model. This measure is interpreted usually without statistical testing and compared as a relative measure between variables. One example study used this method to assess biological, behavioral, and social determinants associated with self-related health, citing decision tree analysis as an opportunity to use the social-ecological model of health, because these different determinants are acknowledged as possibly interacting with one another. [36] The random forest results were described by the variable importance measure for each variable. For example, they found that physical activity, income and education were the most important variables for predicting the outcome. One of the drawbacks of the random forest method is that because it is created by multiple trees, there is no one tree that can be visualized. Because

subgroup identification can be an important goal of using decision trees, the authors paired the analysis with a single classification tree. The single tree analysis resulted in 15 terminal nodes. The characteristics that made up these resulting subgroups, such as family income, physical activity, and education, were described. The authors' comparison between their single classification tree and random forest analysis was that the resulting cross-validated error from the single classification tree was 31% versus an out-of-bag error from the random forest model (average error when assessing model prediction against data not included in each bootstrapped sample) of 26%, giving random forest a slight advantage in terms of prediction accuracy.

2.1.4 Current literature assessing decision tree methods in epidemiological contexts

The benefit of using decision tree methods in intersectionality is that they can concurrently explore many positions or identities. This methodology can identify complex interactions and does not require assumptions about the variable distributions or relationships. From a health equity standpoint, it has been suggested as relevant to policy to identify groups that are most disadvantaged. Limitations of these methods include that there is no estimate of relative strength of variables or interaction effects, sufficiently large sample sizes are required for subgroup identification, trees may over-split and lose their relevance to policy, trees must choose cut offs for continuous variables even if the true effect is linear, and there is limited hypothesis testing. There are concerns around single decision tree methods being unstable in comparison to ensemble methods such as random forest, due to single decision tree models being more prone to drastically change with small changes in the sample data. [37, 38] Because of these limitations, decision tree methods for intersectionality research have been framed by some as a more "exploratory approach", and some studies have supplemented the inclusion of decision trees with traditional regression with interaction terms. Given these strengths and limitations, the next section reviews the current literature assessing the quantitative performance of decision tree methods, compared to traditional epidemiological regression.

To assess how well random forests may work for epidemiology compared to traditional

regression methods, random forest analysis has been compared against logistic regression using study data with a binary outcome. [39] The outcome was being overweight, defined by body mass index (BMI), and 14 sociodemographic and behavioral factors were included as input variables. No interaction terms were included, but separate analyses were conducted for men and women, for both the logistic regression and random forest analyses. Random forest was similar to logistic regression in terms of ability to classify members in the study sample as overweight or not overweight, when comparing true- and false- negatives and positives, and sensitivity and specificity. The two methods identified similar variables as important or significant. The authors stated that these results may be because the variables they chose have a more linear relationship to the outcome, or don't involve interactions. They state the benefit of random forest being that highly correlated variables (such as multiple nutrition factors) can be included in a random forest, but not in a logistic regression. As well, there is no need to pre-specify interaction terms, and creating a single decision tree can be useful for identifying subgroups that can be targetable from a public health perspective. They suggest that the use of random forest may be more beneficial than logistic regression for situations with a greater number of input variables.

Another study compared OLS regression with four machine learning algorithms: repeated linear regression, penalized linear regression, random forest, and neural networks. [40] Random forest was the decision tree method that they chose to incorporate, based on the fact that it has been widely used in the medical literature. They used each method to create predictive models for four continuous variables: systolic blood pressure, BMI, waist circumference and telomere length. Methods were compared via root-mean-square error and R-squared values. They created two regression models, one that was minimal and one that was theory based. Random forest did perform better for prediction than both regression models. Notably, a separate article had commented that the use of R-squared values to compare regression against machine learning methods has limitations. [41]

Finally, CTree, CART, and partially mis-specified regression have been evaluated for prediction of a continuous outcome, using a simulation study. [42] Data were generated

using three different scenarios: a linear regression with no interaction terms, a decision-tree-based outcome (where the outcome is created based off of decision rules), and a hybrid model, which included interactions for specific subgroups of the data. Both decision tree methods were compared to linear regression with no interaction terms, for these three data generation scenarios. Methods were assessed using mean squared error (MSE), calculated from independent test data sets. Using MSE to report prediction accuracy, they found that the decision tree methods performed better than the regression methods, under the decision-tree-based data generation scenario. For the regression-based data generation, the regression method was a better predictor. For the hybrid data generation, the three methods were found to have similar MSE's. Additionally, between CART and CTree, they found that the predictive accuracy of CTree improved with increasing sample sizes from n=30 to n=5000, compared to CART, where improvements plateaued by n=3000. The number of terminal nodes created by CTree increased over increasing sample sizes, to over 200 terminal nodes by n=5000, while the number of terminal nodes resulting from the CART models remained as less than 25 at n=5000. Results from this simulation study demonstrate that there are definite differences in prediction between decision tree methods, and that when compared to regression methods, decision trees were better predictors under circumstances with non-linear interactions.

2.1.5 Summary of decision trees and intersectionality, application to current study

Specifically reviewing the utility of decision trees in epidemiology, Wolfson and Venkatasubramaniam [24] outline three uses for decision trees in epidemiology. The first is for "explanatory modelling", where decision trees can be used as a "variable selector". Here, variables used in the splitting process are acknowledged as those important to the outcome. Decision trees can also be read to understand how a variable may affect the outcome (although this would not be true for random forest, which does not produce a visual tree diagram), especially in the presence of non-linear effects. The second use is for outcome prediction. They note that the limitation here is that sometimes predictions for decision trees can be subject to change with small changes in the data. Methods like

random forest can counteract this by creating multiple trees to prevent overfitting, but lose the interpretability of single decision tree methods. Another limitation is that if the relationship between the explanatory variable and outcome is truly linear, then a regression model will perform better for predictions, because for a tree to make an equivalent prediction, it would have to split many times. The third use is for subgroup identification, which in the context of public health or health equity, can help identify subgroups to be targeted for prevention efforts or treatment.

Resultantly, the evaluation of the decision tree methods in the current thesis addresses the three potential uses for decision trees outlined by Wolfson and Venkatasubramaniam: prediction, explanatory modelling, and subgroup identification. The main outcome, prediction accuracy, addresses how well the decision tree methods perform prediction. For "explanatory modelling", the percent of iterations that variables are correctly identified as important to the outcome is assessed. Finally, to understand subgroup identification, the number of terminal nodes or "leaves" is recorded.

2.2 MAIHDA

2.2.1 What is MAIHDA

MAIHDA (Multilevel analysis of individual heterogeneity and discriminatory accuracy) has recently been proposed as an alternative to traditional regression, to describe outcomes for many intersectional groupings. Specifically, it aims to address the following issues with traditional regression approaches: "scalability, model parsimony, reduced sample size in some intersectional strata, and occasionally, issues of interpretability." [43] The original approach by Evans et. al. [43] uses multilevel models with random intercepts, with individual-level characteristics as fixed effects, no fixed-effect interaction terms, and strata or clusters defined as each intersection. Combinations of the fixed effects form the intersections, therefore membership in the fixed effects fully determines which stratum or intersection an individual belongs to. The fixed effects are interpreted as the main "additive" effects, and the intersection residuals represent intersectional effects, or departures from additivity, and significant residuals can be easily identified as

significant intersectional effects. The variables inputted to create intersections must be categorical or binary, to allow for creation of distinct intersectional groups for the clusters. Models are fitted using Bayesian estimation techniques, with null priors.

Compared to traditional regression models, MAIHDA is suggested as a more parsimonious way to include many intersections, because rather than the number of interaction terms required increasing geometrically with every added social position, for MAIHDA the number of fixed effects increases linearly, with only one extra fixed-effect term required for each additional social position. [43] MAIHDA addresses issues of low sample size in certain intersections by adjusting residual estimates according to sample size of the intersection. The intersection residuals are shrunk towards the population mean with a weighting according to sample size, where a smaller intersection will be weighted more towards the mean. This is seen as preventing the residuals estimated for smaller intersections from being erroneously identified as larger than expected, due to extreme outliers. [43]

2.2.2 Review of MAIHDA studies

A review of current published studies using the MAIHDA methodology was conducted. Table 2.3 outlines the studies and their outcome types, sample sizes, and the number of intersectional positions and final groupings created. There are variable applications with both continuous and binary outcomes and a large range in the total number of intersections, but overall the number of intersectional variables included is notably greater than those in the typical intersectionality literature applying regression or uni- or bi-variate analyses.

Table 2.3: Studies using MAIHDA

Study	**Year published**	**Outcome type (Prevalence for binary outcomes)**	**N**	**Number of intersections**
Evans et. al. [43]	2018	Continuous	32,788	2*3*4*4*4=384
Fisk et. al. [44]	2018	Binary (rare: 0.22%)	2,445,501	2*2*3*2*2*2 = 96

Hernandez-Yumar et. al. [45]	2018	Continuous	14,190	2*3*3*3*2 = 108
Evans and Erickson [46]	2019	Continuous	15,388	2*7*2 = 28
Persmark et. al. [47]	2019	Binary (prevalence not provided)	6,846,106	2*3*3*2*2 = 72
Persmark et. al. [48]	2019	Binary (prevalence not provided)	43,409	2*4*3*3 = 72
Kiadaliri and Englund [49]	2019	Binary (rare: 3.5%, 0.5%, 0.2%, and 0.2%)	342,542	2*2*3*3*2*2 =144
Wemrell et. al. [50]	2019	Binary (rare: 5.6%)	4,334,030	2*5*2*3*2 = 120

The primary article looked at the continuous outcome of BMI, to identify differences across intersectional strata, defined by five variables: gender, race/ethnicity, income, education, and age. [43] This resulted in 384 unique intersectional groups, for which each was considered a stratum for the random effects. Table 2.4 is an example table published in this article. Here the “Null Model” includes only the random intercepts, and no fixed effects, and the full “Main Effects Model” includes all fixed effects, as well as the random intercepts for each intersection. As can be seen from these results, the inclusion of the main effects explains some of the stratum-level effects, as it reduces from 1.823 to 0.643. The remaining between-strata variation is displayed as a percentage, where 35.27% of the between-strata variation was unexplained by main effects. Groups with significant residuals are interpreted as having greater or lesser outcomes than expected from additive effects alone, also known as interaction or intersectional effects.

Table 2.4: Results from Evans et. al. [43] MAIHDA analysis for BMI (kg/m^2)

	Null Model Estimate (95% CI)	Main Effects Model Estimate (95% CI)
Fixed Effects		
Intercept	28.126 (27.965, 28.293)	26.858 (26.433, 27.288)
Gender		
Male (reference)		–
Female		0.081 (−0.149, 0.316)
Race/Ethnicity		
White Non-Hispanic (reference)		–
Black Non-Hispanic		1.791 (1.511, 2.066)
Hispanic/Latino		0.659 (0.383, 0.941)
Education		
Less than high school (reference)		–
Completed high school		0.087 (−0.255, 0.433)
Some college no degree		−0.240 (−0.591, 0.123)
College degree or more		−0.813 (−1.167, −0.460)
Income (% Poverty Threshold in 2000)		
Low income (Below 100%) (reference)		–
Low-middle income (100%–199%)		−0.066 (−0.370, 0.245)
High-middle income (200%–399%)		−0.258 (−0.574, 0.060)
High income (400% or more)		−0.584 (−0.953, −0.210)
Age		
18–29 years (reference)		–
30–44 years		1.282 (0.944, 1.624)
45–59 years		1.814 (1.477, 2.152)
60 + years		0.523 (0.184, 0.862)
Random Effects		
Strata	1.823 (1.503, 2.196)	0.643 (0.488, 0.826)
Individual	34.506 (33.984, 35.035)	34.511 (33.977, 35.047)
Percent of Between-Strata Variation Unexplained by Main Effects		**35.272%**

Note: 95% Credible Intervals in parentheses. *P*-values are associated with frequentist approaches and are not available for Bayesian estimations.

Reprinted from Social Science & Medicine, 203, Clare R. Evans, David R. Williams, Jukka-Pekka Onnela, S.V. Subramanian, A multilevel approach to modeling health inequalities at the intersection of multiple social identities, 70, Copyright (2018), with permission from Elsevier. ©

The original Evans et. al. [43] paper did not explicitly report on the number of significant interactions, instead focusing on general patterns. But subsequent papers have used MAIHDA to identify which intersections have the overall best and worst outcomes, and which intersections report the highest and lowest intersectional effects. The authors Fisk et. al. [44] reported from their analysis that three strata had significant interactions

according to the residual estimates, which they state is what would be expected to be significant (out of 96 strata) due to chance. Hernandez-Yumar et. al. [45] found that 9 out of 108 intersections were significant, Kiadaliri and Englund [49] found 6 out of 144 were significant, while Evans and Erikson [46] and Persmark et. al. [48] reported that none of the residuals were significant (0 out of 28 and 0 out of 72).

MAIHDA has been expanded for use with binary outcomes. [44, 47-50] By using a logistic regression, this creates issues for interpreting any interaction effects, which would be on the multiplicative scale for odds ratios. To use the additive scale, the authors Fisk et. al. [44] used the predicted log-odds to create the predicted probabilities (or incidence) in each stratum, and compared the expected and predicted probabilities. Figure 2.2 presents the identification of significant interaction effects by Fisk et. al. [44], where the significant interactions are calculated by significant differences in the predicted outcome for an intersection between main effects alone, and main effects plus the residual estimate. Persmark et. al. [47, 48] and Kiadaliri and Englund [49] similarly used the logistic model to calculate absolute risk.

Incidence of Chronic Obstructive Pulmonary Disease during 2011 for people aged 45-65 residing in Sweden on Dec 31st 2010, by intersectional strata. Predicted incidences and their 95% CIs based on the total effect (intersectional effects and main effects) and main effects only, in model 3. Interaction effects calculated as total effect minus main effect. Intersectional strata were calculated by categories of age, gender, income based on tertiles in the whole population aged 45-65 years, education, living alone and immigration status. In this table only the five strata with the most negative (protective) and the most positive (hazardous) interaction effects are shown. Intersectional strata are ordered according to their interaction effects with the lowest first and increased interaction effects in descending rows. Strata with 95% CIs excluding 0 are bold. For a full table showing data for all 96 intersectional strata, see Table A2 in Appendix and Figs. 3 and 4.

Age		Gender		Income			Education		Living alone		Immigrant		Model 3					
													Total		Main effects		Total - main effects	
45-54	55-65	Male	Female	High	Medium	Low	High	Low	No	Yes	Yes	No	Incidence	95% CI	Incidence	95% CI	Interaction	95% CI
The five intersectional strata with the most negative (protective) interaction effect																		
✓			✓			✓	✓			✓	✓		0.92	0.77 – 1.07	1.06	0.92 – 1.23	-0.15	-0.35 – 0.06
✓			✓		✓		✓			✓	✓		0.59	0.46 – 0.75	0.72	0.61 – 0.84	-0.13	-0.28 – 0.04
✓			✓			✓	✓		✓		✓		0.45	0.36 – 0.55	0.57	0.49 – 0.65	-0.11	-0.23 – 0.00
✓			✓			✓	✓		✓		✓		0.23	0.17 – 0.31	0.29	0.25 – 0.34	-0.06	-0.12 – 0.02
✓		✓		✓			✓			✓		✓	0.29	0.25 – 0.33	0.34	0.29 – 0.40	-0.05	-0.12 – 0.01
The five intersectional strata with the most positive (hazardous) interaction effect																		
✓		✓		✓			✓			✓		✓	0.23	0.18 – 0.28	0.17	0.15 – 0.20	**0.06**	**0.01 – 0.11**
✓	✓				✓	✓		✓		✓	✓		0.51	0.37 – 0.72	0.45	0.38 – 0.53	0.06	-0.08 – 0.25
✓	✓			✓			✓	✓		✓		✓	0.39	0.29 – 0.50	0.32	0.27 – 0.37	0.07	-0.02 – 0.18
✓		✓			✓		✓			✓	✓		0.29	0.25 – 0.35	0.21	0.18 – 0.25	**0.08**	**0.03 – 0.13**
✓			✓		✓		✓			✓	✓		0.39	0.33 – 0.45	0.26	0.22 – 0.30	**0.13**	**0.07 – 0.20**

Figure 2.2: Results from Fisk et. al. [44] © (https://doi.org/10.1016/j.ssmph.2018.03.005), presenting significant intersectional effects in bold. Figure re-used under the Creative Commons Noncommercial-No Derivatives license. (https://creativecommons.org/licenses/by-nc-nd/4.0/)

Another analysis by Evans and Erickson [46] applied MAIHDA to a longitudinal study, by assessing changes in depression scores over two different waves of a longitudinal dataset. The three outcomes were wave 1 scores, wave 4 scores, and the difference between wave 1 and 4. The authors also incorporated a continuous variable into the analysis by controlling for age, centred at zero.

2.2.3 Measures of discriminatory accuracy

Discriminatory accuracy aims to understand how well the chosen social positions or identities are actually able to predict and account for variation in the outcome. [51] Evans et. al. in the original article [43] does not directly reference discriminatory accuracy, but it has become a focus of subsequent studies applying MAIHDA. Articles include calculations of the intra-class correlation coefficient (ICC), where the ICC is calculated for both the null model (random intercepts and no main effects), and the model which includes main effects. [44-49] The ICC of the null model is seen as representative of the total explanatory power of the intersections for explaining variation in the outcome, and this explanatory power can include additive effects and intersectional effects. The ICC of the model fitted with fixed effects is seen as the remaining variation explained by intersectional, or interaction effects. [44] To interpret the ICC values, the authors Fisk et. al. [44] acknowledge that there is no set scale for these values, and suggest using the same scale used for psychometric tests (where ICC is expressed as a percentage): “non-existent (0–1), poor (>1 to ≤ 5), fair (>5 to ≤ 10), good (>10 to ≤ 20), very good (> 20 to ≤ 30), excellent (> 30)”. Accordingly, a poor ICC is seen as indicative that the chosen positions or identities used to form the intersectional stratum should not be acted on at a public health or policy level, because they contain too much individual heterogeneity to be effective targets. [44]

2.2.4 MAIHDA main effects

The interpretation of the main effects or “additive effects” estimated by MAIHDA is important, because it is based off of these effects that an intersection is evaluated as experiencing significant interactions. Interpretations of the meaning of MAIHDA

additive effects has varied. From the original article by Evans et. al. [43], these are simply described as “main effects” representative of the additive model. Other interpretations have been ambiguous on the meaning of main effects [44, 48, 49], or have simply made no interpretation of the main effects at all [47]. The third study to apply MAIHDA interpreted the coefficients from the MAIHDA model as one would for a traditional regression model, that assumes no missing interaction terms or model misspecification. [45] For example, they interpret the intercept as “The intercept, β0 measures the predicted BMI of the stratum defined when all the dummy variables equal zero (i.e., the reference individuals: 18 to 35-year-old males, with high income and high education and who cohabit).”. [45] An example interpretation of the main effect coefficient for gender is, “The results show women had an average BMI 1.16 units lower than that of men, having controlled for the other variables.”. [45] This interpretation may be the typical interpretation of “additive effects” for traditional regression models and intersectionality research, but further publications have clarified this interpretation does not apply to MAIHDA. In further articles, Evans has clarified that the interpretation of the main effects is not as holding all other variables constant, but is rather an average effect of each variable. For example, they state that for MAIHDA, “the parameter for “black” represents the average difference between black and white respondents, inclusive of all genders. In a model [traditional regression model] inclusive of interaction parameters, as in Table 3, the parameter for “black” represents something else entirely—the average difference between black males and white males.”. [52] Here, the effect of a variable according to MAIHDA is the average effect in the overall population, inclusive of potential interaction effects, and is distinguished from the typical interpretation of additive effects in regression models including interaction terms.

In a commentary by Lizotte, Mahendran, Churchill, and Bauer [53], a short simulation was used to demonstrate that the provided definition of MAIHDA main effects is not a sufficient explanation. Our commentary demonstrated that rather than being the population average effects [52], MAIHDA main effects fall under a different definition; main effects represent the average effects created from a pseudo-population, where clusters or intersections are equally weighted. In this commentary, we argued that this is

the true definition of the MAIHDA main effects, and that this definition falls outside of what is typically interpreted in intersectionality research as "additive effects", which are usually described as effects without the existence of interaction. An example of the difference between traditional additive effects, population average effects, and average effects across equally weighted clusters is given below.

The following is a data generating scenario where x1, x2, and x3 are binary variables, resulting in 8 possible intersections:

$$y = x1 + x2 + x3 + x1*x2$$

where P(x1=1) is 0.7, P(x2=1) is 0.7, and P(x3=1) is 0.5. Assume that the true effect of each variable (x1, x2, and x3) and the interaction term (x1*x2) is 1. The working model that will be fitted for MAIHDA is:

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \mu_j + e_{ij}$$

where for each stratum j, μ_j represents the effect of the random intercept, and e_{ij} is the individual error term. As explained previously, the MAIHDA working model includes no interaction terms.

The expected effect of x1 according to MAIHDA will be 1.5, because for half of the possible intersections where x1=1, the effect of x1 is 1 (when x2=0). For the other half of the intersections where x1=1 (when x2=1), the effect of x1 is 2 (because of the interaction of x1*x2). Therefore (0.5*1)+(0.5*2)=1.5. This is different from the population-level average effects, where the expected effect of x1 would be 1.7. This is because for 70% of those for who x1=1, x2=1 and the effect of x1 is 2. For 30% of those for who x1=1, x2=0, and the effect of x1 is just 1. Therefore (0.7*2) + (0.3*1) = 1.7. Finally, if MAIHDA main effects were only representative of the additive effects (an effect where no interaction is present), then the effect of x1 would simply be equal to 1. There calculations demonstrate that differences in the definition of main effects can lead to different interpretations of the results. Understanding the true interpretation of the main effects is relevant because these main effects form the baseline for determining which intersections experience significant interactions.

2.2.5 Current literature assessing MAIHDA

Comparison between MAIHDA and traditional regression has been conducted using secondary data analysis to assess differences in how the two methods identify significant interactions, for both continuous and binary variables. [52] Traditional regression models were fitted with main effects and all possible interactions. Each of the outcomes had a sample size of approximately 14,000. For each outcome, a varying number of possible intersectional strata were considered, from 6 (created by gender and race) to 91 (created by gender, race, immigrant status, parental education, income, and sexual identity). Generally, MAIHDA identified fewer significant intersections than the corresponding traditional regression model would identify significant interactions. For example, for the binary outcome of fair/poor health, 6 interaction terms were significant from the fixed regression model, but 0 intersections were significant from the MAIHDA model. Evans provides several explanations for why MAIHDA performs more conservatively when identifying significant intersections when compared to traditional regression, including 1) differences in the estimation techniques (frequentist versus Bayesian), 2) MAIHDA adjusts estimates for each intersection towards the grand mean based on intersection sample size, resulting in more conservative intersection estimates for small intersections, and 3) that MAIHDA and conventional methods make "fundamentally different comparisons" when calculating main effects, as explained in the previous section, resulting in intersections requiring to be detected as significant interactions "particularly intense effects in the expected direction or by breaking with general patterns all together".

2.2.6 Summary of MAIHDA and application to current study

MAIHDA has been suggested as the "gold-standard" for studying health disparities [54], that leads to "improved mapping of the risk heterogeneity of and socioeconomic inequalities" in different health outcomes [47]. With its ability to assess a large number of intersections and combinations of social positions with mixes of marginalization and privilege, it has thus been rapidly adopted as a methodology. No existing studies using MAIHDA have compared its predictive performance to traditional regression for the

accuracy of outcome predictions for each intersectional group. Given that one of the stated benefits of MAIHDA is the adjustment of estimates based on the sample size of intersections, which can supposedly lead to improved predictions, prediction accuracy compared to traditional regression or cross-classification should be evaluated. The current thesis uses simulated data to evaluate MAIHDA predictions at the intersection-level. Given the varying interpretations of the meaning of the main effect estimates created by MAIHDA, and their potential impact on the identification of significant intersectional effects [52], secondary outcomes of this thesis include assessing if effect estimates reflect the traditional definition of additive effects (the lower order effects in an interaction model), or if they fall under the definition proposed by Lizotte et. al. [53]. As well, this is an opportunity to further test the calculations suggested in the commentary, as these calculations were performed for only three simulated examples, and only on a continuous outcome.

2.3 Review of the literature and limitations

Applications beyond traditional regression and uni- or bi-variate approaches have been touted to have many benefits for studying interaction and accommodating a larger number of intersections, which is of interest given the wider availability of population-level datasets for intersectional research. CART, CTree, random forest and MAIHDA have been separately studied in comparison to traditional regression. However, it is not clear which methods perform best for intersection-level predictions, or for variable identification. Most studies assessing prediction, with the exception of Venkatasubramaniam et. al. [42], compare methods using secondary data analysis, where the true underlying answer is not always known, unlike in simulations. Without known simulated variable effect sizes and outcomes, the validity of different methods cannot be established, as it is unclear which method approaches the actual "truth". Methods that perform similarly in secondary data analysis may just be similarly over- or under-fitting to the sample data. The current thesis will improve upon the existing literature by using simulated data with known true outcome estimates and known true variable effects, to assess how well the conventional methods (regression and cross-

classification) and novel methods (decision trees and MAIHDA) perform for intersection-level prediction, as well as variable identification.

The authors Kreatsoulas and Subramanian [41] in their review of how social epidemiology stands to benefit from the incorporation of machine learning, reference the concept of “no free lunch” [55], which in this context refers to there not being any one methodology or algorithm that is best suited for prediction in every data scenario. This is why this simulation study explored a multitude of scenarios, varying by sample size, input types, and outcome types, with parameters selected based on the dataset qualities and outcomes relevant to intersectional research. The existing literature has not yet addressed this wider range of data analysis scenarios, which is necessary to understand which quantitative intersectionality methods may work best for different data scenarios.

Chapter 3

3 Methods

This chapter presents the methodology behind this simulation study. Firstly, the primary and secondary study objectives and outcomes are presented in Section 3.1. Section 3.2 outlines the eight quantitative intersectionality methods that were evaluated in this study. The combinations of parameters used to create the simulated data are described in Section 3.3, and procedures to create the simulated data are explained in Section 3.4. All analyses were run on R version 3.6.1. [56]

3.1 Study Objectives

The primary objective of this study was to evaluate the accuracy of population-level predictions created by descriptive quantitative intersectionality methods. Ten data generation models, varying by outcome and input type, were used to evaluate eight methods. The five outcome types were continuous, binary with a rare prevalence, binary with a common prevalence, negative binomial, and multinomial. The two types of inputs were either all categorical, or a mix of categorical and continuous. The ten data generation models were simulated for 4 sample sizes (2,000, 5,000, 50,000, and 200,000), for 1,000 iterations each, resulting in 10*4*1,000=40,000 total simulations. Results from these 40,000 iterations were summarized over each sample size and data generation model. The secondary objectives were to estimate the confidence interval coverage of the single-level regression methods and MAIHDA, the power of the single-level regression methods and MAIHDA, the ability of the decision tree methods to correctly identify important splitting variables, and the ability of the decision tree methods to estimate the number of distinct intersections. The data generation process included five variables (X1 to X5) with true effects on the outcome, one variable (X6) with no true effect, and interaction terms, resulting in 192 intersectional groups, of which 64 were truly different from one another in regard to the outcome Y.

The methods that were assessed are as follows:

1. Regression – best fitted: only the necessary/true interaction terms are included.
2. Regression – over-specified: all possible interaction terms are included in the model.
3. Cross-classification
4. MAIHDA (Multilevel analysis of individual heterogeneity and discriminatory accuracy)
5. CART (Classification and Regression Trees)
6. CTree (Conditional Inference Trees)
7. CHAID (Chi-square Automatic Interaction Detector)
    - only for models with all categorical inputs and outcome
8. Random Forest

Further descriptions of these methods can be found in Section 3.2.

The ten data-generation models used to assess the methods are as described in Table 3.1.

Table 3.1: Description of the ten data generation processes

	Outcome	Input variables
Model 1	Continuous	Categorical only
Model 2	Continuous	Mixed (Categorical and continuous)
Model 3	Binary – Common prevalence	Categorical only
Model 4	Binary – Common prevalence	Mixed (Categorical and continuous)
Model 5	Binary – Rare prevalence	Categorical only
Model 6	Binary – Rare prevalence	Mixed (Categorical and continuous)
Model 7	Multinomial	Categorical only
Model 8	Multinomial	Mixed (Categorical and continuous)
Model 9	Negative binomial	Categorical only
Model 10	Negative binomial	Mixed (Categorical and continuous)

Further description of the creation of these models can be found in section 3.4.

3.1.1 Primary outcome

The primary outcome was the accuracy of each method's intersection-level predictions, evaluated by mean squared error (MSE) and mean absolute percentage error (MAPE). Intersection-level predictions were defined as the prevalence or mean of outcomes in each of the 192 intersections. Note that accuracy was calculated at the intersection-level (for each of the 192 intersections), rather than by comparing individual-level predictions.

While predicting outcome prevalences or means at the intersection level may not be the only outcome of interest for descriptive quantitative intersectionality, it was chosen as the primary outcome for this study because it can be calculated across all the included methods. Secondary outcomes address the outcomes specific to certain methods.
For the continuous and negative binomial outcomes, the MSE for each method was calculated using the difference between the true population mean for each intersection, and the estimated population mean. Mean squared error was calculated as follows:

$$\boldsymbol{MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2}$$

where n is 192 (reflecting the 192 intersections), $\boldsymbol{\widehat{Y}_i}$ is the estimated mean for intersection $\boldsymbol{i}$, and $\boldsymbol{Y_i}$ is the true population mean for intersection $\boldsymbol{i}$. The true mean for the continuous outcome was known by using the same outcome generating formula (see section 3.4) as for the model dataset, without including a random error term. The true mean for the negative binomial outcome was known by using the same outcome generating formula, and not running it through the negative binomial sampling function.

For the binary and categorical outcome, accuracy was assessed using MAPE. The MAPE for each method was calculated using the difference between the true population prevalence for each intersection and the estimated population prevalence, divided by the true population prevalence.

$$\boldsymbol{MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\widehat{P}_i - P_i|}{|P_i|}}$$

For the binary outcomes (rare and common), the MAPE was calculated such that $\boldsymbol{\widehat{P}_i}$ is the estimated prevalence of the outcome Y=1 for intersection $\boldsymbol{i}$, and $\boldsymbol{P_i}$ is the true prevalence of the outcome Y=1 for intersection $\boldsymbol{i}$. The true prevalence of the outcome was known by using the same outcome generating formula, but not running it through the binary sampling function. For the multinomial outcome, three MAPE's were calculated for each method, one for each of the three possible outcomes. Therefore, MAPE was calculated for outcomes Y=1, Y=2, and Y=3.

For each of the 10 models and across the four sample sizes, the estimated MSE or MAPE for each iteration is presented in boxplots.

3.1.2 Secondary outcomes

The secondary outcomes were used to examine other outcomes of interest for quantitative descriptive intersectionality.

The regression methods (both over-specified and best-fitted) were assessed for confidence interval coverage, coefficient significance, and convergence. Confidence interval coverage was defined by the percent of iterations for which the confidence interval contained the true estimate for each variable. Coefficient significance was defined by the percent of iterations that the coefficient estimate was significant at an alpha of 0.05. While this is referred to as "power" over the course of this thesis, we acknowledge that the regression analyses were often under-powered to detect all relevant variables. For the over-specified model, results were only calculated for the variables that appear in the best-fitted model.

MAIHDA was assessed for confidence interval coverage and coefficient significance of the main effects. Because main effects in MAIHDA are defined as "additive effects" differently than in the intuitive definition in regular single-level regression, confidence interval coverage was calculated according to two definitions of main effects: the traditional definition in intersectionality, and the MAIHDA definition, as described by Lizotte et. al. [53]. Further description of these definitions can be found in section 3.2.3.

For the decision tree methods, CART, CTree, and CHAID were assessed for the average number of leaves in the final tree, and the average number of splitting variables used in the final tree, with 2.5^{th} and 97.5^{th} percentiles. The percent of iterations where each variable was used as a splitting variable was also reported. The random forest results were summarized by the average number of leaves in the random forest models. Because there is no set splitting pattern for random forest models, the average variable importance measure for each variable was reported, with 2.5^{th} and 97.5^{th} percentiles.

3.2 Description of eight quantitative intersectionality methods

3.2.1 Regression – best-fitted and over-specified

Best-fitted regression was shown as an analytic method to demonstrate a best-case example of a regression model, where only the necessary interaction terms are included. In reality, this would likely not occur because one does not know all relevant interaction terms *a priori*. In this study, these models included only the interaction terms X1*X2, and X3*X4*X5. In contrast, the over-specified regression method was purposely over-specified, and included all possible interaction terms. While the best-fitted regression represented a best-case scenario, the over-specified regression represented a more realistic scenario in intersectionality studies, where the underlying data structure is unknown, and therefore all possible interaction terms are specified. For the mixed input models this results in 64 estimated coefficients, and for the categorical input models, 192 estimated coefficients. The included interaction terms assumed linear interaction effects. The type of regression was as follows, depending on the outcome:

A. Continuous (normal) outcome: The continuous outcome was analyzed using a linear regression model via the R-core function “lm” which applies ordinary least squares.

B. Binary rare prevalence and binary common prevalence outcomes: The modified Poisson regression was used to analyze binary outcomes. [57] It is widely acknowledged that risk ratios have greater interpretability than odds ratios as a measure of association. [58] Therefore, while intersectionality studies applying regression to binary outcomes will often use logistic regression to produce odds ratios, an alternative option was chosen in this study. The modified Poisson regression is an application of Poisson regression that can be used for binary outcome data with rare or common outcomes, and produce coefficient estimates that are risk ratios, rather than odds ratios. Robust error variance is used to counteract the variance overestimation that occurs from applying a Poisson regression to binary outcome data. The binary outcomes were analyzed using the R-core function “glm” for

generalized linear models. Packages “lmtest” [59] and “sandwich” [60] were used to allow the sandwich estimator to correct the variances.

C. Categorical outcome: Multinomial logistic regression was conducted using the R package “nnet” [61], which uses neural nets to fit the multinomial log-linear models. This package was chosen over the alternative “mlogit” [62], which estimates multinomial logit models using maximum likelihood, due to the mlogit package often failing to converge the over-specified regressions.

D. Negative binomial outcome: Negative binomial regression is used for count data, especially when over-dispersed. The negative binomial outcome was analyzed using the “glm.nb” function from the R package “MASS” [61].

3.2.2 Cross-classification

Cross-classification was the univariate approach of taking either the prevalence or mean of the outcome for each of the unique intersections, with no further statistical adjustments. If there was a cell-size of zero for any of the intersections, this intersection was omitted from the MSE or MAPE calculations for cross-classification. The average number of omitted intersections is reported in the results section, as a reminder that not all intersections were included in the estimate of MSE or MAPE for cross-classification.

3.2.3 MAIHDA

MAIHDA (Multilevel analysis of individual heterogeneity and discriminatory accuracy) is a novel application of multilevel analysis for intersectionality, first introduced by Evans et. al. [43]. Here, intersections are defined at the group level with random intercepts, and for the fixed-effects (or main effects), each of the predictors used to create the intersections is included. No interaction terms are included amongst the fixed effects. The MAIHDA model can be represented as

$$y_{ij} = \beta\gamma_j + \mu_{0j} + e_{0ij}$$

$$\text{Level 2}[\mu_{0j}] \sim N(0, \sigma^2_{strata})$$

$$\text{Level 1}[e_{0ij}] \sim N(0, \sigma^2_{e0})$$

where i is each individual in intersection j, γ_j represents a vector of the intercept and main effect predictors, and β is a vector of the parameter values. The random effects are intercepts for each intersectional group (μ_{0j}). The additional term (e_{0ij}) is for individual-level error. Membership in the fixed-level effects therefore determines which intersectional group (level 2) a respondent belongs to. The interpretation is that the fixed effects represent additive effects, and the random intercepts will represent any additional "intersectional effects".

Similar to cross-classification, if there was a cell-size of zero for any of the intersections, this intersection was omitted from the MSE or MAPE calculations, and the average number of omitted intersections is reported with the primary outcome.

The original application of MAIHDA uses Bayesian models with null priors. However, due to computational power and time restraints, the analysis for this simulation was conducted using frequentist analysis. Appendix A demonstrates results from a short simulation comparing the estimation of main effects between frequentist and Bayesian models with null priors, for a continuous outcome model with five binary predictors. Results from this analysis validated the choice to use the frequentist model, given that the main effects estimates were extremely similar between the two approaches.

The types of multi-level regression were as follows, depending on the outcome:

A. Linear outcome: The continuous outcome was analyzed using linear multilevel regression, using the function lmer from the R package "lme4" [63].

B. Binary outcomes: The binary outcome was analyzed using multilevel Poisson regression, using the function glmer from the R package "lme4" [63]. The Poisson regression method was chosen instead of logistic, to be consistent with the use of the modified Poisson regression for the single-level regression methods (best- and over-

specified). A limitation of this choice is that the current packages in R do not allow for the use of the sandwich estimator to alter variances from glmer objects. Therefore, while intersection-level predictions for the primary outcome should be unaffected, the interpretation of confidence interval coverage and coefficient significance should consider that the variance estimates are likely highly conservative.

C. Categorical outcome: The following functions were attempted but could not run a multinomial multi-level regression with random intercepts only: multinom from package "nnet" [61], mlogit from the package "mlogit" [62], mblogit from the package "mclogit" [64], and clmm2 from the package "ordinal" [65]. Therefore, this was taken as a practical restraint and MAIHDA was not used to analyze the two multinomial outcome models. Alternatively, if users did want to run such an analysis on R, the likely solution would be to run a Bayesian multilevel model using the package "brms" [66].

D. Negative binomial: Negative binomial multilevel regression was run using glmer.nb from the "lme4" [63] package.

When considering confidence interval coverage of the main effects, two definitions were used to define the possible main effects estimands. Definition 1 is additive effects in the traditional intersectionality definition, which is when effects of each social position are not impacted by other positions (the effect of the variable in the presence of no interactions). Definition 2 is the average effect of the variable across equally-weighted clusters (or intersections). Presented in Table 3.2 is an example of the calculations for estimands using both definitions, according to the following formula used to generate the continuous outcome with categorical inputs:

$$Y = b_{intercept} + b_{1.1}\,(\text{if } X_1=1) + b_{1.2}\,(\text{if } X_1=2) + b_{1.3}\,(\text{if } X_1=3) + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6(\text{if } X_1=2 \,\&\, X_2=1) + b_7(\text{if } X_1=3 \,\&\, X_2=1) + b_8X_3X_4X_5 + e\ .$$

Table 3.2: MAIHDA estimand definitions

Estimand	Definition 1	Definition 2
intercept	$b_{intercept}$	The expected intercept from MAIHDA is a result of the other main effects. The calculation for the estimand is currently unknown.
x1.1	$b_{1.1}$	$b_{1.1}$
x1.2	$b_{1.2}$	$(b_{1.2} + (b_{1.2} + b_6))/2$
x1.3	$b_{1.3}$	$(b_{1.3} + (b_{1.3} + b_7))/2$
x2	b_2	$(b_2 + b_2 + (b_2+b_6) + (b_2+b_7))/4$
x3	b_3	$((b_3 + b_8) + b_3 + b_3 + b_3)/4$
x4	b_4	$((b_4 + b_8) + b_4 + b_4 + b_4)/4$
x5	b_5	$((b_5 + b_8) + b_5 + b_5 + b_5)/4$

Because MAIHDA has so far only been applied when intersections are defined by categorical variables, definition 2 was only calculated for models with categorical inputs. For all other models, confidence interval coverage was only calculated using definition 1.

3.2.4 Classification and Regression Trees (CART)

CART analysis was conducted using the function “rpart” from the package “rpart” [67]. CART is a binary decision trees method that can handle both continuous and categorical data. [25] To build a CART model, each variable is assessed for its ability to best split the data into two groups based on the outcome, as defined by a pre-determined splitting criterion (e.g. the Gini index). The variable that performs best according to this criterion is then used to split the data into two groups. The same process of searching for the next best splitting variable is repeated, performed independently for the two groups. This process is continued for each resulting subgroup, until a pre-determined stopping criterion. Stopping criteria can be based on a minimum sample size within a leaf (e.g. a final leaf can have no less than five respondents), or on an improvement criterion. [68] To avoid over-fitting the CART model to the data, cross-validation can then be used to trim the tree.

In this study, the CART splitting criterion for classification trees is based on the Gini rule. The node impurity (or heterogeneity of the outcome within the node) for node A is calculated as,

$$I(A) = \sum_{i=1}^{C} f(p_{iA})$$

where p_{iA} is the proportion respondents in node A whose outcome is "i". [68] The Gini index is the function "f" to measure of node impurity, calculated by f(p)=p(1-p). This is then used to determine the best splitting variable for a node, by calculating the splitting variable that results in the "maximal impurity reduction":

$$\Delta I = p(A)I(A) - p(AL)I(AL) - p(AR)I(AR)$$

where AL is the resulting left node, and AR is the resulting right node. The stopping criterion is based on when no further improvements in impurity reduction can be made. The splitting criterion for regression trees by CART is defined using sum of squares (SS), where $SS_T = \sum(y_i - \bar{y})^2$ is the sum of squares for the parent node. The splitting criterion is $SS_T - (SS_L + SS_R)$, where SS_R and SS_L are for the left and right resulting child node. A better split will have a larger difference between the sum of squares of the parent node and the child nodes.

10-fold cross-validation was performed (k=10 is the default for rpart). Pruning was used to select the complexity parameter with minimal cross-validation error. The complexity parameter is the improvement required by a split to be continued. The default value is 0.01. The minimum size of a node for it to be considered for splitting was 20, which is the default for rpart. Because the simulated datasets are assumed to be representative of the population probability of the outcome, no prior probabilities were specified.

3.2.5 Conditional Inference Trees (CTree)

Conditional inference trees were created using the ctree function from the R-package "partykit" [33]. Like CART, CTree is a binary recursive partitioning method. Unlike CART, CTree works in two stages, first to identify if variables are significant to the outcome, and then secondly, to find a splitting point for the selected splitting variable.

Whether or not variables are significant to the outcome is assessed using univariate regression models. If the global null hypothesis (that none of the available variables are significantly related to the outcome) is rejected, the variable with the greatest association to the outcome is selected as the splitting variable. [33] The global null hypothesis is assessed using p-values with Bonferroni correction for multiple testing, and the p-value is a parameter which can be altered. This study used an alpha of 0.05. The process is then repeated for the subsequent child nodes, until no further splits can be made. Other stopping criteria, like minimum node size, can also be implemented. Minimum node size was kept at the default value of 20.

3.2.6 Chi-square Automatic Interaction Detector (CHAID)

CHAID trees were created using the chaid function from the R-package "CHAID" [69]. CHAID is a non-parametric method that utilizes the significance values from chi-squared analysis as splitting criteria. [70] The best way to partition for each variable is selected, and then each of the best partitions for each variable is compared against one another, and the best of these partitions is used to divide the data into groups. Each of the resulting groups is then separately partitioned again. This method can only be applied for models with all categorical inputs and a categorical outcome (three out of the ten models assessed in this study).

3.2.7 Random Forest

Random forest trees were created using the tuneMtryFast function from the R-package "tuneRanger" [71], which calls from the package "Ranger" [72]. The random forest method for constructing decision trees combines decision trees with bootstrapping methods, with the goal of reducing over-fitting to the data. [73] To create a random forest model, trees are built from bootstrapped subsamples of the data, that are the same size as the original dataset. The parameter "mtry" in this package determines how many variables of the available input variables are assessed to determine the best splitting variable. At each node, the number of variables as specified by "mtry" are randomly

selected and assessed to find the next splitting variable, until the tree is fully grown at the pre-determined stopping criterion. The random forest model is therefore a collection of multiple trees. To combine this collection of trees to predict outcomes for a new dataset, each response is applied to each of the trees. For classification problems, the outcome predicted by the random forest is whichever outcome occurred most of the time (the mode) from the total trees. For example, if out of 500 trees, Respondent 1 is classified in a binary classification problem as A in 150 trees and B in 350 trees, the predicted outcome for Respondent 1 will be B. Regression problems utilize the mean value.

The random forest models were tuned using the tuning parameter mtry. Tuning allows for optimization of the model by adjusting parameters to improve model fit. The standard default for mtry is the square root of the number of input variables. For example if nine input variables are fed into the random forest model, the resulting mtry value is three. By tuning by a step factor of 1, the mtry is increased or decreased by this value, and then the out-of-bag error (average error when assessing model prediction against data not included in each bootstrapped sample) is assessed. If the improvement to out-of-bag error passes a threshold value (0.05), then the mtry value is again increased or decreased by the step factor. This iterative process is continued until improvement to the out-of-bag error does not pass the threshold value. The threshold value for improvement of 0.05 is based on the defaults for tuneRanger, but the step factor of 1 was selected, rather than the default of 2, given that there were a small number of input variables in the simulated datasets. Random forests were fitted with 500 trees, the default for the package, but this parameter could be altered or used for tuning in other applications.

The splitting criterion for the ranger package is similar to CART, where splits that results in the greatest decrease in node impurity are selected. [72] As with CART, node impurity is defined by the Gini index for classification problems, and response variance for regression problems. Because there are multiple different trees created by the random forest procedure, there is no one observable splitting pattern that can be assessed to definitively say if a variable is important to the outcome or not. Instead, the variable importance measure is used to represent this concept, where higher values are considered

more important. There is no limit to the range of possible values, but typically values are greater than zero. Variable importance for our analysis was determine by how much a variable contributes to decreases in node impurity. No minimum node size was set, the default values from the ranger package being 1 for classification and 5 for regression.

3.3 Description of simulation parameters and combinations

Table 3.3 describes the combination of parameters used to create the simulated datasets. By the combination of five outcome types and two input types, a total of ten different data generation scenarios exist. Each of these models was repeated for four different sample sizes. These 40 different types of models were each created with 1000 iterations, and iterations varied by effect sizes of the main and interaction effects. The following section describes the selection of these parameters.

Table 3.3: Parameter combinations for the creation of datasets

Number of options	Parameter	Description
5	Outcome types	1. Continuous 2. Binary variable: Rare prevalence 3. Binary variable: Common prevalence 4. Multinomial 5. Negative binomial
2	Combinations of input variables	1. Four binary and two categorical (categorical inputs only) 2. Four binary and two continuous (mixed inputs)
4	Sample sizes	1. 2,000 2. 5,000 3. 50,000 4. 200,000
1000 iterations	Effect sizes	1,000 iterations vary by effect size of main and interaction effects

3.3.1 Outcome types

Selected outcomes were based on common outcomes of interest in intersectional research, according to consensus by the thesis committee.

1. Continuous outcome: The continuous outcome was simulated to have a normal distribution.

2. Rare binary outcome: A rare binary outcome was simulated with the aim for the prevalence to be less than 5%. The simulated data for this outcome had an average prevalence of 3%.

3. Common binary outcome: A common binary outcome was simulated with the aim for the prevalence to be greater than 10%. The average prevalence was 15%.

4. Multinomial outcome: The multinomial outcome was simulated with three categorical responses, which were treated as nominal data. The three groups were created to have unequal prevalence's, with the average prevalence of groups Y=1, Y=2, and Y=3 being 17%, 33%, and 50%, respectively.

5. Negative binomial: The negative binomial distribution represents a count outcome, with a large number of zeros. This was chosen to evaluate the methods with a more extreme but still common form of data. The dispersion parameter of the outcome distribution was varied slightly using the value theta, where the outcome variance is defined by $mu + mu^2/theta$, and mu is the outcome mean. Theta was sampled from uniform distribution between 0.8 and 1.2.

3.3.2 Input types

Two combinations of input variables were selected: a set of only categorical variables, and a mixed set of categorical and continuous variables. This was chosen to reflect the possibility that different methods may perform better with different sets of predictors. For example, the bias towards continuous variables in decision trees can be better evaluated by using two different types of input variable sets.

Based on the systematic review [19], most published papers included intersections defined by very few social identity/position variables (two to three), likely because of the lack of methods that address the large number of intersectional groupings created from multiple variables. The literature search into MAIHDA and intersectional applications of decision trees however finds that these methods push towards including more intersectional variables in analyses. Therefore, six input variables were selected to create a total number of intersections that is more than the usual, but within the existing limits of the applications of decision trees and MAIHDA.

The mix of binary and categorical or continuous variables were selected based on the most common variables used in quantitative intersectionality research, according to the systematic review of the literature. [19] The simulated combination of input variables can be considered as analogous to the following variables: X1 as income, X2 as ethnicity, X3 as sex, X4 as post-secondary education, X5 as immigrant status, and X6 as age. Table 3.4 presents the prevalences of the simulated binary variables, compared to the Canadian Community Health Survey 2015/2016 [74] prevalences of their real-life counterparts in the Canadian national population data. X1 and X6, or income and age, were treated as continuous variables in the mixed input models, and as quartiles or tertiles in categorical input models.

Table 3.4: Simulated variables drawn from Canadian Community Health Survey (CCHS) prevalences

Variable	CCHS 2015/16 prevalence	Simulated Variable	Simulated Prevalence
Ethnicity (white vs. non-white)	22.5%	X2	20%
Sex (Female)	50%	X3	50%
Education (Post-secondary)	60%	X4	55%
Immigrant status	25%	X5	25%

Tables 3.5 and 3.6 present the two combinations of input variables, categorical inputs and mixed inputs. Inclusion of a mediation relationship between X3 and X4 reflects the reality that social positions are often correlated and can affect the likelihood of belonging to other social positions. For example, being male and high income could each have

individual effects on a health outcome, and the effect of being male could be partially mediated by income, if being male increases the probability of earning a higher income. The total number of intersections was 4*2*2*2*2*3 = 192. Because variable X6 had no true effect on the outcome, the total number of intersections that were distinct from one another regarding the outcome were 4*2*2*2*2 = 64. The resulting intersection sizes are such that there are some missing cells at n=2,000, but all cells are filled with greater sample sizes.

Table 3.5: Predictor combination of categorical inputs

X1	Categorical	P(X1=0) = 0.25 P(X1=1) = 0.25 P(X1=2) = 0.25 P(X1=3) = 0.25
X2	Binary	P(X2=1) = 0.2
X3	Binary	P(X3=1) = 0.5
X4	Binary	Mediation: P(X4=1 | X3 = 0) = 0.4 P(X4=1 | X3=1) = 0.7
X5	Binary	P(X5=1) = 0.25
X6	Categorical	P(X6=0) = 0.33 P(X6=1) = 0.33 P(X6=2) = 0.33

Table 3.6: Predictor combination of mixed inputs

X1	Continuous (split in quartiles to create intersections for prediction)	mean=0, variance=1
X2	Binary	P(X2=1) = 0.2
X3	Binary	P(X3=1) = 0.5
X4	Binary	Mediation: P(X4=1 | X3=0) = 0.4 P(X4=1 | X3=1) = 0.7
X5	Binary	P(X5=1) = 0.25
X6	Continuous (split in tertiles to create intersections for prediction)	mean=0, variance=1

3.3.3 Sample sizes

Four sample sizes were selected for the simulations: 2,000, 5,000, 50,000 and 200,000. The largest sample size is reflective of the availability of large population data sets like the Canadian Community Health Survey. The smaller sample sizes are reflective of the reality that many intersectionality papers, including those used for decision tree methods found in the literature review, use smaller datasets.

3.3.4 Effect sizes

The true effects of the variables were varied in magnitude and direction, to include a diverse set of possible scenarios, where effects can be both positive or negative, and interactions may have greater or smaller effect sizes than their main effects. A minimum effect size was selected based on power calculations, which are described in more detail below.

3.3.4.1 Power calculation for beta coefficients

Power calculations were performed to determine the minimum effect size for the beta coefficients. Separate power calculations were conducted for each of the five outcome types. The input variables for the power calculation were X1 to X6, where all variables were either binary or categorical, based on the predictor combination shown in Table 3.5. Two changes were made to the categorical inputs model that differed from what is shown in Table 3.5, that were justified based on the aim to remain relevant to intersectionality research. Firstly, the models for the power calculations were created and evaluated with main effects only, even though the models in the actual simulations include interaction terms. Presumably if an effect size is significant for an "additive effects" model (additive effects by the intersectionality definition, meaning no interaction), then it is still an important enough size for the detection of interaction terms. Elsewise, much larger effect sizes are required to detect interaction terms for Poisson and logistic models. Secondly, the input variables did not have the same distribution as in the actual simulation models. In the power calculation models, variables X1 to X3, and X5 and X6 were split in equally sized categories. Only X4 was not equally distributed, due to the mediation relationship between X3 and X4. The justification is that in ideal circumstances, calculating outcomes for each intersectional group would not be affected by intersection size, especially when those experiencing marginalization may belong to groups with smaller cell sizes.
Eighty percent power was defined as when in at least 80% of the models, all coefficients for X1 to X5 were significant at $p<0.05$. Power calculations were conducted at n=25,000, an intermediate value between the four sample sizes used in the simulation models

(n=2,000, 5,000, 50,000, 200,000). 100 iterations of each model were used to determine 80% power.

Table 3.7 presents the calculated minimum effect sizes and coefficient distributions for the five outcome types. For the linear outcome, the regression coefficients were on a scale with a possible range of negative to positive infinity, with a null value at zero. The sampling of positive and negative beta coefficients was centred around 1 and -1. Positive coefficients were selected from a truncated normal distribution with a minimum of "minimum effect size" and a maximum of (2 – minimum effect size). The negative coefficients were selected from a truncated normal distribution with a minimum of (-2 + minimum effect size) and a maximum of (-1*minimum effect size). For the binary, multinomial, and negative binomial outcomes, the coefficients were sampled by selecting a true RR, OR, or IRR respectively, and these values were then log-transformed to be applied to the outcome-generation process. The coefficients were therefore limited to a possible range of zero to positive infinity, with a null value at 1. The positive coefficients were selected from a truncated normal distribution, with a minimum of "minimum effect size", and a maximum of 1.8. Similarly, the negative coefficients were selected from a truncated normal distribution, with a minimum of 0.2, and a maximum of (1 – minimum effect size). The distributions had a standard deviation of 0.3.

Table 3.7: Coefficient sampling distributions

Outcome	Coefficient scale	Minimum effect size	Positive coefficient distribution (min, max)	Negative coefficient distribution (min, max)	Distribution standard deviation
Linear	Linear	0.06	(0.06, 1.94)	(- 1.94, -0.06)	1
Binary (rare)	RR	1.24	(1.24, 1.8)	(0.2, 0.76)	0.30
Binary (common)	RR	1.11	(1.11, 1.80)	(0.20, 0.89)	0.30
Categorical	OR	1.24	(1.24, 1.8)	(0.2, 0.76)	0.30
Negative binomial	IRR	1.09	(1.09, 1.8)	(0.2, 0.91)	0.30

3.4 Simulation procedures

3.4.1 Independent variable and effect size selection

The simulations for each of the 10 models were run as follows.

The random seed was set at the beginning of each iteration. The independent variables X1 to X6 were generated according to the distributions and probabilities shown in Table 3.5 or 3.6. The effect sizes for the beta-coefficients were then randomly selected from the distributions described in Table 3.7. For the categorical inputs, the beta-coefficients required are shown in Table 3.8, and for the mixed inputs, in Table 3.9.

Two interactions were included in all data-generation scenarios. One interaction was between variables X3, X4, and X5, creating a three-way interaction between binary variables. The second interaction differed depending on if the inputs variables were mixed or all categorical. For the mixed scenario, X1 (continuous) and X2 (binary) interacted when X1 was greater than 1 and X2 was equal to 1. This was a non-linear interaction based on a cut-off value, to maintain realistic expectations that interactions do not have to function linearly, although this is often assumed in regression models when fitting interaction terms. For the categorical-inputs-only scenario, X1 and X2 interacted if X1 was equal to 2 and X2 was equal to 1, or if X1 was equal to 3 and X2 was equal to 2.

Table 3.8: Categorical inputs coefficients

Variable	Coefficient	Coefficient distribution type
X1 = 1	$\beta_{1.1}$	Positive
X1 = 2	$\beta_{1.2}$	Positive
X1 = 3	$\beta_{1.3}$	Positive
X2 = 1	β_2	Positive
X3 = 1	β_3	Negative
X4 = 1	β_4	Negative
X5 = 1	β_5	Negative
Interaction: if X1= 2 & X2=1	β_6	Negative
Interaction: if X1= 3 & X2=1	β_7	Negative
Interaction: if X3=1 and X4=1 and X5=1	β_8	Positive
X6 = (0, 1, 2)	0	No true effect

Table 3.9: Mixed inputs coefficients

Variable	Coefficient	Coefficient distribution type
X1	β_1	Positive
X2 = 1	β_2	Positive
X3 = 1	β_3	Negative
X4 = 1	β_4	Negative
X5 = 1	β_5	Negative
Interaction: if X1> 1 & X2=1	β_6	Negative
Interaction: if X3=1 and X4=1 and X5=1	β_7	Positive
X6	0	No true effect

For the binary, categorical, and negative binomial outcome, the beta coefficients selected were transformed to the natural log-scale before inclusion in the outcome generation formulas presented in the next section.

3.4.2 Outcome variable generation

Two formulas provided the underlying process of outcome generation for the five different outcome types, with variations to allow for transformation to each outcome. Table 3.10 outlines the formulas used in the outcome generation process, which included variables X1 to X5, and two- and three-way interactions. Described below is how the outcome variable "Y" was generated using either the categorical or mixed inputs formula.

Table 3.10: Outcome generation formulas for each type of outcome

Outcome	**Inputs**	**Formula**
Continuous	**Categorical**	$Y = \text{intercept}^a + \beta_{1.1}(\text{if } X1=1) + \beta_{1.2}(\text{if } X1=2) + \beta_{1.3}(\text{if } X1=3) + \beta_2X2 + \beta_3X3 + \beta_4X4 + \beta_5X5 + \beta_6(\text{if } X1=2 \ \& \ X2=1) + \beta_7(\text{if } X1=3 \ \& \ X2=1) + \beta_8X3*X4*X5 + e$
	Mixed	$Y = \text{intercept}^a + \beta_1 X1 + \beta_2X2 + \beta_3X3 + \beta_4X4 + \beta_5X5 + \beta_6 X1*X2 (\text{if } X1>1 \ \& \ X2=1) + \beta_7X3*X4*X5 + e$
Binary (rare or common prevalence) and negative binomial	**Categorical**	$z = \text{intercept}^a + \beta_{1.1}(\text{if } X1=1) + \beta_{1.2}(\text{if } X1=2) + \beta_{1.3}(\text{if } X1=3) + \beta_2X2 + \beta_3X3 + \beta_4X4 + \beta_5X5 + \beta_6(\text{if } X1=2 \ \& \ X2=1) + \beta_7(\text{if } X1=3 \ \& \ X2=1) + \beta_8X3*X4*X5$
	Mixed	$z = \text{intercept}^a + \beta_1X1 + \beta_2X2 + \beta_3X3 + \beta_4X4 + \beta_5X5 + \beta_6 X1*X2 (\text{if } X1>1 \ \& \ X2=1) + \beta_7X3*X4*X5$

Multinomial	**Categorical**	mu1 = **1.6** + $\beta_{1.1}$ (if X1=1) + $\beta_{1.2}$ (if X1 = 2) + $\beta_{1.3}$ (if X1 = 3) + β_2X2 + β_3X3 + β_4X4 + β_5X5 + β_6(if X1= 2 & X2=1) + β_7(if X1= 3 & X2=1) + β_8X3*X4*X5 mu2 = **2** + $\beta_{1.1.2}$ (if X_1=1) + $\beta_{1.2.2}$ (if X1 = 2) + $\beta_{1.3.2}$ (if X1 = 3) + $\beta_{2.2}$X2 + $\beta_{3.2}$X3 + $\beta_{4.2}$X4 + $\beta_{5.2}$X5 + $\beta_{6.2}$(if X1= 2 & X2=1) + $\beta_{7.2}$(if X1= 3 & X2=1) + $\beta_{8.2}$ X3*X4*X5
	Mixed	mu1 = **1.6** + β_1 X1 + β_2X2 + β_3X3 + β_4X4 + β_5X5 + β_6 X1*X2 (if X1> 1 & X2=1) + β_7X3*X4*X5 mu2 = **2** + $\beta_{1.2}$ X1 + $\beta_{2.2}$X2 + $\beta_{3.2}$X3 + $\beta_{4.2}$X4 + $\beta_{5.2}$X5 + $\beta_{6.2}$ X1*X2 (if X1> 1 & X2=1) + $\beta_{7.2}$ X3*X4*X5

"*e":* individual error

[a] Where intercept = -3 for the rare binary outcome, -1.5 for the common binary outcome, and 0 for the continuous and negative binomial outcomes

A. Continuous outcome: The data generation formula for the continuous outcome directly generates the outcome Y. The individual error "*e*" was created with a mean of 0 and standard deviation of 1.

B. Binary outcomes (rare and common): The value "z" from Table 3.10 was converted into probabilities, where P(Y=1) = exp (z). The outcome Y was then sampled from this probability. This process creates known RR's for each variable, which are the beta-coefficients (β_x) exponentiated. [75] Therefore, analysis with the modified Poisson method should create comparable beta-coefficient estimates. Because the probability cannot exceed 1, any "z" that resulted in a probability greater than 1 had all coefficients resampled until the probability was less than or equal to 1.

C. Multinomial outcome: The values "mu1" and "mu2" from Table 3.10 were used to create the probabilities of outcomes Y=1, Y=2, and Y=3. A total score from the three possible outcomes was calculated as Denominator=1+exp(mu1)+exp(mu2), and the probability of each of the outcomes was calculated as follows:

P(Y=1) = 1/Denominator

P(Y=2) = exp(mu1)/Denominator

P(Y=3) = exp(mu2)/Denominator

These probabilities were used to sample for outcome Y. This process created known

OR's for each variable, which are the beta-coefficients (β_x) exponentiated. Therefore analysis by multinomial logistic regression should create comparable beta-coefficient estimates.

D. **Negative binomial outcome**

The value "z" from Table 3.10 was converted to the mean of count outcome Y, via mu = exp(z). The outcome Y was then selected via the rnegbin function from the R-package "MASS" [63], using parameters mu and the distribution theta, which was randomly sampled between 0.8 to 1.2 under a uniform distribution.

3.4.3 Simulation feasibility testing

To assess feasibility, run times were recorded for each analysis method, using single iterations of each outcome type, created with categorical input variables. For single-level regression analyses, a single analysis included fitting the regression model, and obtaining standard errors and confidence intervals for each coefficient. For decision tree methods, a single analysis included fitting the decision tree and pruning or tuning when applicable. For MAIHDA, a single analysis included model fitting and estimation of main and random effects, as well as confidence interval construction for main and random effects. A single iteration of cross-classification was calculation of the average value of the outcome for each intersection. These trials were performed using a typical office computer with an Intel Core i5-3470 and 8 GB of RAM, running the 64-bit version of Windows 10.

Chapter 4

4 Results

This chapter will report the performance of the eight analytic methods when applied to 10 different simulated data scenarios, representing descriptive data with intersectional variables, across four sample sizes. First, section 4.1 presents the primary result, prediction accuracy for intersectional groups, for each method across the ten data generation scenarios and four sample sizes. Section 4.2 presents a summary of percent significance of coefficients and confidence interval coverage from best-fitted and over-specified regression analyses. Section 4.3 similarly presents a summary of percent significance of coefficients and confidence interval coverage from select MAIHDA analyses. Section 4.4 presents secondary outcomes for the decision tree models, including the average number of leaves and splitting variables, and probability of splitting on each variable for CART, CHAID, and CTree, and the average number of leaves and variable importance measures for random forest. Results are presented for all ten data generation scenarios and four sample sizes. Finally, section 4.5 presents run times for single iterations of each analysis.

4.1 Primary results

The primary outcome (MSE for continuous or negative binomial outcome models, MAPE for binary or categorical outcome models) is presented for each data generation scenario in Figures 4.1 to 4.10, via boxplots presenting the distribution of the primary outcome across the 1000 iterations.

For all ten scenarios, the accuracy for each method improved with increasing sample size, except for CART. For prediction at the largest sample size (n=200,000), CART performed the poorest with the highest prediction error. At the larger sample sizes, other methods performed relatively the same and approached an MSE or MAPE of zero. One

exception is that the single-level regression methods performed worse than the other methods, but better than CART, for the continuous outcome model with mixed inputs. The following results summarize the worst and best predictors at the smallest sample sizes (n=2000, n=5000), where there were the greatest differences in accuracy between methods. At the smaller sample sizes, the least accurate predictors (highest values of MSE or MAPE) overall were CART, over-specified regression, and cross-classification. For the continuous outcomes overall, the best performers at the smaller sample sizes were MAIHDA, best-fitted regression, and random forest. For the binary outcome models, MAIHDA, best-fitted regression, and CTree (and CHAID when applicable) performed better at smaller sample sizes. For the multinomial outcomes, best-fitted regression performed best, followed by CTree, CHAID, and random forest. For the negative binomial outcome, best-fitted regression and MAIHDA performed well, with the decision tree's close behind and performing similarly to one another. CHAID and CTree performed similarly to one another across sample sizes for all three models that CHAID was applied to.

There are issues of model convergence and missing values to consider when interpreting the primary results. The over-specified regression, especially at smaller sample sizes, resulted in iterations which did not converge, presented in Table 4.1. Therefore, for the models shown in Table 4.1, primary and secondary results from the over-specified regression are not from all 1000 iterations, but rather only from models that converged. Boxplots do not include outliers due to the over-specified regression presenting extreme outliers for some models. Additionally, at the smaller sample sizes, not all 192 intersections were filled with every iteration. Measures of accuracy were calculated with equal weight given to each intersection, regardless of intersection size. Table 4.2 presents the mean number of intersections that remained unfilled for the two input types: mixed and categorical. Therefore, for the smaller sample sizes, calculations of MSE or MAPE did not always include all intersections for MAIHDA and cross-classification, because these methods require a minimum cell size of one to produce predictions. The other methods still produce estimates for intersections with a cell size of zero, and therefore always calculated the MSE or MAPE using predictions from all 192 intersections.

Table 4.1: Number of converged over-specified regression models over 1000 iterations by sample size for select models

	N=2000	N=5000	N=50000	N=200000
Model 3: Common binary outcome, categorical inputs	167	830	1000	1000
Model 4: Common binary outcome, mixed inputs	998	1000	1000	1000
Model 5: Rare binary outcome, categorical inputs	480	855	1000	1000
Model 6: Rare binary outcome, mixed inputs	989	998	1000	1000
Model 9: Negative binomial outcome, categorical inputs	225	878	1000	1000

Table 4.2: Mean and 2.5th percentile and 97.5th percentile of number intersections with cells size zero by the two input data generation models

	N=2000	N=5000	N=50000	N=200000
Categorical inputs	7.788	0.734	0	0
	(3, 13)	(0, 3)	(0, 0)	(0, 0)
Mixed inputs	7.816	0.737	0	0
	(3, 13)	(0, 3)	(0, 0)	(0, 0)

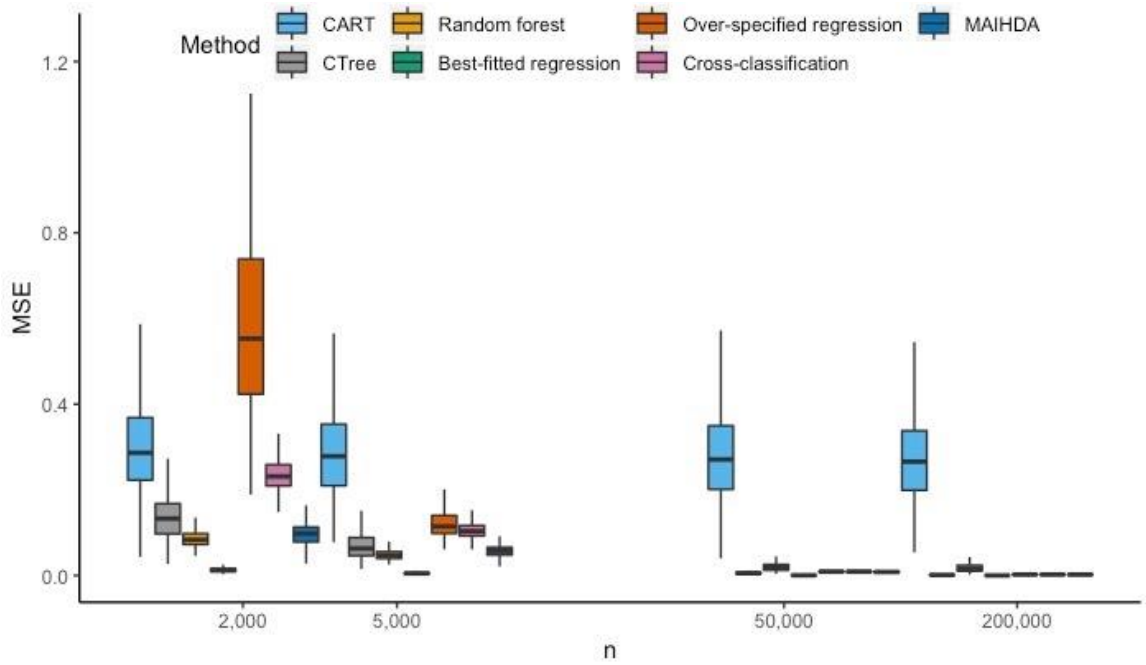

Figure 4.1: Boxplots of intersection prediction MSE for Model 1 (continuous outcome, categorical inputs) across four sample sizes (graph excludes outliers)

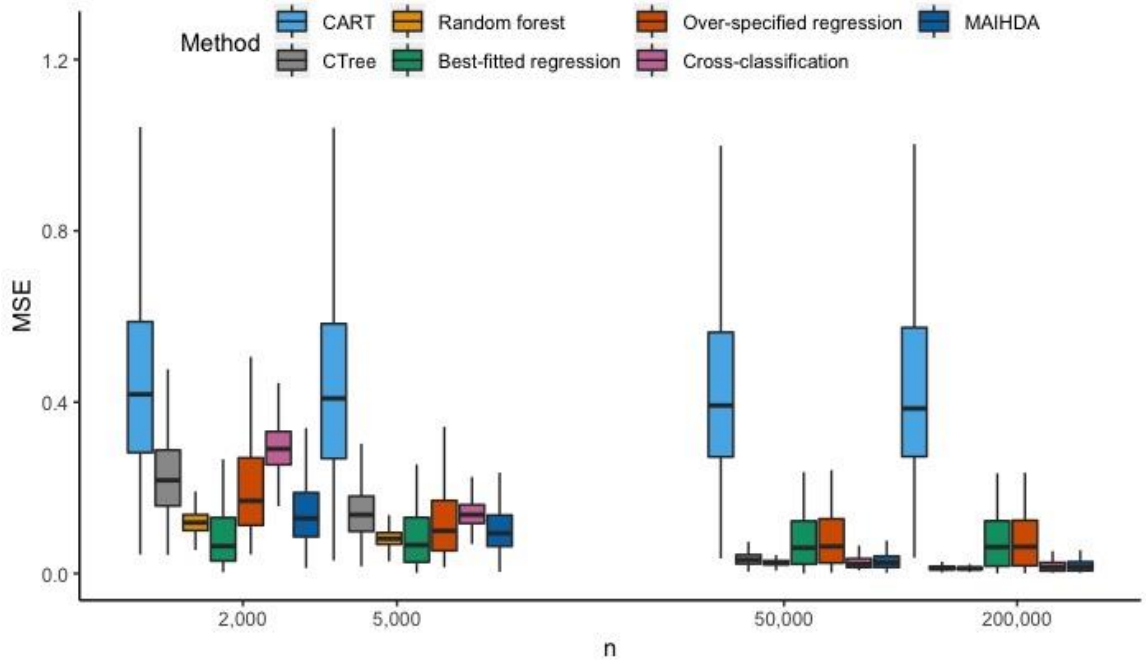

Figure 4.2: Boxplots of intersection prediction MSE for Model 2 (continuous outcome, mixed inputs) across four sample sizes (graph excludes outliers)

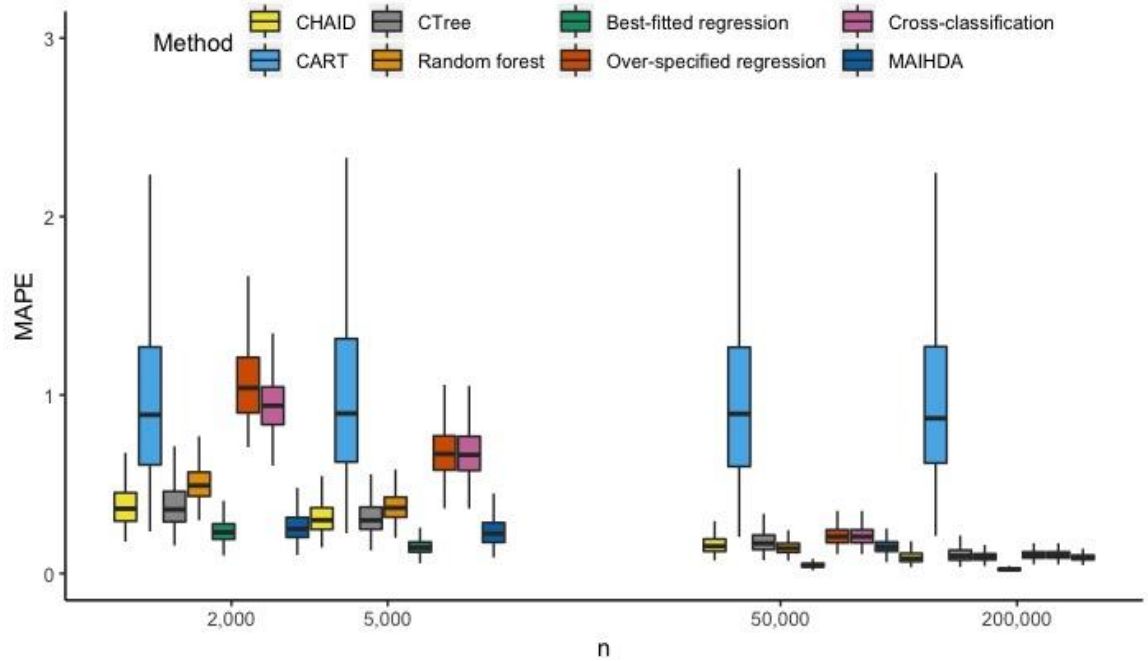

Figure 4.3: Boxplots of intersection prediction MAPE for Model 3 (common binary outcome, categorical inputs) across four sample sizes (graph excludes outliers)

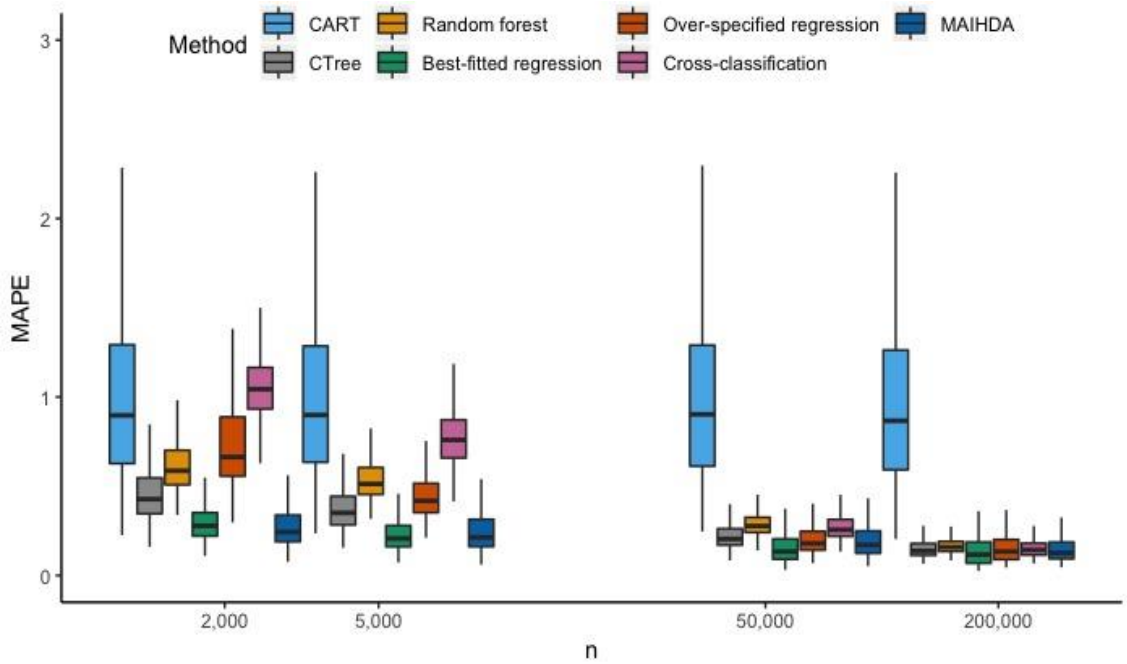

Figure 4.4: Boxplots of intersection prediction MAPE for Model 4 (common binary outcome, mixed inputs) across four sample sizes (graph excludes outliers)

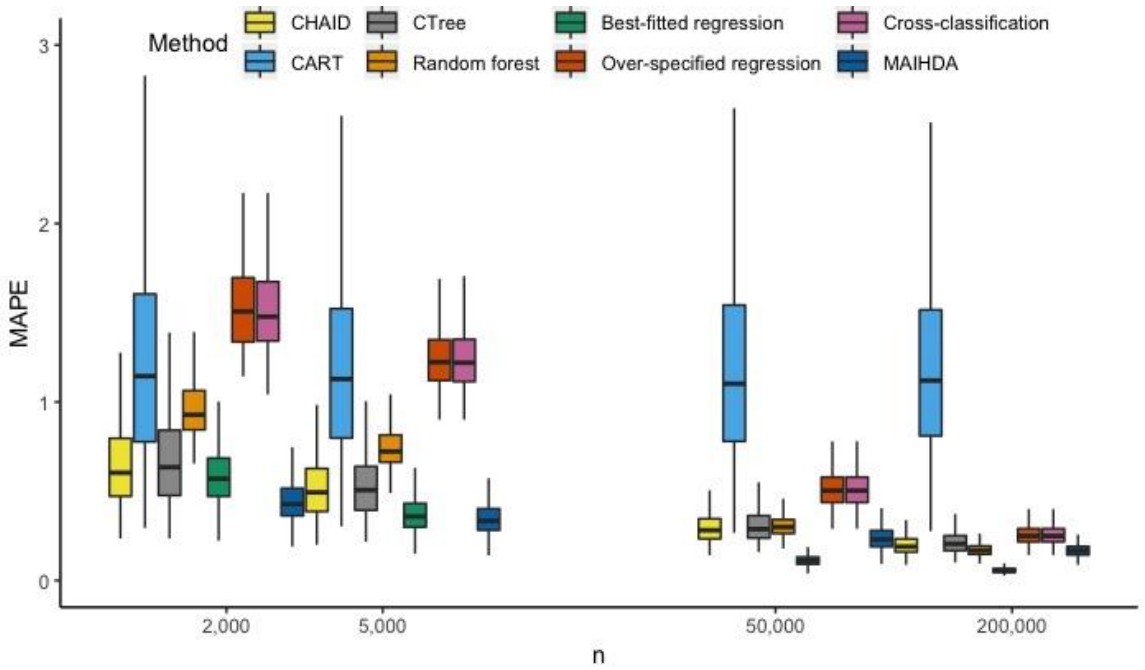

Figure 4.5: Boxplots of intersection prediction MAPE for Model 5 (rare binary outcome, categorical inputs) across four sample sizes (graph excludes outliers)

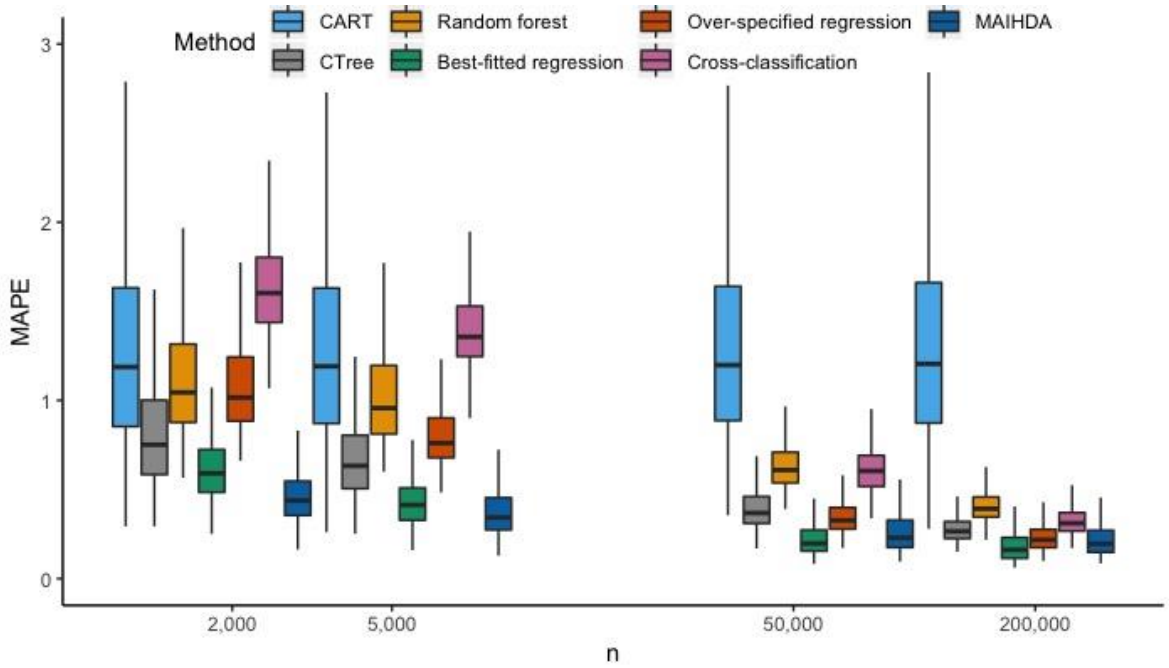

Figure 4.6: Boxplots of intersection prediction MAPE for Model 6 (rare binary outcome, mixed inputs) across four sample sizes (graph excludes outliers)

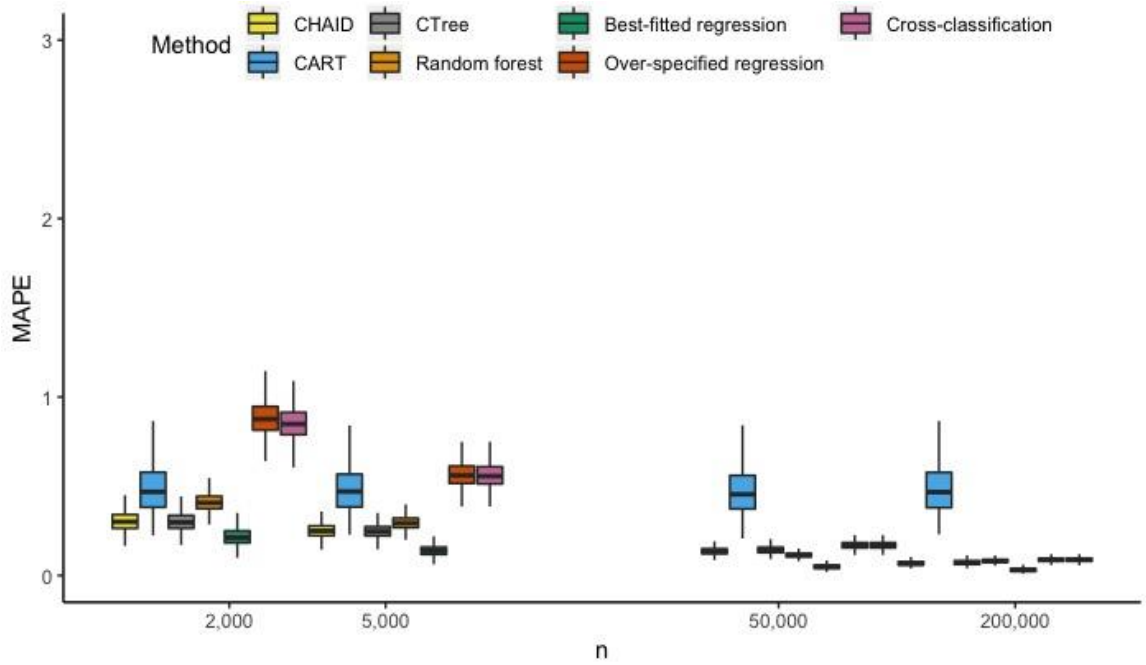

Figure 4.7: Boxplots of intersection prediction MAPE for Model 7 (multinomial outcome, categorical inputs) when y=1, across four sample sizes (graph excludes outliers)

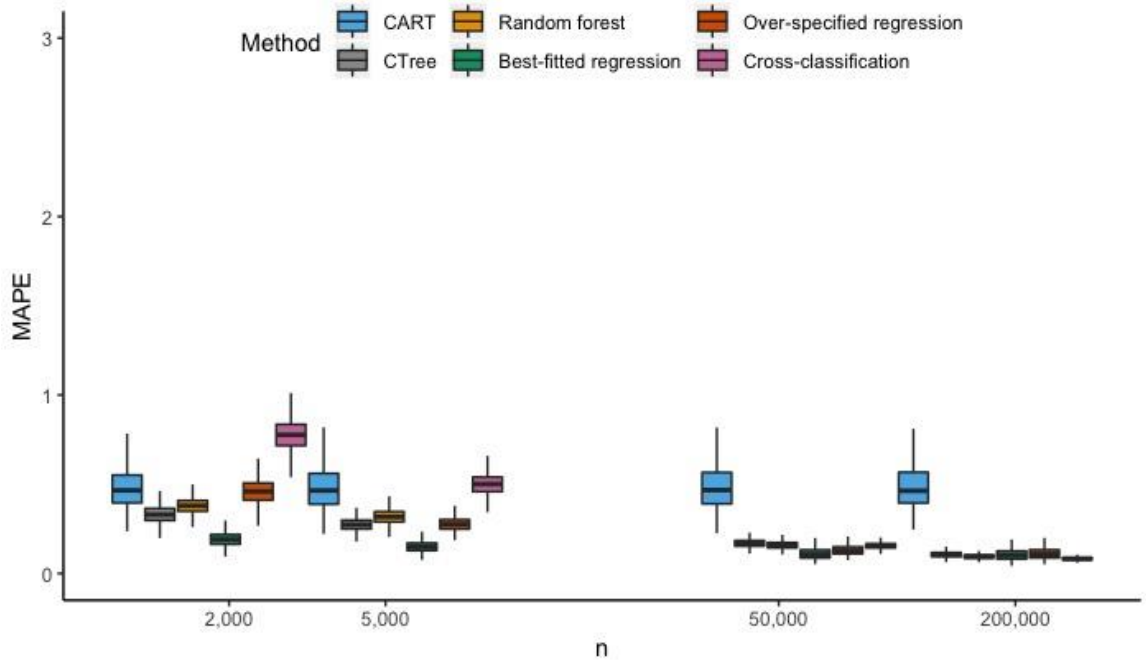

Figure 4.8: Boxplots of intersection prediction MAPE for Model 8 (multinomial outcome, mixed inputs) when y=1, across four sample sizes (graphs excludes outliers)

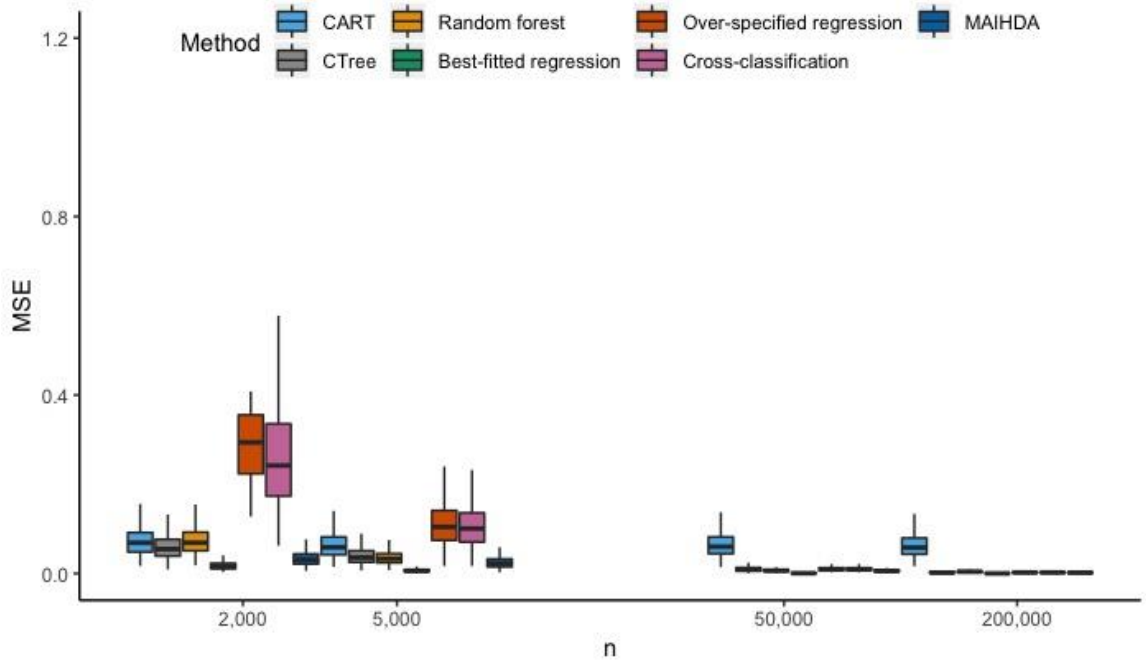

Figure 4.9: Boxplots of intersection prediction MSE for Model 9 (negative binomial outcome, categorical inputs), across four sample sizes (graph excludes outliers)

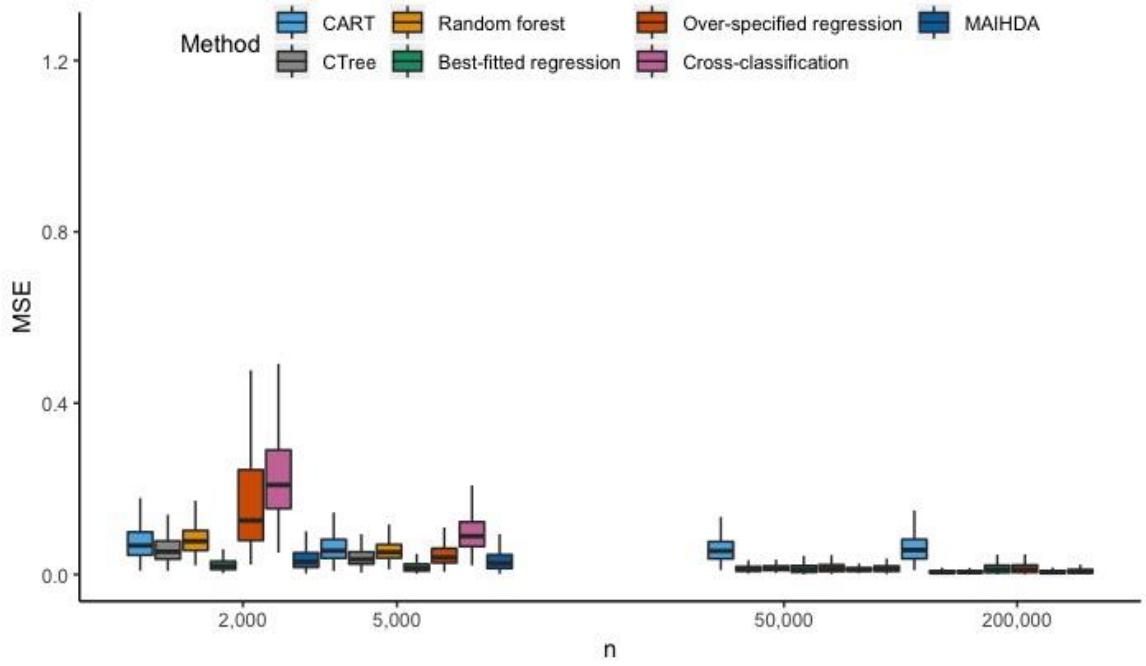

Figure 4.10: Boxplots of intersection prediction MSE for Model 10 (negative binomial outcome, mixed inputs), across four sample sizes (graph excludes outliers)

4.2 Regression secondary results

Results from the over-specified and best-fitted regression analyses are presented by the percentage of completed iterations that important coefficients (the intercept, X1 to X6, X1*X2 and X3*X4*X5) are detected as significant, and the confidence interval coverage for these coefficients. Generally, the over-specified regression required larger sample sizes for coefficient significance and confidence interval coverage to resemble that of the best-fitted regressions. Full results for all ten models are presented in Appendix B. Select results are discussed below.

For models with categorical inputs, confidence interval coverage was approximately 95% for all important coefficients. Results differed for the mixed input models. Table 4.3 presents the confidence interval coverage of over- and best-fitted regression analysis for Model 2 (continuous outcome, mixed inputs). The confidence interval coverage was approximately 95% for most main effects from the categorical input models, across all sample sizes. However, because the simulated outcome was formed with a non-linear interaction between X1 and X2 (where the interaction between continuous variable X1 and binary variable X2 only begins when X1 is greater than 1), the confidence interval coverage for X2 and X1*X2 was poor and decreased with increasing sample size. Similar results are seen for the other mixed input models, where confidence interval coverage was poorest for X2 and X1*X2.

Table 4.3: Model 2 **(continuous outcome, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

		Intercept	x1	x2	x3	x4	x5	x1:x2	x3:x4:x5
Over-specified	N = 2000	96	95.7	53.8	96	96	95	17.3	95
	N = 5000	95.7	96.3	28.7	95.7	96.4	95.3	9.1	95
	N = 50000	95.7	97.1	6.1	95.5	94.5	96.2	1.1	95.8
	N = 200000	96	95.9	1.4	95.7	94.9	95.7	0.1	95.8
Best-fitted	N = 2000	96.7	95.8	21.1	96	95.8	94.9	6.2	96.1
	N = 5000	95.1	95.9	8.9	94.3	95.5	95.3	2.8	94.4
	N = 50000	95.2	96	1.2	94.6	95.1	96.2	0	95.2
	N = 200000	95.3	95.2	0	96.4	94.9	95.3	0	95.1

Table 4.4 presents the coefficient significance of best- and over-specified regression for Model 2 (continuous outcome, mixed inputs). As expected, with increasing sample size coefficient significance approached 100% for variables relevant to the outcome (X1 to X5, X1*X2, and X3*X4*X5), and not for the intercept and X6, which had true values of zero and were expected to only be significant in approximately 5% of models. There were however surprising results concerning coefficient significance of the three-way interaction term for models aside from Model 2. Table 4.5 and 4.6 show the coefficient percent significance for the three-way interaction term X3*X4*X5, for the over-specified regression and best-fitted regression respectively. When using an over-specified regression model on binary outcomes with either categorical or mixed inputs, and multinomial outcomes with categorical inputs, significance of the three-way interaction did not consistently increase, but rather fluctuated or decreased with increasing sample size. The same result was also observed for models with a rare binary outcome fitted with a best-fitted regression. These results demonstrate that increasing sample size does not always result in better identification of significant interactions, even in a best-fitted regression model.

Table 4.4: Model 2 **(continuous outcome, mixed inputs)** regression coefficient significance (% of iterations)

		Intercept	x1	x2	x3	x4	x5	x6	x1:x2	x3:x4:x5
Expected		5	100	100	100	100	100	5	100	100
Over-specified	N = 2000	4	97.5	80	93.9	96.2	92.2	4.7	69.5	79.9
	N = 5000	4.3	99	87.4	96.4	98.7	96.6	3.5	84.4	89.3
	N = 50000	4.3	100	96.8	100	100	99.9	3.8	97.2	98.2
	N = 200000	4	100	97.7	100	100	100	5.6	99.7	99.7
Best-fitted	N = 2000	3.3	99.5	88.5	94.2	96.3	93.2	4.4	90.3	83.7
	N = 5000	4.9	100	94.6	96.7	99	97.2	4.5	95.2	91.8
	N = 50000	4.8	100	98.8	100	100	99.8	4.9	99.5	98.4
	N = 200000	4.7	100	99.2	100	100	100	4.9	100	99.8

Table 4.5: Over-specified regression % significance for 3-way interaction (x3*x4*x5)

	N=2000	N=5000	N=50000	N=200000
Model 1: Continuous outcome, categorical inputs	22.7	48.1	86.9	94.4

Model 2: Continuous outcome, mixed inputs	79.9	89.3	98.2	99.7
Model 3: Common binary outcome, categorical inputs	65.8	56.7	9.6	24.5
Model 4: Common binary outcome, mixed inputs	24.4	13.9	50.0	84.8
Model 5: Rare binary outcome, categorical inputs	40.2	62.4	40.4	10.3
Model 6: Rare binary outcome, mixed inputs	68.0	44.4	20.0	47.4
Model 7: Multinomial outcome, categorical inputs (Y=2)	76.3	21.2	13.6	36.2
Model 8: Multinomial outcome, mixed inputs (Y=2)	10.3	15.7	74.6	98.5
Model 9: Negative binomial outcome, categorical inputs	1.8	4.2	19.2	49.1
Model 10: Negative binomial outcome, mixed inputs	15.9	23.6	79.6	95.5

Table 4.6: Best-fitted regression % significance for 3-way interaction (x3*x4*x5)

	N=2000	N=5000	N=50000	N=200000
Model 1: Continuous outcome, categorical inputs	82.0	90.2	98.5	99.8
Model 2: Continuous outcome, mixed inputs	83.7	91.8	98.4	99.8
Model 3: Common binary outcome, categorical inputs	9.3	14.7	68.4	92.8
Model 4: Common binary outcome, mixed inputs	11.3	14.1	57.7	88.9
Model 5: Rare binary outcome, categorical inputs	48.7	21.0	26.6	64.8
Model 6: Rare binary outcome, mixed inputs	50.7	25.3	23.5	61.2
Model 7: Multinomial outcome, categorical inputs (Y=2)	9.6	17.3	79.2	98.9
Model 8: Multinomial outcome, mixed inputs (Y=2)	10.4	17.2	80.2	99.2
Model 9: Negative binomial outcome, categorical inputs	15.3	29.8	86.8	98.6
Model 10: Negative binomial outcome, mixed inputs	16.7	27.4	85.5	97.3

4.3 MAIHDA

Results from MAIHDA are presented by the coefficient percent significance and confidence interval coverage of the main effects (X1 to X6). MAIHDA estimates for main effects differ from the typical definition of main effects. There were two possible definitions of main effects to use when determining confidence interval coverage: 1) main effects capture the additive effects only, and 2) main effects as the average effects of the variable across equally weighted clusters. The calculations for definition 2 of MAIHDA main effects are only possible for models with only categorical inputs. Therefore, coefficient significance and confidence interval coverage results are presented below for the four models that both MAIHDA estimands can be calculated for (Model 1, 3, 5, and 9). For models with mixed inputs (Model 2, 4, 6, and 10), confidence interval coverage could only be assessed using definition 1. Significance of coefficients and confidence interval coverage using definition 1 is presented for these mixed input models in Appendix C.

Tables 4.7 to 4.10 present the coefficient significance of main effects for the continuous, common binary, rare binary, and negative binomial outcome models with categorical inputs. Compared to the single-level regression methods, the coefficient significances of main effects from MAIHDA were farther away from the expected values. Coefficients for variables X1.2, X1.3, and X2 consistently had lower percent significance than other variables. Across all models, MAIHDA did not consistently identify variables included in the data generating process as significant.

Table 4.7: Model 1 **(Continuous outcome, categorical inputs)** MAIHDA coefficient significance (% of iterations)

	Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5	x6.1	x6.2
Expected	0	100	100	100	100	100	100	100	0	0
N = 2000	12.9	94.3	83.8	78.8	84.3	88.3	90.1	84.8	0.8	0.9
N = 5000	19.0	94.4	83.4	80.2	86.5	89.4	89.7	87.6	0.3	0.3
N = 50000	32.0	96.6	83.4	83.5	88.1	88.4	89.7	89.4	0.0	0.2
N = 200000	33.9	95.9	84.2	83.2	85.7	90.5	91.4	89.4	0.0	0.0

Table 4.8: Model 3 **(Common binary outcome, categorical inputs)** MAIHDA coefficient significance (% of iterations)

	Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5	x6.1	x6.2
Expected	100	100	100	100	100	100	100	100	0	0
N = 2000	100.0	62.2	36.6	33.6	22.9	86.5	88.4	77.1	3.0	1.8
N = 5000	100.0	82.5	56.8	54.3	38.6	96.1	96.4	87.8	2.0	2.4
N = 50000	100.0	98.2	66.6	66.2	70.0	97.8	98.3	95.8	0.4	0.2
N = 200000	100.0	99.0	70.1	65.4	73.9	97.6	97.3	96.1	0.2	0.0

Table 4.9: Model 5 **(Rare binary outcome, categorical inputs)** MAIHDA coefficient significance (% of iterations)

	Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5	x6.1	x6.2
Expected	100	100	100	100	100	100	100	100	0	0
N = 2000	100.0	20.6	9.2	9.7	7.9	69.1	72.1	49.3	4.4	3.5
N = 5000	100.0	46.1	20.9	23.4	12.8	88.3	90.4	78.6	3.1	4.3
N = 50000	100.0	98.4	64.3	67.9	43.8	100	99.9	99.4	3.1	2.8
N = 200000	100.0	99.8	63.3	62	58.5	100	100	99.9	0.7	0.3

Table 4.10: Model 9 **(Negative binomial outcome, categorical inputs)** MAIHDA coefficient significance (% of iterations)

	Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5	x6.1	x6.2
Expected	0	100	100	100	100	100	100	100	0	0
N = 2000	5.3	81.3	56.2	52.9	37.6	91.6	93.2	84.8	3.8	3.6
N = 5000	7.2	90.0	64.4	64.7	53.2	96.0	94.7	88.7	3.2	2.4
N = 50000	17.5	98.9	70.2	68.5	72.8	95.6	95.7	94.3	0.1	0.2
N = 200000	27.5	99.4	71.3	69.4	76.9	95.9	95.5	95.0	0.1	0.1

Tables 4.11 to 4.14 present the confidence interval coverage according to definitions 1 and 2 for the continuous, common binary, rare binary, and negative binomial outcome with categorical inputs. The confidence interval coverage did not approach 95% for definition 1, indicating that the traditional definition of additive effects (definition 1) does not apply to MAIHDA. This is true for models presented in the Appendix as well. Confidence interval coverage was sufficient only for variable X1.1 by definition 1, because this variable was not involved in any interactions, therefore the average effect of this variable across equally weighted clusters was the same as the additive effect. When

observing Table 4.11 for the continuous and Table 4.14 for the negative binomial outcome, the confidence interval coverage for definition 2 was almost or approached 100%, demonstrating that definition 2 is more aligned with the true MAIHDA estimand. As previously mentioned in the Methods section, the estimand for the intercept according to definition 2 is unknown, and therefore confidence interval coverage was not calculated for the intercept. Notably, the confidence interval coverage in Table 4.11 is extremely conservative, by exceeding 95%. This is in alignment with previous results from Lizotte et. al. (2019), that MAIHDA results in very large SE's.

Regarding Tables 4.12 and 4.13 for common binary and rare binary outcomes, while definition 1 did not hold, the confidence interval coverage for definition 2 was also not near 95%. Overall, results show that the typical interpretation of additive effects under intersectionality theory (definition 1) does not apply to MAIHDA, regardless of outcome type. While MAIHDA main effects can be interpreted by definition 2 for the continuous and negative binomial outcome, it is unclear what the defined main effects for binary outcomes would be.

Table 4.11: Model 1 **(Continuous outcome, categorical inputs)** MAIHDA confidence interval coverage by definition 1 (typical additive effects) and definition 2 (MAIHDA additive effects) (% of iterations)

		Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5
Definition 1	N = 2000	87.1	99.3	28.8	27.9	3.8	33.3	38.8	18.3
	N = 5000	81.0	99.9	18.4	19.4	1.6	21.9	23.6	15.8
	N = 50000	68.0	100.0	9.5	10.2	0.7	13.4	14.4	12.7
	N = 200000	66.1	100.0	11.2	8.4	0.4	14.4	15.2	15.0
Definition 2	N = 2000	-	99.3	62.1	63.2	98.2	95.6	90.1	94.2
	N = 5000	-	99.9	82.7	81.8	99.4	98.8	97.6	98.7
	N = 50000	-	100.0	99.4	99.7	100.0	100.0	100.0	99.9
	N = 200000	-	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 4.12: Model 3 (Common binary outcome, categorical inputs) MAIHDA confidence interval coverage by definition 1 (typical additive effects) and definition 2 (MAIHDA additive effects) (% of iterations)

		Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5
Definition 1	N = 2000	94.6	98.1	89.9	88.5	46.8	96.9	96.8	91.4
	N = 5000	94.9	97.4	75.0	74.7	17.4	94.8	96.5	88.4
	N = 50000	86.6	99.6	16.8	16.0	0.2	73.4	80.4	49.4
	N = 200000	78.1	100.0	6.6	5.9	0.0	43.5	46.2	27.7
Definition 2	N = 2000	-	98.1	74.1	74.1	94.6	95.1	93.4	96.0
	N = 5000	-	97.4	54.8	57.3	91.6	90.0	87.1	95.5
	N = 50000	-	99.6	41.6	42.2	91.1	81.1	76.0	96.0
	N = 200000	-	100.0	76.2	72.7	99.5	92.7	88.3	98.7

Table 4.13: Model 5 (Rare binary outcome, categorical inputs) MAIHDA confidence interval coverage by definition 1 (typical additive effects) and definition 2 (MAIHDA additive effects) (% of iterations)

		Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5
Definition 1	N = 2000	96.0	96.1	94.1	93.5	84.0	95.4	95.7	93.4
	N = 5000	95.1	95.7	90.7	90.2	56.6	94.8	95.5	93.2
	N = 50000	88.8	97.4	43.2	45.0	1.2	90.7	94.3	79.1
	N = 200000	85.2	99.6	9.6	9.6	0.0	79.1	82.7	55.4
Definition 2	N = 2000	-	96.1	91.6	91.5	94.9	94.9	94.2	95.4
	N = 5000	-	95.7	80.5	80.8	93.9	93.7	93.9	94.7
	N = 50000	-	97.4	38.2	40.6	85.2	81.1	75.3	92.9
	N = 200000	-	99.6	43.0	41.1	93.8	80.8	79.3	96.0

Table 4.14: Model 9 (Negative binomial outcome, categorical inputs) MAIHDA confidence interval coverage by definition 1 (typical additive effects) and definition 2 (MAIHDA additive effects) (% of iterations)

		Intercept	x1.1	x1.2	x1.3	x2	x3	x4	x5
Definition 1	N = 2000	94.7	96.6	78.1	77.5	20.3	90.3	93.3	80.1
	N = 5000	92.8	96.6	52.6	54.3	7.0	87.4	91.5	66.9
	N = 50000	82.5	99.8	10.8	13.2	0.1	50.7	51.5	31.8
	N = 200000	72.5	100.0	6.2	6.6	0.0	27.8	30.0	22.8
Definition 2	N = 2000	-	96.6	59.2	58.5	89.6	90.9	87.8	94.3
	N = 5000	-	96.6	42.7	44.6	90.0	86.3	82.7	94.5
	N = 50000	-	99.8	64.6	68.9	99.2	92.7	92.1	99.0
	N = 200000	-	100.0	95.7	96.5	99.9	98.6	97.5	99.8

4.4 Decision tree outcomes

Decision tree methods were evaluated using the following criteria. For CART, CTree, and CHAID, the ideal method will split on variables X1 to X5 for 100% of iterations, but not on X6, which has no true effect on the outcome. For random forest, the variable importance measure should be higher for variables X1 to X5 than for X6. For all methods, the number of leaves can be seen as the number of unique "intersectional groups" identified by the decision tree. For the categorical input models, there are 192 possible intersections, 64 of which are actually distinct from one another. For models with mixed inputs there is no defined number of leaves that we would expect to see. Results for CART, CTree, and CHAID are presented for all ten models in Tables 4.15 to 4.24. Results for random forest for all ten models are presented in Tables 4.25 to 4.34.

4.4.1 CART

For the continuous outcome models, CART split on variables X1 to X5, and did not split on the null variable X6. Splitting was not near 100% of iterations for variables X2 to X5, and did not improve with increasing sample size. Resultantly, the number of leaves was much lower than 64. For the binary outcomes, both of common and rare prevalence, there was almost no splitting at all. Therefore, for many of the iterations, predictions were only based on the population prevalence, and were equal across all intersections. For the multinomial outcome, there was some splitting on variables X3, X4, and X5 for 20 to 35% of iterations and even less splitting on X1 and X2. No variables approached a 100% split rate. X6 was very rarely used as a splitting variable. For the negative binomial outcome, X1 to X5 were used as splitting variables, and there was no splitting on X6, but the splitting rate was not near 100% for variables X1 to X5. Overall, amongst those outcomes where there was a sizable amount of splitting (continuous, multinomial, and negative binomial) a noticeable pattern was that X2 was split on less often, which may be a result of the interaction between X2 (a binary variable) and X1 (a continuous or categorical variable). Generally, CART would correctly avoid splitting on the null

variable X6, but was not guaranteed to split on all relevant variables for continuous, categorical, or negative binomial models, and mostly did not split at all for the binary outcomes.

4.4.2 CTree

For the continuous outcome, CTree split on variables X1 to X5 at an almost 100% rate, even at the lowest sample size (n=2,000). Splitting on variable X6 approached 100% with increasing sample size, and occurred even at the lowest sample size, for both when X6 was categorical (categorical input models) and when it was continuous (mixed input models). For the binary, multinomial, and negative binomial outcomes, splitting on X1 to X5 increased with increasing sample size, and reached a 100% splitting rate by the larger sample sizes. The X6 splitting rate was lower than for variables X1 to X5, but also increased with increasing sample size. The number of leaves for the categorical input models appears to be too low given the number of splitting variables used. For example, the analysis of continuous outcomes with categorical inputs at n=200,000 resulted in approximately 62 leaves, even though variables X1 to X6 were used in splitting, which would result in a total possible 192 intersections. Therefore, not all possible splits were performed using the given split variables. When comparing between categorical input models at the n=200,000, the resulting number of leaves was greatest for the continuous outcome, and lowest for the rare binary outcome. Similar to CART, the variable X2 was not split on as often as variables X1 and X3 to X5.

4.4.3 CHAID

CHAID results were very similar to results from CTree. A notable difference was that the splitting rates for all variables X1 to X6 was slightly higher for CHAID across all three models (Models 3, 5, and 7), when starting at n=2,000.

4.4.4 Random forest

Random forest uses the variable importance measure to illustrate which variables are more important to the outcome. This value is compared between variables, rather than

statistically analysed. For the continuous outcome model with categorical inputs, the variable importance was lowest for X6 across all sample sizes. For the models with categorical inputs that had common binary, rare binary, and multinomial outcomes, X6 was only the least important at the larger sample sizes (50,000, 200,000, and 50,000 respectively). For the binary, multinomial, and negative binomial models where X6 is continuous (mixed inputs), X6 was the second most important variable, after X1 (the other continuous variable), even at the largest sample size. For categorical input models, the average number of leaves produced by the random forest model was always between 90 and 100 by n=200,000. Compared to CTree, random forest produced more leaves for both mixed and categorical input models. From these results it appears that one could only reliably infer that X6 is the variable with no true effect for continuous outcome models with categorical inputs. For categorical input models with binary, multinomial, or negative binomial outcomes, large sample sizes are required to correctly identify the least important variable. For mixed input models, the variable importance measure prescribes continuous variables with greater importance even if they have no true effect, and does not reliably identify variables relevant to the outcome.

Table 4.15: Model 1 **(continuous outcome, categorical inputs)** CART and CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	8.843 (5, 13)	3.914 (3, 5)	96.3	55.4	79.5	81.9	78.3	0.0
	5000	8.762 (5,13)	3.865 (2,5)	96.2	53.5	79.2	79.7	77.9	0.0
	50000	8.454 (4,12)	3.743 (2,5)	96.7	51.1	76.3	77.0	73.2	0.0
	200000	8.597 (5,13)	3.768 (2,5)	96.7	50.3	77.5	78.1	74.2	0.0
CTree	2000	23.915 (12, 36)	5.385 (5,6)	100.0	98.6	99.3	99.4	99.4	41.8
	5000	33.829 (19, 49)	5.64 (5,6)	100.0	100.0	100.0	99.9	100.0	64.1
	50000	55.63 (39,68)	5.91 (5,6)	100.0	100.0	100.0	100.0	100.0	91.0
	200000	61.945 (50,70)	5.945 (5,6)	100.0	100.0	100.0	100.0	100.0	94.5

[a] Means presented with 2.5^{th} and 97.5^{th} percentiles

Table 4.16: Model 2 (**continuous outcome, mixed inputs**) CART and CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	9.846 (5, 14)	3.399 (2, 5)	93.0	56.0	66.5	67.5	56.9	0.0
	5000	9.701 (5,14)	3.412 (2,5)	92.5	54.0	67.7	69.8	57.2	0.0
	50000	9.441 (5,14)	3.234 (2,5)	91.3	49.7	63.5	64.3	54.6	0.0
	200000	9.56 (5,14)	3.255 (2,5)	92.3	52.8	64.4	62.7	53.3	0.0
CTree	2000	33.819 (12, 59)	5.251 (4, 6)	99.6	98.2	96.4	98.3	98.0	34.6
	5000	56.349 (19, 95)	5.538 (5, 6)	99.9	99.7	99.7	99.6	99.9	55.0
	50000	162.499 (47, 276)	5.925 (5, 6)	100.0	100.0	100.0	100.0	100.0	92.5
	200000	278.53 (88, 449)	5.988 (6, 6)	100.0	100.0	100.0	100.0	100.0	98.8

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.17: Model 3 (**Common binary outcome, categorical inputs)** CART, CTree, and CHAID outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	1.027 (1, 1)	0.027 (0, 0)	0.7	0.7	0.5	0.6	0.2	0.0
	5000	1.014 (1,1)	0.014 (0,0)	0.4	0.4	0.2	0.2	0.2	0.0
	50000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
	200000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
CTree	2000	4.838 (2, 8)	3.231 (1, 5)	50.6	23.9	83.3	85.6	72.8	6.9
	5000	7.476 (4,12)	4.253 (2,6)	84.3	52.4	94.7	93.7	88.4	11.8
	50000	22.564 (13,34)	5.472 (5,6)	100.0	98.4	100.0	100.0	99.9	48.9
	200000	37.243 (23,51)	5.767 (5,6)	100.0	100.0	100.0	100.0	100.0	76.7
CHAID	2000	6.592 (3, 11)	4.076 (2, 6)	64.8	50.0	91.5	93.0	85.2	23.1
	5000	10.071 (5,16)	4.908 (3,6)	89.9	76.3	98.1	97.4	94.6	34.5
	50000	28.6 (17,42)	5.674 (5,6)	100.0	99.7	100.0	100.0	100.0	67.7
	200000	43.204 (28,58)	5.844 (5,6)	100.0	100.0	100.0	100.0	100.0	84.4

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.18: Model 4 **(Common binary outcome, mixed inputs)** CART and CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	1.087 (1, 1)	0.059 (0, 0)	2.4	0.5	0.6	1.3	0.4	0.7
	5000	1.011 (1, 1)	0.009 (0, 0)	0.4	0.1	0.1	0.2	0.1	0.0
	50000	1 (1, 1)	0 (0, 0)	0.0	0.0	0.0	0.0	0.0	0.0
	200000	1 (1, 1)	0 (0, 0)	0.0	0.0	0.0	0.0	0.0	0.0
CTree	2000	5.681 (3, 10)	3.445 (1, 5)	82.5	38.2	76.4	79.8	61.8	5.8
	5000	8.824 (4,14)	4.351 (3,6)	93.6	64.2	90.2	90.7	83.7	12.7
	50000	26.52 (13,42)	5.404 (5,6)	100.0	99.0	99.9	100.0	99.8	41.7
	200000	48.417 (25,74)	5.644 (5,6)	100.0	100.0	100.0	100.0	100.0	64.4

[a] Means presented with 2.5^{th} and 97.5^{th} percentiles

Table 4.19: Model 5 **(rare binary outcome, categorical inputs)** CART, CTree, and CHAID outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	1 (1, 1)	0 (0, 0)	0.0	0.0	0.0	0.0	0.0	0.0
	5000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
	50000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
	200000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
CTree	2000	2.621 (1, 4)	1.612 (0, 3)	7.9	5.0	56.3	62.6	27.4	2.0
	5000	3.833 (2, 6)	2.624 (1,5)	21.6	11.0	77.6	82.7	65.1	4.4
	50000	10.871 (6, 17)	4.896 (4,6)	98.6	71.7	100.0	100.0	99.5	19.8
	200000	20.962 (12, 32)	5.438 (5,6)	100.0	98.5	100.0	100.0	100.0	45.3
CHAID	2000	3.687 (2, 7)	2.524 (1, 5)	18.5	20.5	72.9	76.2	52.6	11.7
	5000	5.23 (3, 9)	3.483 (2,6)	36.8	32.1	89.5	92.1	80.0	17.8
	50000	14.557 (8, 22)	5.366 (4,6)	99.2	90.4	100.0	100.0	99.8	47.2
	200000	26.577 (15,39)	5.667 (5,6)	100.0	99.9	100.0	100.0	100.0	66.8

[a] Means presented with 2.5^{th} and 97.5^{th} percentiles

Table 4.20: Model 6 **(rare binary outcome, mixed inputs)** CART and CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	1.005 (1, 1)	0.004 (0, 0)	0.1	0.0	0.1	0.1	0.0	0.1
	5000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
	50000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
	200000	1 (1,1)	0 (0,0)	0.0	0.0	0.0	0.0	0.0	0.0
CTree	2000	2.91 (1, 5)	1.834 (0, 4)	49.6	10.3	48.8	52.2	19.3	3.2
	5000	4.588 (2,8)	3.008 (1,5)	78.8	21.9	73.3	75.0	47.3	4.5
	50000	16.029 (9,24)	5.137 (4,6)	100.0	93.7	99.6	99.6	98.0	22.8
	200000	33.658 (19,49)	5.502 (5,6)	100.0	99.9	100.0	100.0	100.0	50.3

[a] Means presented with 2.5^{th} and 97.5^{th} percentiles

Table 4.21: Model 7 **(multinomial outcome, categorical inputs)** CART, CTree, and CHAID outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	1.896 (1, 5)	0.863 (0, 3)	5.8	3.0	27.3	27.8	21.8	0.6
	5000	1.806 (1,5)	0.772 (0,3)	3.3	1.4	25.3	25.7	21.4	0.1
	50000	1.849 (1,4)	0.82 (0,3)	1.8	1.8	28.0	26.9	23.5	0.0
	200000	1.769 (1,4)	0.752 (0,3)	0.6	0.8	26.1	24.6	23.1	0.0
CTree	2000	5.54 (3, 9)	3.277 (2, 5)	36.2	22.2	89.1	88.6	84.1	7.5
	5000	8.722 (5,14)	4.286 (3,6)	77.8	47.4	97.6	96.4	95.2	14.2
	50000	32.039 (19,45)	5.654 (5,6)	100.0	100.0	100.0	100.0	100.0	65.4
	200000	53.035 (40,65)	5.891 (5,6)	100.0	100.0	100.0	100.0	100.0	89.1
CHAID	2000	7.589 (4, 13)	4.193 (2,6)	57.2	49.2	94.8	96.2	94.3	27.6
	5000	11.454 (6, 19)	5.012 (3,6)	87.9	74.9	99.3	98.9	98.7	41.5
	50000	38.557 (25, 52)	5.821 (5,6)	100.0	100.0	100.0	100.0	100.0	82.1
	200000	57.345 (46,66)	5.911 (5,6)	100.0	100.0	100.0	100.0	100.0	91.1

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.22: Model 8 **(multinomial outcome, mixed inputs)** CART, CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	2.452 (1, 6)	1.311 (0, 4)	31.2	2.3	34.2	33.9	25.8	3.7
	5000	2.247 (1,5)	1.172 (0, 4)	23.6	0.4	34.4	33.1	25.5	0.2
	50000	2.091 (1,5)	1.05 (0, 4)	16.3	0.3	32.6	33.8	22.0	0.0
	200000	2.114 (1,5)	1.08 (0, 4)	16.5	0.2	34.8	31.8	24.7	0.0
CTree	2000	7.49 (4, 12)	3.919 (2, 6)	95.1	27.9	89.9	88.7	82.1	8.2
	5000	12.344 (7,19)	4.683 (3,6)	99.5	60.2	97.9	97.9	93.5	19.3
	50000	44.828 (28,60)	5.617 (5,6)	100.0	100.0	100.0	100.0	100.0	61.7
	200000	86.055 (60, 107)	5.853 (5,6)	100.0	100.0	100.0	100.0	100.0	85.3

[a] Means presented with 2.5^{th} and 97.5^{th} percentiles

Table 4.23: Model 9 (**negative binomial outcome, categorical inputs**)
CART and CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	3.438 (2, 6)	2.419 (1, 5)	33.2	9.1	69.8	72.3	55.5	2.0
	5000	3.450 (2,5)	2.447 (1, 4)	30.0	6.3	73.2	73.7	61.4	0.1
	50000	3.275 (2,5)	2.275 (1, 4)	23.2	2.9	70.6	72.2	58.6	0.0
	200000	3.279 (2,5)	2.278 (1,4)	23.1	1.2	68.6	75.4	59.5	0.0
CTree	2000	6.303 (3, 10)	3.707 (2, 5)	69.4	37.0	86.9	89.1	80.1	8.2
	5000	10.077 (5, 16)	4.649 (3,6)	94.0	70.3	95.8	95.0	93.9	15.9
	50000	30.498 (18, 44)	5.635 (5,6)	100.0	99.8	100.0	100.0	100.0	63.7
	200000	46.187 (29, 61)	5.838 (5, 6)	100.0	100.0	100.0	100.0	100.0	83.8

[a] Means presented with 2.5^{th} and 97.5^{th} percentiles

Table 4.24: Model 10 **(negative binomial outcome, mixed inputs)** CART, CTree outcomes

	N	Leaves[a]	Total splitting variables[a]	Iterations that split on each variable (%)					
				x1	x2	x3	x4	x5	x6
CART	2000	4.056 (2, 9)	2.52 (1, 5)	71.9	10.9	59.2	61.4	38.1	10.5
	5000	4.129 (2,7)	2.705 (1, 5)	75.5	10.5	64.5	68.2	48.7	3.1
	50000	3.894 (2,6)	2.644 (1, 4)	74.7	5.7	64.8	67.2	52.0	0.0
	200000	3.801 (2,6)	2.614 (1,4)	76.4	5.2	64.6	66.2	49.0	0.0
CTree	2000	8.401 (4, 14)	3.989 (2, 6)	91.9	57.0	83.2	83.8	71.6	11.4
	5000	13.947 (6,23)	4.772 (3,6)	97.7	82.5	92.8	94.5	91.1	18.6
	50000	47.354 (21,74)	5.581 (5,6)	100.0	99.9	100.0	99.9	100.0	58.3
	200000	88.641 (39, 130)	5.855 (5, 6)	100.0	100.0	100.0	100.0	100.0	85.5

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.25: Model 1 **(continuous outcome, categorical inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	90 (82, 96)	616 (190, 1159)	231 (57, 679)	651 (66, 1820)	698 (66, 1849)	339 (54, 983)	72 (60, 85)
5000	96 (88, 102)	1465 (380, 2797)	545 (94, 1734)	1607 (124, 4698)	1685 (121, 4638)	868 (90, 2544)	81 (68, 95)
50000	96 (89, 101)	14316 (3340, 27892)	4851 (518, 16613)	15146 (832, 45276)	15588 (803, 46735)	7948 (428, 24138)	85 (72, 100)
200000	95 (88, 100)	57416 (13203, 110020)	19919 (1926, 63578)	65800 (2822, 181636)	64650 (3346, 180842)	31567 (1592, 98846)	85 (71, 99)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.26: Model 2 (**continuous outcome, mixed inputs**) random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	367 (228, 474)	2731 (484, 6882)	308 (60, 881)	642 (76, 1778)	635 (78, 1791)	325 (66, 940)	552 (379, 691)
5000	669 (376, 935)	6429 (937, 17104)	761 (126, 2164)	1671 (166, 4598)	1632 (168, 4530)	819 (137, 2307)	1062 (667, 1386)
50000	2475 (1151, 3995)	60562 (4298, 161504)	7597 (710, 21520)	16163 (892, 47476)	16464 (941, 46274)	8416 (708, 25399)	4524 (2319, 6766)
200000	4865 (2151, 8009)	239953 (12550, 629907)	31906 (2139, 90952)	63981 (3170, 185378)	63664 (3426, 186800)	32870 (2005, 97326)	9558 (4586, 14714)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.27: Model 3 **(Common binary outcome, categorical inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	74 (61, 85)	18 (11, 28)	8 (5, 13)	16 (7, 37)	17 (7, 38)	11 (6, 22)	12 (8, 17)
5000	88 (74, 98)	26 (15, 44)	12 (7, 21)	35 (10, 82)	38 (10, 86)	21 (8, 50)	14 (9, 21)
50000	96 (89, 101)	124 (47, 255)	51 (15, 126)	308 (42, 822)	338 (44, 862)	166 (21, 486)	17 (10, 26)
200000	95 (89, 100)	442 (151, 963)	170 (34, 449)	1294 (157, 3504)	1361 (170, 3360)	645 (58, 1889)	17 (10, 26)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.28: Model 4 **(Common binary outcome, mixed inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	136 (94, 182)	112 (71, 166)	8 (4, 12)	12 (6, 25)	13 (6, 27)	9 (5, 15)	101 (67, 144)
5000	257 (172, 373)	226 (135, 353)	14 (8, 23)	26 (11, 58)	27 (12, 61)	17 (9, 32)	204 (129, 309)
50000	875 (501, 1452)	938 (497, 1672)	61 (26, 126)	211 (48, 525)	223 (52, 567)	124 (34, 302)	782 (434, 1353)
200000	1472 (804, 2547)	1902 (943, 3607)	184 (56, 449)	822 (124, 2087)	879 (146, 2234)	435 (72, 1131)	1392 (755, 2512)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.29: Model 5 **(rare binary outcome, categorical inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	48 (36, 61)	4 (2, 6)	2 (1, 3)	2 (1, 3)	2 (1, 3)	2 (1, 2)	3 (2, 4)
5000	66 (52, 80)	4 (3, 7)	2 (1, 3)	3 (2, 5)	3 (2, 5)	2 (1, 3)	3 (2, 5)
50000	92 (84, 99)	9 (5, 15)	4 (2, 7)	19 (5, 43)	21 (6, 45)	10 (3, 24)	4 (2, 6)
200000	95 (90, 99)	25 (12, 45)	10 (4, 19)	77 (21, 169)	82 (22, 177)	39 (9, 89)	4 (3, 6)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.30: Model 6 **(rare binary outcome, mixed inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	58 (41, 79)	33 (22, 47)	1 (1, 2)	1 (1, 2)	1 (1, 2)	1 (1, 2)	32 (21, 45)
5000	121 (84, 167)	74 (50, 103)	3 (1, 4)	3 (1, 4)	3 (2, 4)	2 (1, 3)	70 (48, 97)
50000	535 (332, 838)	376 (227, 594)	10 (5, 16)	15 (7, 28)	16 (8, 31)	11 (6, 17)	352 (217, 549)
200000	1025 (662, 1608)	793 (465, 1248)	21 (10, 37)	54 (19, 112)	58 (21, 118)	32 (13, 65)	712 (429, 1121)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.31: Model 7 **(multinomial outcome, categorical inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	84 (80. 89)	38 (31, 45)	18 (14, 22)	23 (13, 53)	24 (13, 53)	20 (13, 38)	29 (24, 33)
5000	94 (89, 100)	49 (39, 61)	24 (19, 32)	45 (17, 114)	46 (17, 125)	35 (17, 89)	35 (29, 40)
50000	94 (88, 100)	147 (82, 262)	73 (34, 171)	403 (46, 1181)	400 (46, 1193)	264 (33, 849)	41 (34, 48)
200000	93 (87, 99)	471 (213, 914)	220 (66, 623)	1577 (173, 4557)	1548 (156, 4819)	1030 (73, 3295)	42 (34, 49)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.32: Model 8 **(multinomial outcome, mixed inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	225 (191, 255)	243 (209, 275)	22 (18, 26)	27 (17, 55)	27 (17, 55)	24 (18, 40)	227 (194, 253)
5000	420 (336, 513)	489 (394, 581)	40 (32, 47)	59 (29, 139)	56 (29, 128)	46 (30, 86)	450 (368, 539)
50000	1376 (1015, 1910)	2044 (1495, 2714)	135 (100, 194)	473 (117, 1296)	468 (109, 1232)	299 (104, 900)	1672 (1264, 2238)
200000	2323 (1700, 3271)	4438 (2921, 6156)	334 (199, 594)	1886 (289, 5162)	1770 (268, 4986)	1143 (214, 3559)	3002 (2248, 4130)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.33: Model 9 **(negative binomial outcome, categorical inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	88 (80, 95)	129 (63, 243)	58 (29, 115)	138 (45, 328)	156 (45, 381)	86 (34, 200)	79 (42, 143)
5000	95 (87, 102)	197 (87, 381)	87 (37, 187)	330 (70, 831)	351 (70, 864)	184 (50, 489)	89 (45, 155)
50000	95 (89, 101)	1190 (391, 2495)	483 (120, 1272)	2950 (341, 8151)	3271 (378, 8718)	1628 (135, 4906)	103 (53, 189)
200000	94 (88, 99)	4429 (1355, 9862)	1702 (298, 4556)	12409 (1326, 35011)	13415 (1405, 34690)	6264 (468, 19009)	102 (53, 182)

[a] Means presented with 2.5th and 97.5th percentiles

Table 4.34: Model 10 **(negative binomial outcome, mixed inputs)** random forest outcomes

N	Leaves[a]	Variable importance[a]					
		x1	**x2**	**x3**	**x4**	**x5**	**x6**
2000	278 (193, 171)	840 (396, 1614)	62 (28, 119)	108 (45, 254)	115 (46, 258)	69 (33, 140)	675 (367, 1180)
5000	506 (329, 730)	1698 (752, 3401)	119 (51, 232)	254 (82, 634)	268 (88, 624)	151 (65, 344)	1336 (696, 2487)
50000	1754 (895, 2942)	9022 (2901, 21228)	671 (204, 1575)	2245 (398, 5826)	2370 (430, 6085)	1288 (264, 3387)	5743 (2334, 12256)
200000	3146 (1561, 5339)	24544 (6258, 62111)	2428 (470, 6204)	8783 (1257, 23036)	9571 (1307, 24395)	4798 (617, 13137)	11732 (4425, 26733)

[a] Means presented with 2.5th and 97.5th percentiles

4.5 Run time assessment

To assess feasibility, run times for single analyses of categorical input models with four different outcomes were analysed. Results are presented in Table 4.35. Most analyses were fairly quick, with run times around 5 minutes or less, and the majority running in less than one second. Notably, regression methods for the negative binomial outcome took much longer with increasing sample size, with the over-specified regression method requiring over 8 hours to be completed.

Table 4.35: Run time (HH:MM:SS) for a single iteration with categorical inputs

	Method	N=2000	N=5000	N=50000	N=200000
Continuous	CART	< 1 sec	< 1 sec	< 1 sec	00:00:02
	CTree	< 1 sec	< 1 sec	< 1 sec	00:00:01
	Random forest	< 1 sec	< 1 sec	00:00:04	00:00:29
	Best-fitted regression	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	Over-specified regression	< 1 sec	< 1 sec	00:00:02	00:00:07
	Cross-classification	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	MAIHDA	00:00:03	00:00:05	00:00:42	00:03:32
Binary - common prevalence	CART	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	CTree	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	Random forest	< 1 sec	< 1 sec	00:00:03	00:00:23
	CHAID	< 1 sec	< 1 sec	00:00:06	00:00:23
	Best-fitted regression	< 1 sec	< 1 sec	< 1 sec	00:00:02
	Over-specified regression	00:00:01	00:00:02	00:00:22	00:00:48
	Cross-classification	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	MAIHDA	00:00:01	00:00:02	00:00:25	00:02:18
Multinomial	CART	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	CTree	< 1 sec	< 1 sec	< 1 sec	00:00:02
	Random forest	< 1 sec	< 1 sec	00:00:03	00:00:23
	CHAID	< 1 sec	< 1 sec	00:00:06	00:00:33
	Best-fitted regression	< 1 sec	< 1 sec	00:00:07	00:00:29
	Over-specified regression	00:00:13	00:00:34	00:04:45	00:22:04
	Cross-classification	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	MAIHDA	-	-	-	-
Negative binomial	CART	< 1 sec	< 1 sec	< 1 sec	00:00:02
	CTree	< 1 sec	< 1 sec	< 1 sec	00:00:01
	Random forest	< 1 sec	< 1 sec	00:00:03	00:00:28
	Best-fitted regression	00:00:02	00:00:03	00:00:25	00:01:45
	Over-specified regression	00:01:19	00:06:21	01:52:22	08:28:12
	Cross-classification	< 1 sec	< 1 sec	< 1 sec	< 1 sec
	MAIHDA	00:00:37	00:01:25	00:14:25	00:52:55

Chapter 5

5 Discussion

This section will summarize the capabilities of each method, including ability to create accurate predictions for intersectional groupings, and variable identification. Recommendations are provided for the application of each method and the context for using these methods in descriptive intersectionality is further discussed. Study strengths, limitations, and further points for future work are also considered.

5.1 Primary outcome recommendations of methods

Results from this thesis indicate that there are many options for quantitative intercategorical-intersectional analyses, and choices may vary based on the dataset being analyzed. To begin the discussion of the results, Table 5.1 summarizes what each method produces, as initially outlined in the introduction. This can be contrasted with Table 5.2, which summarizes results from this study regarding performance for prediction, effect size estimation, variable identification, and type 1 error. Variable identification occurs in different manners for the different methods: observing results of significance testing for single-level regression and MAIHDA, by identifying splitting variables for CART, CTree, and CHAID, and by comparing the variable importance measure for random forest. Further description on performance for prediction is described below, while further detailed in Section 5.2 are the circumstances under which methods are and are not suitable for the other applications described in Table 5.2.

Table 5.1 Summary of method characteristics

	Regression with interaction terms	Cross-classification	MAIHDA	CART	CTree	CHAID	Random forest
Create intersection predictions	X	X	X	X	X	X	X
Hypothesis testing	X		X		X	X	
Effect size estimates	X		X				
Not cumbersome to include high number of intersections			X	X	X	X	X
Use continuous variables without categorization	X			X	X		X
Visual subgroup identification				X	X	X	

Table 5.2 Summary of key study results

	Regression using interaction terms	Cross-classification	MAIHDA	CART	CTree	CHAID	Random forest
Prediction at small sample size (n=2,000)	N	N	Y	N	Y	Y	Y
Prediction at large sample size (n=200,000)	Y	Y	Y	N	Y	Y	Y
Validity of estimates- 1st level effects (and interactions)	M	NA	M	NA	NA	NA	NA
Variable identification (detect variables significant or important to outcome with increasing sample size)	M	NA	N	N	Y	Y	M
Low type 1 error	Y	NA	Y	Y	N	N	M

Y – Yes: Generally suitable N – No: Not recommended M – Maybe: Okay under certain circumstances

With increasing sample size, all methods improved their performance for creating accurate predictions, with the exception of CART. At the smaller sample sizes, it appears more critical to select appropriate methods for accurate predictions. Table 5.3 highlights the top performers for creating accurate intersectional predictions, by outcome type and input type as well as sample size. This table includes best-fitted regression only for theoretical reasons rather than for practical recommendation, given that researchers cannot know that they are specifying this model. Generally, random forest and MAIHDA created the most accurate predictions at small sample sizes. Results from this project reveal that the most common quantitative intersectional analysis methods, regression with interaction terms and cross-classification, are not reliable methods to create accurate intersectional estimates when sample size is small, and the number of intersections is large. Additionally, over-specified regression is not a viable option at smaller sample sizes for binary or negative binomial outcomes, given the convergence issues when estimating a large number of coefficients. While Table 5.3 presents recommendations at two extreme sample sizes, for most outcomes prediction accuracy of the methods was similar between sample sizes 50,000 and 200,000, except for the rare binomial outcome. Somewhere between n=5,000 and 50,000 many methods equalize in prediction, although that specific point in this instance was not determined.

Table 5.3 Methods that performed well for prediction

	Categorical inputs (n=2,000)	Mixed inputs (n=2,000)	Categorical inputs (n=200,000)	Mixed inputs (n=200,000)
Continuous outcome	Best-fitted regression Random forest MAIHDA	Random forest	CTree Random forest Best-fitted regression Over-specified regression Cross-classification	CTree Random forest Cross-classification MAIHDA

			MAIHDA	
Binary outcome (common)	Best-fitted regression MAIHDA	Best-fitted regression MAIHDA	CTree Random forest Best-fitted regression Over-specified regression Cross-classification MAIHDA	CTree Random forest Best-fitted regression Over-specified regression Cross-classification MAIHDA
Binary outcome (rare)	MAIHDA	MAIHDA	CHAID CTree Random forest Best-fitted regression MAIHDA	CTree Random forest Best-fitted regression Over-specified regression Cross-classification MAIHDA
Multinomial	CTree CHAID Random forest Best-fitted regression	CTree Random forest Best-fitted regression	CHAID CTree Random forest Best-fitted regression Over-specified regression	CTree Random forest Best-fitted regression Over-specified regression

			Cross-classification	Cross-classification
Negative binomial	MAIHDA Best-fitted regression	MAIHDA	CTree Random forest Best-fitted regression Over-specified regression Cross-classification MAIHDA	CTree Random forest Best-fitted regression Over-specified regression Cross-classification MAIHDA

The conclusion that random forest is superior for prediction compared to regression and single classification tree methods is in agreement with the existing decision tree literature. When conducting secondary data analysis for continuous outcomes, random forest has been shown to be superior to linear regression when comparing R-squared and root mean squared error. [40] For binary outcomes, while one study found random forest and logistic regression to perform similarly for classification [39], other studies have found that random forest creates more accurate predictions compared to CART [22, 36, 79], CHAID [76], and logistic regression [22, 79]. The poor predictive performance of CART is a striking result considering the use of CART as the primary decision tree method in the current intersectionality literature, and is discussed further in section 5.2.4. Results are however in agreement with those found by Venkatasubramaniam et. al. [42] where a simulation study found that CART did not improve in accuracy (measured by MSE), with increasing sample size beyond n=3,000, unlike CTree.

5.2 Summary and recommendations for each method

5.2.1 Regression (Over-specified)

Over-specified regression models require a cumbersome number of interaction terms

when interested in the potential for multiple social positions to experience unique intersectional effects, given that researchers cannot know *a priori* which interactions to include and exclude. This study evaluated a variety of over-specified regression models: OLS, modified Poisson, multinomial logistic, and negative binomial. Results from this study found that across sample sizes, the validity of estimates for main and interaction effects estimations were sufficient for models with only categorical input variables. Because the mixed input models were created with a non-linear interaction term which was not specified in the fitted regression models, prediction of the main effects and interactions involved in these interactions were suboptimal. The power to detect variables and interaction terms significant to the outcome was also low at small sample sizes, which is not surprising given the number of coefficients (64 or 192) being estimated. For the interaction terms, sometimes the probability of being detected as significant would decrease with increasing sample size, contrary to the expectation that power should increase with increasing sample size. A possible explanation is that with increasing sample size, the main effects became more likely to be detected, and resultantly lowering the probability of the interaction terms to be detected as significant. Hypothetically, we might expect that with even greater sample sizes the detection of the higher-level interaction terms would also approach 100%. Type 1 error was not an issue at the large sample sizes, but at smaller sample sizes the variable X6 could sometimes be detected more than 5% of the time. For example, for the multinomial outcome with categorial inputs at n=2,000, variable X6 was the detected as significant around 22% of the time (see Appendix B). This shows that fitting an over-specified regression does not protect results from Type 1 error at smaller sample sizes.

5.2.1.1 Recommendations for application

Overall, over-specified regression is not suggested as a viable option for prediction or variable selection, at least at smaller sample sizes when estimating high-dimensional intersections, formed with more than 2 or 3 intersectional variables. There are a few lessons however to take away from these results. The issues of non-linear effects and interactions can be accounted for in regression models, for example using generalized

additive models. [77] This can lead to better estimation of non-linear interaction effects, and better estimation of the main effects involved in these interactions as well. This is an important consideration for intersectionality research, because there is no reason to assume that interaction effects will always be linear. Secondly, the issues of poor power to detect interactions, and general lack of convergence for these over-specified models, could be addressed by applying stepwise regression or backwards selection, to remove insignificant coefficients, improve convergence issues, and reduce over-fitting. For example via backwards elimination, insignificant variables are removed one by one until all remaining variables are significant to the outcome. [78] However, it is cautioned that this data-driven approach can result in the selection of variables that have spurious associations, or the removal of variables or interaction terms that are actually significant to the outcome. [79] This methodology should be applied with caution as it can result in overfitting to the sample and make prediction for populations outside of the sample less accurate. [79]

5.2.2 Cross-classification

Cross-classification has the advantage of being a simple, easy to understand descriptive method for predicting outcomes for intersectional groupings, by simply calculating the average value of the outcome in each intersection, with no further adjustment. But when looking at multiple intersecting positions, cell sizes run the risk of becoming too small to create accurate predictions. With no mechanism to account for the random error that is likely with small sample sizes, cross-classification is prone to outliers for continuous and count outcomes, or does not have enough events to approximate the true proportions for binary and categorical outcomes.

5.2.2.1 Recommendations for application

Use of cross-classification is best when there are a small number of intersectional groupings being created, or a relatively large sample size to reduce the likelihood of inaccurate or misleading estimates, especially for small intersections. A viable alternative to cross-classification, if the goal is to simply describe each intersectional group, could be

a MAIHDA analysis where only the intersection predictions are interpreted. Both create predictions for each intersectional grouping based on the observed sample, but the shrinkage of residuals in MAIHDA results in much greater prediction accuracy.

5.2.3 MAIHDA

This study was the first to evaluate MAIHDA for its effectiveness for prediction, and it is evident that predictions with shrinkage are effective for improving prediction accuracy, even for datasets with small sample sizes. Regarding validity of the main effect estimates, this study verified that the estimates do not follow those of a traditional single-level regression model with interaction terms, and the stratum-average effects interpretation as proposed by Lizotte et. al. [53] is only true for continuous and negative binomial regression models. It is unknown what the expected main effect estimate would be for binary outcomes. Regarding the power to detect significant main effects, because the estimates do not follow the traditional definition, they also were not always significant when expected. The type 1 error was however reliably low.

5.2.3.1 Recommendations for application:

Given the good performance for prediction but not for variable effect estimation or power, the suggested approach would be to use MAIHDA for prediction and outcome mapping, but to not interpret the main effects. For continuous or negative binomial outcomes, main effects could technically be interpreted as outlined by Lizotte et. al., but even so there is no established definition of what the intercept effect would be. Given that the intersectional residuals are to represent any intersectional effects beyond the additive model, it is also not recommended to interpret the intersection residuals on their own, given that it is unclear what the baseline additive effects mean.

While the main effects are not for interpretation, they should still be included in the MAIHDA model, as opposed to a null model being fitted with only the random intercepts. This is because of the impact of the main effects on residual shrinkage. Bell et. al. [80] have noted that the residuals break the assumption of being independent and

identically distributed, and instead may be related to one another since they are determined by the main effects. Therefore, the authors assessed if shrinkage is truly able to account for multiple testing, by focusing on the significance of residuals. The shrinkage formula for the residuals is

$$u_j = r_j \times \frac{\sigma_u^2}{\sigma_u^2 + (\frac{\sigma_e^2}{n_j})}$$

where for intersection j, u_j is the shrunken residual, r_j is the raw (unshrunken) residual, n_j is the cell size of the intersection, σ_u^2 is the level 2 between-intersection variation, and σ_e^2 is the level one within-intersection variance. Bell et. al. [80] conclude via a simulation study that shrinkage is better able to reduce the spurious detection of significant residuals if the fitted model includes main effects, because inclusion of the main effects reduces the level 2 variance. They note however that if there are true interaction effects between variables, (e.g. a two-way interaction), this variance will not be included in the level 2. They suggest that to have optimal shrinkage, interaction effects should be added to the model fixed effects (first two-way, three-way, etc.), until level 2 variance reaches zero. If through this process all interactions are included until the highest level (e.g. four way), and level 2 variance is still greater than zero, then there is an intersectional effect occurring between all positions. One problem here is that issues with multiple testing are re-introduced when adding more interactions to the fixed effects, but the authors think it unlikely that the number of included interactions will reach a point where this is a major concern. In comparison to the current study, Bell et. al. [80] focus on significance of residuals rather than prediction, but since shrinkage applies to prediction accuracy, it is possible that prediction of MAIHDA could also be improved by the systematic inclusion of interaction terms. However, prediction without the inclusion of interaction terms seemed sufficient for creating accurate predictions for the simulations conducted in this study. Regardless, it is at least important to include the main effects in MAIHDA models for prediction to maintain shrinkage of the residuals. This is a point worth making because other applications of MAIHDA have presented intersection predictions from the null model, rather than the model including all main effects. [45, 48]

One limitation of MAIHDA is that continuous variables must be categorized to be included in the formation of the intersectional groups. For example, MAIHDA studies have typically split a continuous income variable into tertiles or quartiles. [44] However, continuous variables can be kept as continuous, if simply being used to adjust the models and resulting predictions, rather than used in the formation of stratum. [46] This assumes that the effect of a certain variable remains consistent between all intersections.

Applications of MAIHDA have also extended beyond the original model proposed by Evans et. al. [43]. Evans [81] has suggested an update to the MAIHDA method, where the intersectional strata are also created with contextual factors, using group-level variables. For example, a combination of gender, race, parental education, along with neighbourhood- and school-level poverty can be used to create MAIHDA intersectional groupings. [81] Another study has used MAIHDA where the second-level effects represent the more traditional application of multi-level modelling, by using country as the random intercept for multi-country data. [82] Intersections in this case were not determined by the first-level effects, but the authors still considered this an application of MAIHDA because they calculated the model's discriminatory accuracy. Our discussion of MAIHDA is limited to only when the random intercepts are fully determined by the first-level effects, and include no other contextual variables.

5.2.4 CART

While CART has been the dominant application of decision trees in the intersectionality literature thus far, this study found that CART usually performed poorly for both creating accurate intersection predictions and variable selection. CART models often did not split at all for binary classification problems, regardless of if outcomes were of a rare or common prevalence. These results were surprising given the use of CART in the intersectionality literature, to successfully create binary classification models. For example, one study with a sample size of less than 10,000 presented a resulting tree with eight splitting variables and 13 terminal nodes. [32] Additionally, given that for example the previously mentioned study [32] used the same R package and pruning criteria to

create CART models as in this study, it is unlikely that the CART model-building criteria is the reason for the lack of splitting results in this study. An alternative explanation is that results from the literature review differ from those found in this study due to the simulation data generation process. The range of effect sizes allowed in the binary outcome simulations may not have been enough to trigger splitting for the CART algorithm, given that power calculations were centered around what was detectable for a regression model, not a CART model. For example, when looking at a study which included both a CART and regression model, the main effects split on by the CART model often had a higher ratio of effect size to standard error. [31] Further exploration of the simulations after the completion of this study have suggested that increasing the simulated effect sizes does lead to an increase in splitting, but this remains to be further explored. One advantage was that CART did have the lowest type 1 error compared to CTree and CHAID. Given the overall poor variable selection, especially for binary outcomes but also for other outcome types, CART was effective neither for prediction (at small or large sample sizes), or subgroup identification. Potentially, CART may be effective if effect sizes are of a larger magnitude, but this threshold would be higher than for other decision tree methods or single-level regressions.

5.2.4.1 Recommendations for application

Overall, results from this study do not suggest using CART for prediction of individual intersection outcomes. If the primary interest of a user is to look at splitting patterns or subgroup identification, using CART models for a single decision tree analysis will likely provide a more conservative splitting pattern and fewer subgroups compared to the other single decision tree methods, or may result in no splitting at all. If type 1 error is a large concern, CART may be the safest decision tree option, however we would still suggest contrasting results from CART to other decision tree algorithms or pairing it with another type of analysis, such as regression, to obtain a more representative picture of variables influencing the outcome. As well, the type 1 error of CART when modelling binary outcomes should be further explored when simulated effect sizes are larger

5.2.5 CTree and CHAID

Overall, prediction accuracy for CTree and CHAID was as good as random forest and MAIHDA at larger sample sizes (at least n=50,000), and was sufficient, but not as accurate, at smaller sample sizes. Splitting patterns were highly similar between CHAID and CTree. Splitting by CTree and CHAID was more sensitive than CART, resulting in better identification of relevant variables. For subgroup identification, one issue is that these methods were likely not correctly identifying all subgroups, given that the number of final nodes was lower than expected. For example, while all variables could be used in splitting, this does not mean that each variable was split on one another to create every unique intersection. While it has been suggested that CTree, under the conditional inference framework, would result in less over-fitting and reduce the selection bias of splitting on continuous variables [33], both CTree and CHAID resulted in a high type 1 error, especially with increasing sample size. This was true for variables with no true effect that were continuous or three categories.

5.2.5.1 Recommendations for application

While both methods perform well for prediction and subgroup exploration, especially when compared to CART, users should be wary that not all variables included in the tree are necessarily relevant to the outcome. For CTree, the issue of high type 1 error may be mitigated if with larger sample sizes, a lower p-value is used for the selection process (e.g. $p<0.01$ or $p<0.001$). Across all outcome types, between n=50,000 and 200,000 there was minimal improvement in the selection of variables important to the outcome, but the type-1 error continued to increase. Lowering the p-value threshold at least for n=50,000 and beyond may still allow for sufficient splitting on significant effects, but limit the increases in Type 1 error. For CHAID, tuning was not performed in this study, because it is not readily available within the R package. However, tuning can be performed using the "caret" package [83], and similarly may result in lower type 1 error if thresholds are adjusted. However, both these options would have to be further explored.

5.2.6 Random forest

Random forest performed very well for creating accurate intersection predictions and was a reliable decision tree method for this application. Variable identification using the variable importance measure worked well across sample sizes for the continuous outcome model with categorical inputs, or binary, multinomial, and negative binomial outcomes with categorical inputs at larger sample sizes. However, the variable importance measure appears to fall under a similar bias as has been reported for CART, where continuous variables are favoured during splitting. [84] Strobl et. al. [84] found that the variable importance measure for the R package "randomforest", arguably the most popular random forest package in R, is biased towards splitting on variables with more splitting options, which could be continuous or categorical variables with many categories. While this study did not use the package randomforest, the package "ranger" does similarly rely on the Gini Index for splitting criteria, which the authors of this paper identify as the potential issue.

5.2.6.1 Recommendation for application

While the application of random forest models in the current study functioned well for creating predictions, there are a few alternatives to consider if researchers are concerned regarding the capacity for variable identification. Strobl et. al. [84] suggest creating random forest models using the package "cforest", which is based on the conditional inference framework, and show that the variable importance measure for this algorithm is less biased, if also combined with subsampling replacement when creating each bootstrapped sample. Alternatively, Altmann et. al. [85] provides a method for correcting the variable importance measure bias, without requiring trees be built under the conditional inference framework. The adjustment by Altmann et. al. [85] also provides a p-value to each variable importance, to improve interpretability when trying to identify significant variables.

5.2.7 General comments on the application of decision trees

While decision trees do not isolate the effects of any one variable, the visualization created by single decision trees such as CART, CTree, and CHAID can accompany regression models as seen in the examples in the introduction. [29, 31, 32] Random forest also does not create one single decision tree, so visualization would require accompaniment by a single decision tree method. The pairing of decision trees with regression models may help visualize subgroups, inform interaction terms to be included in the model, or also in the case of random forest, inform on which variables to include in the regression model. If researchers are interested in centering the analysis around differences between certain groups (like male and female), separate decision trees can be created for each subsample. [32] One issue with the application of decision trees is that splitting on a continuous variable can create hundreds of final nodes, as seen in the CTree and random forest results in this study, when dealing with a combination of categorical and continuous input variables. This is more categories than can be feasibly visualized or are informative for subgroup identification. There is however the possibility to "adjust for covariates" by making the outcome the residuals from an adjusted regression model. [42] This is a way to include continuous variables and reduce the number of nodes required to create an accurate prediction, but only under the assumption that effects of the variable are strictly linear with no interaction between any other variables. Finally, it should be acknowledged that the application of decision trees in this study was limited to only six input variables, differing from the more typical applications of decision trees which use a much longer list of input variables. The application of decision trees to intersectionality assumes a level of theory-based decision making regarding which variables to input into the model, while decision trees in reality are usually given a long list of input variables to be narrowed down from a data-driven perspective. Researchers should keep mind the balance between theory- and data-driven variable selection when using decision tree methods.

5.3 Considerations for applying methods to intersectionality research

While the previous sections discussed how each method can be individually applied, this next section will discuss general considerations in the application of descriptive intercategorical-intersectionality. The primary outcome of this study focused on the accuracy of predicting outcomes for each intersectional grouping, which can be used to understand the extent of a problem in different intersectional groups, and to identify groups for further study or further intervention. But before considering these groups intervenable from a policy or public health perspective, other considerations need to be taken into account.

The first consideration is that variables (e.g. gender, ethnicity) from an intersectional perspective are meant to represent structural-level effects, rather than individual effects. This is in alignment with the goals of ecosocial theory, to consider the contextual and structural factors that each individual interacts with in their environment. [86] Exploration of outcomes for marginalized groups can lead to further stigmatization if social positions or identities are seen as individual-level variables, by placing the responsibility for the outcome on the individual. MAIHDA explicitly frames the second-level effects as representations of the intersectional or contextual effects for the social positions. [47] This thought process can still be applied when looking at variables inputted in a decision tree or regression model, by being mindful that they represent a greater contextual effect.

The second consideration is the interpretability or relevance of identified subgroups. Subgroups created by decision trees to identify risk profiles may not be fully interpretable or applicable to actual policy intervention, and may not represent targetable groups that relate to identity or community. Either a method may create too many subgroups by splitting many times on a continuous outcome (e.g. from the mixed input models, creating hundreds of final nodes), or the splitting points for continuous variables are too specific to act on. For example, the study by Sridharan et. al. [31] looking at antenatal

care usage presented a CART model involving two nodes where the splitting criteria was a household wealth index being less than or equal to 4.839 or greater than 4.839. This resulted in nodes with outcome proportions of 58.3% and 82.4%. The splitting on 4.839 is somewhat arbitrary, and although the difference in outcome between the two groups is quite large, it is not necessarily a cut-off informative for use in clinical practice or in designing public health interventions. The values of these cut-offs may change due to the instability of single decision trees, and decision rules could change with small changes in the sample data used to build the model. [38] Subgroups created from decision trees should not be seen as definitive, and an important step to ensuring validity outside of the sample data is to assess if the final decision tree model is still sufficient when applied to a different dataset. [41]

The third consideration is to consider not only the average effects in each intersection, but also the size of the variation within each one. Merlo et. al. [87] refers to this as the "tyranny of averages", where focus on mean outcomes within a group can potentially be harmful or further stigmatizing if the within group variation is disregarded. For example, specific interventions for "high risk" groups may be inappropriate if the within group variation is large, and a large portion of the members of this subgroup are not at greater risk than the general population. To look at the heterogeneity in subgroups created by decision trees, terminal nodes can for example be represented by boxplots to visualize the outcome distribution in each node. Venkatasubramaniam et. al. [42] created a visualization tool to look at the subgroups created by a decision tree, to better understand the outcome spread as well as the variables that make up the nodes. While not addressed in this study, assessment of discriminatory accuracy is a key component of MAIHDA papers, evaluating the heterogeneity within versus between intersections. [44] MAIHDA authors have suggested that intersections that are calculated to have a low discriminatory accuracy should not be regarded as intervenable targets, because these categories are not good at predicting the outcome for all individuals in an intersection, due to a large level of heterogeneity. [44]

There is also an important distinction to be made between prediction and causality. The

authors Kreatsoulas and Subramanian [41] touch on the challenges faced when incorporating machine learning into social epidemiology, and discuss how the goals of the analysis should be compatible with the "the underlying mathematical skeleton of the optimization theory". For methods like random forest which are more complex to understand mathematically, it may be difficult to directly understand the impact of variables on the outcome, so it should not be used with this goal in mind. While machine learning methods perform well for creating predictions, they are not designed to understand causal relationships, and results of analyses should be interpreted with this in mind. Additionally, descriptive studies often work with cross-sectional data that lack temporality, and certain variables representing identities or social positions, such as gender, or race/ethnicity, are non-intervenable. While the current study assessed methods for their ability to create predictions, these predictions were ultimately descriptive. When identifying intervenable factors, such as discrimination, methods such as intersectional mediation analysis can be applied to conduct analytic intersectionality research, and assess causality. [88]

Finally, researchers should be aware that biases in the data can perpetuate existing disadvantage. [41, 91] For example, if the data are not representative of the population due to selection or reporting biases, those biases will be maintained in the predictive models. Additionally for machine learning methods, there are limitations to how well cross-validation can explain model performance and generalizability. Because cross-validation usually occurs using a validation set that is a random selection from the same dataset used to create the model, any biases or underrepresentations in the data will remain undetected. One of the solutions for this, as previously mentioned, is to validate the completed model against a different dataset. [41] Additionally, a marker of fairness when applying machine learning for the purposes of health equity is equal performance. [89] Equal performance means that outcomes are estimated with equal accuracy for advantaged and disadvantaged groups. The implications of a low sensitivity, specificity, or positive predictive value may be harmful and further marginalize already disadvantaged groups, by either over- or under-stating the outcome. This can apply not only to machine learning methodologies, but to the other methods used in this study for

prediction. We attempted to consider equal performance by equally weighting the performance of each intersection when calculating accuracy, to avoid prioritizing accuracy for larger intersections.

5.4 Survey of method feasibility

This project also served as a survey of the current capabilities of analyses in R. We identified a lack of R packages for frequentist multilevel multinomial regression analysis for random-intercept only models. Although not used in this study due to the practical time constraints of conducting thousands of Bayesian analyses, applications of MAIHDA for a multinomial outcome could still be conducted using R using Bayesian analysis under the package "brms" [66]. Additionally, application of a modified Poisson multi-level regression in R is also limited, given the lack of packages allowing for the appropriate adjustment of standard errors. Regarding the feasibility of running these analyses, none of the analyses required a prohibitive amount of computing resources if running a single iteration. When assessing run times for a single analysis, the longest was for the over-specified negative binomial regression, but even this at a sample size of 200,000 could be conducted overnight on a standard PC. The major reason for the extensive run time of this analysis was the creation of confidence intervals for all the coefficients. Computationally, there is little limitation for researchers to consider these alternative methods.

5.5 Strengths and limitations

This study had notable strengths and limitations worthy of discussion. A major strength of this study was the use of simulated data to assess the different quantitative methods. With simulated data, the true outcomes for each intersection and effect size are known, and we are able to assess both the validity and accuracy of estimates. There are many studies in the literature which compare methods using secondary data analysis, where the true outcomes are unknown. [39, 40, 52] In these cases, it is unknown which method is actually approaching the true population estimates, because two methods may hold the

same biases or under-/over-fitting issues. As well, each simulated data scenario and sample size were iterated 1,000 times with varying effect sizes, and this allows for greater confidence that the observed results and patterns are consistent.

This is the first study of its kind to look across the intersectionality literature and assess these different methods of varying complexities. While decision tree methods have been compared against one another [36, 42], and decision trees and MAIHDA have been separately compared to regression analysis [39, 40, 52], no comparison has yet been drawn between cross-classification, regression, MAIHDA, and decision trees. We focused not only on accuracy of predictions, but also on variable identification and effect size estimation, to reflect how these studies are applied and interpreted in the literature. This allowed for improved understanding of when and how to apply certain methodologies, because methods performed differently for different objectives. For example, while this study identified that MAIHDA performs well for the prediction of binary outcomes, it was also further identified that the interpretation of the main effects of MAIHDA is still unknown for binary outcomes, and do not fall under any previously proposed definitions.

An extensive variety of dataset qualities were considered in this analysis, to reflect the variety of datasets used in intersectional and population health research. The existing simulation studies in this study's literature review, while varying by sample size and data generation processes, were limited to continuous outcomes. [42, 82] This study focused on looking across different sample sizes, outcome types and input types, because these are all dataset qualities that a researcher will know *a priori*. We were able to see that there is no singular method best-suited for all data scenarios, and provide a more comprehensive guide for researchers based on particular dataset qualities and research objectives.

There are also certain limitations that should be discussed. Regarding the process of creating the simulations, only five variables had any true effect on the outcome, and all five of these variables were included in the fitting of the regression, MAIHDA, cross-

classification and decision tree models. This differs from the reality of intersectionality research, where the selected social identities or positions are likely not the sole explanatory variables for an outcome. It is typically expected that there are other individual- and structural-level effects not accounted for in the analysis. We did incorporate some unknown elements, like correlation between variables X3 and X4 via mediation, and non-linear interaction effects, but all variables with an effect on the outcome were included when fitting the analysis models, which is an unrealistic expectation. Therefore, the accuracy of the intersection-level predictions in the results of this study should not be expected to be the true accuracy of the individual-level predictions, if conducting analysis on an actual dataset. This simulation represents the best-case scenario, and a real-world analysis would have unaccounted for predictors increasing the variation between individuals within an intersection. An additional concern regarding the data simulations is that for the simulation of the binary outcome, resampling of the variable effect sizes was required if the probability of the outcome exceed 100%. Because this was required for approximately half of the 1,000 iterations when simulating the common binary outcome with mixed inputs, the random sampling of effect size estimates for this data scenario were possibly not as random as for the other scenarios. Finally, there are some concerns when using MAPE for the assessment of accuracy for the binary and multinomial outcomes. Because the difference between the predicted and actual prevalence is standardized over the actual prevalence, MAPE can create excessively large errors when predicting outcomes with a very low true prevalence, such as the rare binary outcome in this simulation. Therefore, the measure may be biased towards favouring methods that perform particularly well for predicting outcomes for intersections with smaller outcome prevalences, because the impact of a poor prediction for a rare outcome is larger than that for a common outcome.

Regarding the application of the decision tree methods, the way the analyses were performed in this simulation are not meant to be definitive or seen as the best possible approach. In the building of decision trees, it would be typical to have a fitting process involving a training and test dataset, where model fit can be assessed and adjusted using different tuning parameters. Due to the size of the study and the requirement for many

iterations, our tuning process was limited to only one parameter for CART and random forest analysis. The use of different R packages, tuning parameters, and building parameters (like stopping rules) may yield different results. Finally, regarding the application of MAIHDA in this study, one notable limitation was the use of frequentist analysis rather than a Bayesian model with null priors, due to time and computational restraints. Although results from a short simulation presented in Appendix A indicate that main effect coefficient estimates are similar between the two approaches, it cannot be definitively said that the results found in this study regarding the accuracy of predictions will be the same under a Bayesian approach. Additionally, this study applied MAIHDA for a binary outcome using a multilevel Poisson regression, to mirror the use of the modified Poisson for the single-level regression. However, MAIHDA analyses for binary outcomes have typically used multilevel logistic regression. It is unclear whether differences in these methods would impact the predictive performance of MAIHDA for binary outcomes.

5.6 Directions for future work

While the use of simulated data is beneficial for understanding the accuracy and validity of estimates, results from this thesis would benefit from a demonstration of each of the methods under a real-life dataset, comparing the significant intersections identified by MAIHDA, the significant interactions identified by regression models, and the subgroups identified by CART, CTree, and CHAID. There were certain issues identified by this thesis that could also be further explored in future research. Given the concern when creating high-dimensional intersections that smaller intersections will suffer in terms of accuracy, further exploration of the predictions can specifically look at which methods perform well for predictions for smaller intersections, and if there are any clear patterns regarding if outcomes for small intersections are typically over- or under-estimated. This studied also identified that the binary effects produced by MAIHDA do not fall under the definitions of main effects that have been proposed by either Evans et. al. [52] or Lizotte et. al. [53]. Given the quick adaptation of MAIHDA in the intersectionality literature, and

the regular analysis of binary outcomes, the interpretation of these main effects should undergo further investigation.

In this thesis, the assessment of methods for quantitative intersectionality was strictly quantitative. However, as suggested in the discussion, interpretability of the methods also plays an important role for researchers choosing to do quantitative intersectional health research, given that the end goal of this work is to inform decisions in public health. Therefore, future work will include a qualitative analysis, assessing interpretability and usefulness of methods. Points to consider under qualitative analysis include how well the method is in agreement with intersectionality theory and the goals of intercategorical-intersectionality, how visible effects are for each intersection (is each intersection equally prioritized), and how well a large number of intersections can be incorporated into the method. While methods like decision trees may be helpful for visualization, these data-driven approaches may also provide subgroups that are not useful for further study. There are also alternative ways to conduct some of the methods that should be further explored, both for their predictive performance and impact on the interpretability of the results. For example, alternative methods for random forest analysis that produce less biased variable importance measures [86, 87] and provide p-values for the interpretation of the variable importance measure [85], may make the random forest method more user-friendly. The inclusion of interaction terms in the fixed-effect for MAIHDA as suggested by Bell et. al. [80], or the use of Generalized Additive Models for regression [77], may improve prediction, but may also contribute to a loss of interpretability. The tradeoff between accuracy and interpretability remains to be further explored.

5.7 Conclusion

This study aimed to understand how to best incorporate an intercategorical-intersectional perspective into quantitative health research, with a particular focus on methods able to assess a large number of high-dimensional intersections at the same time. Methods were assessed using simulated data scenarios varying by outcome type, input type, and sample size. Assessment of methods included prediction accuracy, identification of variables

important or significant to the outcome, and type 1 error. Different methods outperformed others, depending on both the data scenario and the objective. All methods improved in prediction accuracy with increasing sample size with the exception of CART, which often performed poorly at both large and small sample sizes. Random forest and MAIHDA generally created the most precise predictions at small sample sizes. CTree and CHAID were also generally suitable for creating predictions at small sample sizes, but typically less accurate.

CART did not perform well for variable selection for all outcome types, and especially for binary outcomes. These results were surprising given the use of CART for binary classification problems in the existing intersectionality literature. One explanation for this observed difference is that our simulations may not have had large enough effect sizes to pass CART's threshold for splitting. Variable selection was better for CHAID and CTree, but consistently faced a high type 1 error. Variable selection by random forest, according to the variable importance measure, worked well if input variables were all categorical, but if presented with continuous variables would result in a high type 1 error, due to bias towards selecting continuous variables. While MAIHDA performed well for prediction, MAIHDA coefficients had worse confidence interval coverage and lower power than traditional regression models with interaction terms. We identified that the definition of main effects for MAIHDA models with binary outcomes is unknown, and requires further investigation.

From this study emerge recommendations for researchers looking to use these methodologies for quantitative intersectionality research. We recommend that MAIHDA can be used for outcome mapping, but researchers should refrain from interpreting the main effects, or residual estimates. Random forest is also a viable option to create intersectional predictions, but the variable importance measure is biased if looking to identify variables significant to the outcome. Alternative random forest methods using either the conditional inference framework or corrections to the variable importance measure may be of interest to researchers. CTree and CHAID are more likely to identify relevant subgroups than CART, but given their high type 1 error, it may be of use to pair

these methods with a regression analysis. Finally, while regressions with interaction terms and cross-classification are the most common methodologies in the current intersectionality literature, they are not recommended for calculating outcomes for a large number of intersections, unless sample size is sufficiently large.

The goal of this work is to ultimately create a guide for quantitative intersectionality research. Accordingly, future research should combine this quantitative evaluation with a qualitative evaluation of the interpretability and usefulness of these different methods, to encourage the use of methods that are both statistically sound and in line with the theoretical basis of intersectionality research.

References

1. Marmot M, Allen JJ. Social determinants of health equity. American Journal of Public Health. 2014;104(S4):S517-9.

2. Link BG, Phelan J. Social conditions as fundamental causes of disease. Journal of health and social behavior. 1995 Jan 1:80-94.

3. Rose G. The strategy of preventive medicine. Oxford, Oxford University Press, 1992.

4. Greenland S. Commentary: Interactions in Epidemiology: Relevance, Identification, and Estimation. Epidemiology. 2009 Jan 1;20(1):14-7.

5. Crenshaw K. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. u. Chi. Legal f.. 1989:139.

6. Collins PH. Black feminist thought: Knowledge, consciousness, and the politics of empowerment.

7. Choo HY, Ferree MM. Practicing intersectionality in sociological research: A critical analysis of inclusions, interactions, and institutions in the study of inequalities. Sociological theory. 2010 Jun;28(2):129-49.

8. Rosenthal L. Incorporating intersectionality into psychology: An opportunity to promote social justice and equity. American Psychologist. 2016 Sep;71(6):474.

9. Schudde L. Heterogeneous effects in education: The promise and challenge of incorporating intersectionality into quantitative methodological approaches. Review of Research in Education. 2018 Mar;42(1):72-92.

10. Bowleg L. The problem with the phrase women and minorities: intersectionality—an important theoretical framework for public health. American journal of public health. 2012 Jul;102(7):1267-73.

11. Bauer GR. Incorporating intersectionality theory into population health research methodology: challenges and the potential to advance health equity. Social science & medicine. 2014 Jun 1;110:10-7.

12. Hancock AM. When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm. Perspectives on politics. 2007 Mar;5(1):63-79.

13. McCall L. The Complexity of Intersectionality. Signs: Journal of Women in Culture and Society. 2005 Mar;30(3):1771-800.

14. Dubrov JK. How Can We Account for Intersectionality in Quantitative Analysis of Survey Data? Empirical Illustration for Central and Eastern Europe. ASK. Research and Methods. 2008(17):65-100.

15. Veenstra G. Race, gender, class, and sexual orientation: intersecting axes of inequality and self-rated health in Canada. International journal for equity in health. 2011 Dec;10(1):3.

16. Bright LK, Malinsky D, Thompson M. Causally interpreting intersectionality theory. Philosophy of Science. 2016 Jan 1;83(1):60-81.

17. Jackson JW, Williams DR, VanderWeele TJ. Disparities at the intersection of marginalized groups. Social psychiatry and psychiatric epidemiology. 2016 Oct 1;51(10):1349-59.

18. Hancock AM. Empirical intersectionality: A tale of two approaches. UC Irvine L. Rev.. 2013;3:259.

19. Churchill SM, Rueda AV, Walwyn C, Mahendran M, Bauer GR. Quantitative intersectionality research methods: A systematic review. University of Western Ontario; 2019 (Unpublished).

20. Cummings JL, Braboy Jackson P. Race, gender, and SES disparities in self-assessed health, 1974-2004. Research on Aging. 2008 Mar;30(2):137-67.

21. Patterson AC, Veenstra G. Black-White health inequalities in Canada at the intersection of gender and immigration. Canadian Journal of Public Health. 2016 May 1;107(3):e278-84.

22. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. InProceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 161-168).

23. Kotsiantis SB. Decision trees: a recent overview. Artificial Intelligence Review. 2013 Apr 1;39(4):261-83.

24. Wolfson J, Venkatasubramaniam A. Branching Out: Use of Decision Trees in Epidemiology. Current Epidemiology Reports. 2018 Sep 1;5(3):221-9.

25. Breiman L, Friedman J, Olshen R, Stone C. Classification and regression trees. Wadsworth Int. Group. 1984;37(15):237-51.

26. Greene MZ, Hughes TL, Hanlon A, Huang L, Sommers MS, Meghani SH. Predicting cervical cancer screening among sexual minority women using

Classification and Regression Tree analysis. Preventive medicine reports. 2019 Mar 1;13:153-9.

27. Shaw LR, Chan F, McMahon BT. Intersectionality and disability harassment: The interactive effects of disability, race, age, and gender. Rehabilitation Counseling Bulletin. 2012 Jan;55(2):82-91.

28. Cairney J, Veldhuizen S, Vigod S, Streiner DL, Wade TJ, Kurdyak P. Exploring the social determinants of mental health service use using intersectionality theory and CART analysis. J Epidemiol Community Health. 2014 Feb 1;68(2):145-50.

29. Zufferey J. Investigating the migrant mortality advantage at the intersections of social stratification in Switzerland: The role of vulnerability. Demographic Research. 2016 Jan 1;34:899-926.

30. Dey A, Hay K, Afroz B, Chandurkar D, Singh K, Dehingia N, Raj A, Silverman JG. Understanding intersections of social determinants of maternal healthcare utilization in Uttar Pradesh, India. PloS one. 2018;13(10).

31. Sridharan S, Pereira A, Hay K, Dey A, Chandurkar D, Veldhuizen S, Nakaima A. Heterogeneities in utilization of antenatal care in Uttar Pradesh, India: the need to contextualize interventions to individual contexts. Global health action. 2018 Jan 1;11(1):1517929.

32. Villanti AC, Gaalema DE, Tidey JW, Kurti AN, Sigmon SC, Higgins ST. Co-occurring vulnerabilities and menthol use in US Young adult cigarette smokers: findings from wave 1 of the PATH Study, 2013–2014. Preventive medicine. 2018 Dec 1;117:43-51.

33. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics. 2006 Sep 1;15(3):651-74.

34. Salis KL, Kliem S, O'Leary KD. Conditional inference trees: a method for predicting intimate partner violence. Journal of marital and family therapy. 2014 Oct;40(4):430-41.

35. Breiman L. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA. 2002;1:58.

36. Nayak S, Hubbard A, Sidney S, Syme SL. A recursive partitioning approach to investigating correlates of self-rated health: The CARDIA Study. SSM-population health. 2018 Apr 1;4:178-88.

37. Banerjee M, Reynolds E, Andersson HB, Nallamothu BK. Tree-Based Analysis: A Practical Approach to Create Clinical Decision-Making Tools. Circulation: Cardiovascular Quality and Outcomes. 2019 May;12(5):e004879.

38. Li RH, Belford GG. Instability of decision tree classification algorithms. InProceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining 2002 Jul 23 (pp. 570-575).

39. Kanerva N, Kontto J, Erkkola M, Nevalainen J, Männistö S. Suitability of random forest analysis for epidemiological research: Exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design. Scandinavian journal of public health. 2018 Jul;46(5):557-64.

40. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. SSM-population health. 2018 Apr 1;4:95-9.

41. Kreatsoulas C, Subramanian SV. Machine learning in social epidemiology: learning from experience. SSM-population health. 2018 Apr;4:347.

42. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerging themes in epidemiology. 2017 Dec 1;14(1):11.

43. Evans CR, Williams DR, Onnela JP, Subramanian SV. A multilevel approach to modeling health inequalities at the intersection of multiple social identities. Social Science & Medicine. 2018 Apr 1;203:64-73.

44. Fisk SA, Mulinari S, Wemrell M, Leckie G, Vicente RP, Merlo J. Chronic obstructive pulmonary disease in Sweden: an intersectional multilevel analysis of individual heterogeneity and discriminatory accuracy. SSM-population health. 2018 Apr 1;4:334-46.

45. Hernandez-Yumar A, Wemrell M, Alesson IA, Lopez-Valcarcel BG, Leckie G, Merlo J. Socioeconomic differences in body mass index in Spain: An intersectional multilevel analysis of individual heterogeneity and discriminatory accuracy. PloS one. 2018;13(12).

46. Evans CR, Erickson N. Intersectionality and depression in adolescence and early adulthood: a MAIHDA analysis of the national longitudinal study of adolescent to adult health, 1995–2008. Social Science & Medicine. 2019 Jan 1;220:1-1.

47. Persmark A, Wemrell M, Zettermark S, Leckie G, Subramanian SV, Merlo J. Precision public health: Mapping socioeconomic disparities in opioid dispensations at Swedish pharmacies by Multilevel Analysis of Individual Heterogeneity and Discriminatory Accuracy (MAIHDA). PloS one. 2019;14(8).

48. Persmark A, Wemrell M, Evans CR, Subramanian SV, Leckie G, Merlo J. Intersectional inequalities and the US opioid crisis: challenging dominant narratives and revealing heterogeneities. Critical Public Health. 2019 Jun 19:1-7.

49. Kiadaliri A, Englund M. Intersectional inequalities and individual heterogeneity in chronic rheumatic diseases: An intersectional multilevel analysis. Arthritis care & research. 2019 Nov 15.

50. Wemrell M, Bennet L, Merlo J. Understanding the complexity of socioeconomic disparities in type 2 diabetes risk: a study of 4.3 million people in Sweden. BMJ Open Diabetes Research and Care. 2019 Nov 1;7(1).

51. Merlo J, Mulinari S. Measures of discriminatory accuracy and categorizations in public health: a response to Allan Krasnik's editorial. The European Journal of Public Health. 2015 Dec 1;25(6):910-.

52. Evans CR. Adding interactions to models of intersectional health inequalities: comparing multilevel and conventional methods. Social Science & Medicine. 2019 Jan 1;221:95-105.

53. Lizotte DJ, Mahendran M, Churchill SM, Bauer GR. Math versus meaning in MAIHDA: a commentary on multilevel statistical models for quantitative intersectionality. Social Science & Medicine. 2020 Jan 1;245:112500.

54. Merlo J. Multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) within an intersectional framework. Social Science & Medicine. 2018 Apr 1;203:74-80.

55. Wolpert DH, Macready WG. No free lunch theorems for optimization. IEEE transactions on evolutionary computation. 1997 Apr;1(1):67-82.

56. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2018.

57. Zou G. A modified poisson regression approach to prospective studies with binary data. American journal of epidemiology. 2004 Apr 1;159(7):702-6.

58. Davies HT, Crombie IK, Tavakoli M. When can odds ratios mislead?. Bmj. 1998 Mar 28;316(7136):989-91.

59. Zeileis A, Hothorn T. Diagnostic checking in regression relationships. R News 2: 7–10. Available at (accessed August 2011). http://CRAN. R-project. org/doc/Rnews/(http://CRAN. R-project. org/doc/Rnews/). 2002.

60. Zeileis A. Object-oriented Computation of Sandwich Estimators. Journal of Statistical Software. 2006 Aug 15;16(i09).

61. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer-Verlag. New York. 2002.

62. Croissant Y, Croissant MY. Package 'mlogit'. Technical report; 2019 Dec 9.

63. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. Journal of Statistical Software. 2015;67(1).

64. Elff M. mclogit: Mixed conditional logit models. R package version 0.4. 2016;4:718.

65. Christensen RHB. ordinal—Regression Models for Ordinal Data.R package version 2019.12-10. 2019.

66. Bürkner P. Advanced Bayesian Multilevel Modeling with the R Package brms. The R Journal. 10(1), 395–411. 2018.

67. Therneau T, Atkinson B, Ripley B. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. 2018.

68. Atkinson EJ, Therneau TM. An introduction to recursive partitioning using the RPART routines. Rochester: Mayo Foundation. 2000 Feb 11.

69. The FoRt Student Project Team. CHAID: CHi-squared Automated Interaction Detection R package version 0.1-2. 2015.

70. Kass GV. An exploratory technique for investigating large quantities of categorical data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1980 Jun;29(2):119-27.

71. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019 May;9(3):e1301.

72. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software. 2017 Mar 31;77(i01).

73. Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

74. Statistics Canada. Canadian Community Health Survey, 2015-2016: Annual component[Public use microdata file and codebook]. Ottawa, ON: Statistics

Canada. 2015.

75. Yelland LN, Salter AB, Ryan P. Relative risk estimation in randomized controlled trials: a comparison of methods for independent observations. The international journal of biostatistics. 2011 Jan 1;7(1):1-31.

76. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I. Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. BMC research notes. 2011;4:299-.

77. Wood S, Wood MS. Package 'mgcv'. R package version. 2015 Dec 12;1:29.

78. Austin PC. The large-sample performance of backwards variable elimination. Journal of Applied Statistics. 2008 Dec 1;35(12):1355-70.

79. Smith G. Step away from stepwise. Journal of Big Data. 2018 Dec 1;5(1):32.

80. Bell A, Holman D, Jones K. Using Shrinkage in Multilevel Models to Understand Intersectionality. Methodology. 2019 May 27.

81. Evans CR. Reintegrating contexts into quantitative intersectional analyses of health inequalities. Health & Place. 2019 Nov 1;60:102214.

82. Ivert AK, Gracia E, Lila M, Wemrell M, Merlo J. Does country-level gender equality explain individual risk of intimate partner violence against women? A multilevel analysis of individual heterogeneity and discriminatory accuracy (MAIHDA) in the European Union. European Journal of Public Health;. 2019.

83. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Team RC, Benesty M. Package 'caret'. The R Journal. 2020 Jan 7.

84. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC bioinformatics. 2007 Dec 1;8(1):25.

85. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010 May 15;26(10):1340-7.

86. Krieger N. Methods for the scientific study of discrimination and health: an ecosocial approach. American journal of public health. 2012 May;102(5):936-44.

87. Merlo J, Mulinari S, Wemrell M, Subramanian SV, Hedblad B. The tyranny of the averages and the indiscriminate use of risk factors in public health: The case

of coronary heart disease. SSM-population health. 2017 Dec 1;3:684-98.

88. Bauer GR, Scheim AI. Methods for analytic intercategorical intersectionality in quantitative research: discrimination as a mediator of health inequalities. Social Science & Medicine. 2019 Apr 1;226:236-45.

89. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018 Dec 18;169(12):866-72.

Appendices

Appendix A. Comparison of MAIHDA by Bayesian versus frequentist analysis

100 simulations were conducted for each of the three scenarios below. Sample sizes of 10,000 were used for each model. The Bayesian multilevel models were calculated using the R brms package [66]. Bayesian (B) multilevel models were performed each with 1000 burn ins, 2000 total. Frequentist (F) multilevel models were created with package lme4, using R version 3.5.3. Presented below are 0.025 and 0.975 percentiles of estimates from the 100 simulations. Results are compared against OLS regressions with and without the necessary interaction terms, and MAIHDA analyses (Bayesian and frequentist) with and without the necessary interaction terms.

Scenario 1: y = x1 + x2 + x3 + x4 + x5 + x1*x2

P(x1=1) = **50%;** P(x2=1) = **50%;** P(x3=1) = 50%; P(x4=1) = 50%; P(x5=1) = 50%;

Appendix A Table 1. 0.025 and 0.975 percentiles of Scenario 1 from 100 simulations

	OLS	OLS with interaction	MAIHDA (B)	MAIHDA (F)	MAIHDA (B) with interaction	MAIHDA (F) with interaction
Intercept	(-0.298 , -0.197)	(-0.056 , 0.052)	(-0.299 , -0.201)	(-0.297, -0.200)	(-0.056 , 0.052)	(-0.056, 0.052)
x1	(1.466 , 1.538)	(0.936 , 1.056)	(1.459 , 1.534)	(1.463, 1.532)	(0.936 , 1.056)	(0.935, 1.056)
x2	(1.454 , 1.540)	(0.948 , 1.062)	(1.458 , 1.539)	(1.456, 1.538)	(0.948 , 1.062)	(0.948, 1.062)
x3	(0.970 , 1.037)	(0.970 , 1.040)	(0.969 , 1.039)	(0.969, 1.041)	(0.970 , 1.041)	(0.970, 1.041)

x4	(0.966 , 1.038)	(0.968 , 1.036)	(0.969 , 1.036)	(0.969, 1.036)	(0.968 , 1.036)	(0.968, 1.036)
x5	(0.959 , 1.042)	(0.961 , 1.038)	(0.960 , 1.039)	(0.961, 1.039)	(0.961 , 1.039)	(0.961, 1.038)
x1:x2	-	(0.924 , 1.071)	-	-	(0.923 , 1.072)	(0.924, 1.071)

Scenario 2: y = x1 + x2 + x3 + x4 + x5 + x1*x2

P(x1=1) = **70%;** P(x2=1) = **70%;** P(x3=1) = 50%; P(x4=1) = 50%; P(x5=1) = 50%;

Appendix A Table 2. 0.025 and 0.975 percentiles of Scenario 2 from 100 simulations

	OLS	OLS with interaction	MAIHDA (B)	MAIHDA (F)	MAIHDA (B) with interaction	MAIHDA (F) with interaction
Intercept	(-0.552 , -0.441)	(-0.078 , 0.064)	(-0.322 , -0.208)	(-0.326, -0.211)	(-0.078 , 0.063)	(-0.077, 0.064)
x1	(1.663 , 1.736)	(0.925 , 1.082)	(1.465 , 1.555)	(1.461, 1.554)	(0.925 , 1.082)	(0.925, 1.081)
x2	(1.648 , 1.749)	(0.927 , 1.084)	(1.461 , 1.558)	(1.461, 1.559)	(0.927 , 1.083)	(0.927, 1.084)
x3	(0.968 , 1.041)	(0.97 , 1.04)	(0.963 , 1.039)	(0.963, 1.042)	(0.969 , 1.039)	(0.969, 1.039)
x4	(0.967 , 1.037)	(0.968 , 1.036)	(0.959 , 1.044)	(0.961, 1.042)	(0.969 , 1.037)	(0.967, 1.036)
x5	(0.962 , 1.045)	(0.961 , 1.039)	(0.96 , 1.044)	(0.957, 1.045)	(0.959 , 1.039)	(0.958, 1.040)
x1:x2	-	(0.889 , 1.097)	-	-	(0.889 , 1.097)	(0.889, 1.097)

Scenario 3: y = x1 + x2 + x3 + x4 + x5 – 2(x1*x2)

P(x1=1) = **20%;** P(x2=1) = **20%;** P(x3=1) = 50%; P(x4=1) = 50%; P(x5=1) = 50%;

Appendix A Table 3. 0.025 and 0.975 percentiles of Scenario 3 from 100 simulations

	OLS	OLS with interaction	MAIHDA (B)	MAIHDA (F)	MAIHDA (B) with interaction	MAIHDA (F) with interaction
Intercept	(0.040 , 0.118)	(-0.039 , 0.033)	(0.432 , 0.537)	(0.432, 0.535)	(-0.039 , 0.033)	(-0.038, 0.033)
x1	(0.546 , 0.653)	(0.949 , 1.053)	(-0.048 , 0.077)	(-0.045, 0.078)	(0.949 , 1.054)	(0.949, 1.053)
x2	(0.544 , 0.657)	(0.949 , 1.056)	(-0.051 , 0.081)	(-0.051, 0.085)	(0.950 , 1.056)	(0.949, 1.056)
x3	(0.964 , 1.040)	(0.970 , 1.040)	(0.947 , 1.060)	(0.946, 1.060)	(0.967 , 1.039)	(0.966, 1.040)
x4	(0.968 , 1.038)	(0.968 , 1.036)	(0.935 , 1.057)	(0.936, 1.057)	(0.968 , 1.037)	(0.968, 1.037)
x5	(0.957 , 1.036)	(0.961 , 1.038)	(0.941 , 1.062)	(0.943, 1.058)	(0.958 , 1.039)	(0.959, 1.038)
x1:x2	-	(-2.123 , -1.889)	-	-	(-2.122 , -1.888)	(-2.123, -1.889)

Appendix B. Over-specified and best-fitted regression results

Appendix B Table 1. Model 1 **(continuous outcome, categorical inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	5.0	4.5	3.6	4.9	5.0	5.1	5.5	5.8
x1.1	81.7	87.1	98.0	99.8	96.5	97.7	100.0	100.0
x1.2	80.6	85.8	98.6	99.6	96.4	98.1	100.0	100.0
x1.3	76.2	87.2	98.5	99.9	96.1	98.2	100.0	100.0
x2	61.2	76.6	94.7	98.5	93.3	97.0	99.8	100.0
x3	70.0	84.9	95.6	99.7	96.3	98.2	100.0	100.0
x4	76.0	86.0	97.6	99.8	97.5	98.6	100.0	100.0
x5	66.6	81.5	96.4	99.2	95.6	97.6	100.0	100.0
x6.1	3.9	5.8	4.5	5.2	4.5	6.2	5.0	4.5
x6.2	5.1	4.1	3.9	4.3	5.7	6.2	4.5	4.7
x1.1:x2	4.4	4.6	4.7	5.0	6.7	4.2	5.3	5.4
x1.1:x2	42.1	62.5	92.6	97.6	87.7	93.6	99.4	100.0
x1.3:x2	42.7	63.8	92.7	98.0	87.2	94.2	99.3	100.0
x3:x4	5.8	5.7	5.3	5.4	5.4	5.9	4.3	4.0
x3:x5	5.9	6.3	4.9	5.0	5.9	3.9	4.6	5.6
x4:x5	4.8	6.1	4.7	5.5	5.0	5.1	4.4	5.1
x3:x4:x5	22.7	48.1	86.9	94.4	82.0	90.2	98.5	99.8

Appendix B Table 2. Model 1 **(continuous outcome, categorical inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	95.0	95.5	96.4	95.1	95.0	94.9	94.5	94.2
x1.1	94.3	93.7	97.0	96.0	95.3	94.6	95.1	95.2
x1.2	95.2	94.4	94.0	96.2	95.1	96.5	95.1	95.6
x1.3	95.9	94.8	95.6	95.8	94.4	94.5	94.8	95.2
x2	94.8	95.2	95.4	95.5	94.8	94.3	94.9	96.2
x3	94.0	94.8	95.6	93.9	94.2	94.9	95.4	95.2
x4	95.6	94.3	95.4	95.2	94.8	94.2	95.9	96.6
x5	94.9	94.3	94.9	94.6	95.7	94.9	95.1	94.5
x1.1:x2	94.9	95.4	95.6	95.5	95.5	94.8	94.1	94.4
x1.3:x2	94.8	95.1	95.2	95.1	93.9	95.4	95.5	95.3
x3:x4:x5	94.5	93.4	95.1	94.8	95.2	94.5	95.4	95.1

Appendix B Table 3. Model 2 **(continuous outcome, mixed inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	4.0	4.3	4.3	4.0	3.3	4.9	4.8	4.7
x1	97.5	99.0	100.0	100.0	99.5	100.0	100.0	100.0
x2	80.0	87.4	96.8	97.7	88.5	94.6	98.8	99.2
x3	93.9	96.4	100.0	100.0	94.2	96.7	100.0	100.0
x4	96.2	98.7	100.0	100.0	96.3	99.0	100.0	100.0
x5	92.2	96.6	99.9	100.0	93.2	97.2	99.8	100.0
x6	4.7	3.5	3.8	5.6	4.4	4.5	4.9	4.9
x1:x2	69.5	84.4	97.2	99.7	90.3	95.2	99.5	100.0
x3:x4	5.0	4.3	4.4	4.3	5.4	5.0	4.6	3.7
x3:x5	4.3	4.2	4.3	4.8	4.6	4.5	5.1	5.3
x4:x5	5.2	4.2	4.0	4.6	4.8	4.4	4.3	4.5
x3:x4:x5	79.9	89.3	98.2	99.7	83.7	91.8	98.4	99.8

Appendix B Table 4. Model 2 **(continuous outcome, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	96.0	95.7	95.7	96.0	96.7	95.1	95.2	95.3
x1	95.7	96.3	97.1	95.9	95.8	95.9	96.0	95.2
x2	53.8	28.7	6.1	1.4	21.1	8.9	1.2	0.0
x3	96.0	95.7	95.5	95.7	96.0	94.3	94.6	96.4
x4	96.0	96.4	94.5	94.9	95.8	95.5	95.1	94.9
x5	95.0	95.3	96.2	95.7	94.9	95.3	96.2	95.3
x1:x2	17.3	9.1	1.1	0.1	6.2	2.8	0.0	0.0
x3:x4:x5	95.0	95.0	95.8	95.8	96.1	94.4	95.2	95.1

Appendix B Table 5. Model 3 **(common binomial outcome, categorical inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	97.2	99.6	100.0	100.0	100.0	100.0	100.0	100.0
x1.1	12.0	29.2	90.4	98.8	59.2	80.6	99.4	100.0
x1.2	7.5	30.1	88.8	98.0	60.2	81.3	99.6	100.0
x1.3	18.3	25.8	88.4	98.4	57.6	80.5	99.7	100.0
x2	17.2	22.3	73.4	93.7	34.8	61.9	97.1	99.7
x3	15.2	19.3	83.0	98.1	84.1	95.3	100.0	100.0
x4	10.7	26.5	86.8	98.7	86.8	95.0	100.0	100.0
x5	30.0	8.8	77.8	95.7	76.2	91.2	100.0	100.0
x6.1	2.3	5.4	5.1	5.1	5.7	4.5	5.2	6.0
x6.2	5.2	5.0	3.9	5.9	5.1	5.6	5.6	5.1
x1.1:x2	10.1	4.5	4.2	4.1	4.3	4.6	5.3	4.0
x1.1:x2	12.8	22.0	78.0	93.7	41.6	67.1	97.6	99.9
x1.3:x2	18.6	17.9	78.2	95.1	40.2	65.9	97.3	100.0
x3:x4	36.0	6.8	5.7	5.7	5.2	5.4	4.4	4.4
x3:x5	45.2	39.9	5.4	5.3	5.1	4.7	5.2	4.4
x4:x5	50.5	35.3	4.5	5.6	4.0	5.4	5.7	6.6
x3:x4:x5	65.8	56.7	9.6	24.5	9.3	14.7	68.4	92.8

Appendix B Table 6. Model 3 **(common binomial outcome, categorical inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	93.8	95.1	94.4	94.9	94.7	96.0	94.4	94.8
x1.1	94.9	95.5	94.8	95.5	96.1	95.7	94.4	96.0
x1.2	96.5	94.6	95.6	95.1	95.3	94.1	94.0	95.4
x1.3	95.4	95.7	96.3	94.8	94.2	94.6	93.9	95.6
x2	92.3	95.7	95.0	95.4	95.1	94.9	94.3	95.8
x3	81.1	94.1	95.0	96.0	95.7	94.5	95.8	94.4
x4	84.9	96.2	94.6	93.7	94.7	96.2	94.8	94.0
x5	64.0	92.2	95.2	94.2	94.6	95.5	95.1	95.2
x1.1:x2	88.5	94.5	94.1	95.3	96.0	94.9	94.9	95.5
x1.3:x2	82.6	94.1	95.9	94.5	94.8	94.7	95.1	95.2
x3:x4:x5	34.2	42.9	94.4	94.2	93.6	94.1	96.0	95.2

Appendix B Table 7. Model 4 **(common binomial outcome, mixed inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
x1	81.2	94.3	100.0	100.0	93.8	98.7	100.0	100.0
x2	30.0	51.1	84.9	93.0	47.8	67.7	89.7	96.0
x3	67.0	87.1	99.9	100.0	76.6	93.1	100.0	100.0
x4	74.1	88.9	100.0	100.0	81.4	91.7	100.0	100.0
x5	61.6	83.6	99.4	100.0	70.6	87.1	99.6	100.0
x6	6.0	6.3	5.0	4.0	4.8	4.5	4.8	4.0
x1:x2	28.9	52.0	95.1	99.8	47.8	73.8	98.3	99.9
x3:x4	5.7	4.5	5.3	5.0	4.2	4.4	5.2	4.4
x3:x5	15.3	8.3	5.6	3.7	6.8	5.3	5.8	4.9
x4:x5	12.1	5.7	4.4	6.1	5.7	4.0	4.6	5.6
x3:x4:x5	24.4	13.9	50.0	84.8	11.3	14.1	57.7	88.9

Appendix B Table 8. Model 4 **(common binomial outcome, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	95.9	94.4	94.7	95.1	95.2	93.8	93.8	95.7
x1	95.7	95.0	94.3	95.3	95.6	95.9	93.1	94.5
x2	90.2	83.3	26.9	7.6	84.6	69.3	14.5	2.6
x3	94.5	95.9	93.6	95.1	93.1	95.4	95.1	94.8
x4	95.4	95.4	94.4	95.3	95.3	94.8	94.4	95.0
x5	93.9	94.8	94.7	94.3	95.2	95.7	95.1	93.5
x1:x2	47.3	28.2	3.9	0.1	31.6	15.9	1.0	0.0
x3:x4:x5	78.2	92.6	94.6	94.2	91.1	94.2	93.5	94.4

Appendix B Table 9. Model 5 **(rare binomial outcome, categorical inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	96.4	99.6	100.0	100.0	100.0	100.0	100.0	100.0
x1.1	13.4	4.7	53.9	93.3	16.3	37.8	97.5	100.0
x1.2	11.4	4.7	51.5	91.5	14.5	35.6	97.1	100.0
x1.3	12.3	4.3	51.2	92.6	14.0	37.0	97.1	100.0
x2	31.5	23.4	29.7	69.8	14.2	20.8	84.5	99.1
x3	36.0	25.1	52.2	90.9	52.2	78.2	99.9	100.0
x4	32.1	18.3	60.9	94.3	60.4	85.5	100.0	100.0
x5	54.5	41.3	41.2	85.4	33.0	72.3	99.8	100.0
x6.1	17.8	2.1	3.9	5.6	4.7	3.6	5.4	4.5
x6.2	19.1	3.1	3.0	5.8	4.1	5.1	4.9	4.8
x1.1:x2	38.5	26.4	5.3	3.7	6.4	4.1	5.3	5.0
x1.1:x2	48.7	40.8	41.7	79.2	17.8	28.4	88.4	99.8
x1.3:x2	42.7	37.0	39.4	81.9	18.3	28.3	89.0	99.1
x3:x4	42.3	45.9	3.7	5.1	5.2	5.4	4.4	3.6
x3:x5	41.8	50.5	29.5	6.4	36.1	13.1	5.1	5.5
x4:x5	44.8	52.5	20.6	4.3	27.0	10.0	5.1	4.8
x3:x4:x5	40.2	62.4	40.4	10.3	48.7	21.0	26.6	64.8

Appendix B Table 10. Model 5 **(rare binomial outcome, categorical inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	86.7	96.5	95.3	95.4	96.0	95.8	94.8	94.8
x1.1	87.3	96.9	95.4	96.2	96.1	95.2	94.8	96.1
x1.2	88.8	97.3	95.7	95.5	95.7	94.6	94.7	94.8
x1.3	87.9	98.0	95.0	95.9	96.1	95.7	94.2	94.2
x2	69.5	82.6	95.8	95.3	94.5	95.5	95.1	94.2
x3	63.7	73.2	97.1	95.3	96.0	94.6	94.4	96.1
x4	66.9	80.0	95.5	94.0	95.1	95.8	94.4	94.8
x5	43.3	55.6	95.3	96.0	94.2	95.5	94.0	94.5
x1.1:x2	50.6	59.6	96.4	96.5	85.8	94.4	96.1	94.8
x1.3:x2	56.9	63.6	94.6	95.2	85.7	96.0	94.9	94.0
x3:x4:x5	59.8	37.6	58.9	92.8	50.7	79.7	95.4	94.6

Appendix B Table 11. Model 6 **(rare binomial outcome, mixed inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
x1	55.1	79.6	100.0	100.0	75.3	93.7	100.0	100.0
x2	10.5	15.5	71.0	91.1	16.9	31.5	86.4	96.7
x3	32.5	64.5	99.1	100.0	40.4	75.1	99.7	100.0
x4	37.5	69.4	99.7	100.0	48.1	79.9	100.0	100.0
x5	32.0	50.0	98.6	100.0	25.9	62.1	99.5	100.0
x6	8.5	6.4	5.3	5.2	5.0	5.1	4.5	4.5
x1:x2	21.2	24.6	86.5	99.7	24.5	39.5	96.8	100.0
x3:x4	17.1	6.9	5.1	5.6	6.4	5.5	5.0	4.2
x3:x5	59.9	31.4	5.8	4.6	44.5	16.8	5.4	5.2
x4:x5	51.2	23.8	5.1	5.7	32.9	11.5	4.5	5.6
x3:x4:x5	68.0	44.4	20.0	47.4	50.7	25.3	23.5	61.2

Appendix B Table 12. Model 6 **(rare binomial outcome, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	95.2	95.1	95.2	95.1	94.5	95.3	95.8	95.9
x1	91.8	95.4	95.2	95.5	92.9	94.3	95.5	95.1
x2	92.8	91.7	68.1	25.1	94.0	89.1	49.5	11.0
x3	90.2	93.6	95.5	94.7	95.6	95.7	94.7	94.9
x4	90.4	93.1	95.4	94.5	94.9	94.2	94.3	96.1
x5	81.6	93.0	95.1	94.5	92.7	95.3	96.2	94.9
x1:x2	65.3	49.2	9.5	0.6	57.4	36.0	2.3	0.0
x3:x4:x5	32.4	57.2	93.9	94.2	49.1	75.6	95.3	93.9

Appendix B Table 13. Model 7 **(multinomial outcome y=2, categorical inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	63.8	97.9	100.0	100.0	100.0	100.0	100.0	100.0
x1.1	30.0	7.3	47.7	90.5	49.0	78.7	99.9	100.0
x1.2	32.3	5.8	46.2	90.1	47.7	77.4	100.0	100.0
x1.3	30.6	6.7	46.9	90.1	46.8	78.8	100.0	100.0
x2	74.0	33.6	15.3	60.0	20.4	44.7	99.5	100.0
x3	29.5	23.6	79.0	98.1	63.8	83.5	100.0	100.0
x4	24.0	23.9	82.4	98.8	69.1	87.9	100.0	100.0
x5	39.3	23.6	73.5	95.7	56.4	78.3	99.8	100.0
x6.1	21.7	3.3	5.4	3.8	5.8	4.2	4.9	5.2
x6.2	22.3	3.2	4.1	5.5	6.3	4.5	4.8	5.3
x1.1:x2	88.9	60.1	4.2	4.1	4.7	4.3	5.9	5.0
x1.1:x2	85.5	48.6	36.9	72.6	36.6	59.3	99.1	100.0
x1.3:x2	85.3	48.3	36.1	73.3	35.8	63.8	98.3	100.0
x3:x4	29.9	6.3	4.3	4.6	4.9	4.6	5.3	4.8
x3:x5	66.6	16.8	4.8	4.0	4.7	4.9	4.6	5.6
x4:x5	58.2	10.6	4.8	4.1	5.2	4.6	5.5	5.3
x3:x4:x5	76.3	21.2	13.6	36.2	9.6	17.3	79.2	98.9

Appendix B Table 14. Model 7 **(multinomial outcome y=2, categorical inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	84.3	96.8	95.7	95.3	94.2	95.5	95.0	93.4
x1.1	70.5	94.6	94.9	96.5	95.3	94.1	95.9	93.7
x1.2	67.5	95.6	95.5	95.1	95.6	95.3	94.9	94.1
x1.3	69.3	95.1	94.3	95.8	93.8	94.3	95.1	95.3
x2	25.7	64.1	96.3	96.5	96.0	95.4	95.9	95.1
x3	75.6	94.8	95.4	95.9	94.2	94.5	94.3	93.4
x4	80.9	95.4	95.6	95.5	96.0	93.6	95.2	95.2
x5	63.8	93.7	95.2	95.6	95.3	95.1	94.1	94.6
x1.1:x2	14.2	51.7	95.6	96.0	94.6	94.9	95.3	95.6
x1.3:x2	14.9	51.1	95.8	95.8	95.3	94.7	94.8	95.6
x3:x4:x5	23.8	78.7	95.8	95.4	95.3	95.3	95.1	94.6

Appendix B Table 15. Model 7 **(multinomial outcome y=3, categorical inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	93.4	99.9	100.0	100.0	100.0	100.0	100.0	100.0
x1.1	29.3	6.5	46.4	90.3	51.5	81.5	100.0	100.0
x1.2	32.1	5.9	49.4	90.0	48.2	79.6	100.0	100.0
x1.3	30.8	8.2	46.5	91.2	49.7	82.0	100.0	100.0
x2	73.2	33.9	16.2	63.5	21.3	47.3	98.8	100.0
x3	29.7	25.3	78.2	97.3	62.8	83.1	100.0	100.0
x4	23.9	25.3	83.3	99.0	70.4	85.5	100.0	100.0
x5	37.7	24.7	75.1	97.0	58.8	78.3	99.9	100.0
x6.1	21.6	3.7	5.6	3.9	5.3	5.3	5.4	6.3
x6.2	22.0	3.4	4.5	5.7	6.9	6.1	5.0	6.4
x1.1:x2	87.9	60.3	4.9	4.8	4.5	4.4	5.5	3.8
x1.1:x2	86.1	48.7	35.8	73.4	37.8	65.2	98.6	100.0
x1.3:x2	84.4	48.5	37.0	71.9	41.8	68.3	98.9	100.0
x3:x4	27.6	5.7	4.7	6.0	6.5	5.4	4.8	5.9
x3:x5	60.7	11.5	3.9	4.0	5.3	4.7	4.4	3.9
x4:x5	51.5	9.6	4.7	3.5	5.1	5.2	3.7	5.6
x3:x4:x5	70.6	15.5	14.4	42.8	11.1	19.5	83.8	99.6

Appendix B Table 16. Model 7 **(multinomial outcome y=3, categorical inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	84.6	97.5	94.4	95.7	93.5	95.4	94.7	93.6
x1.1	70.7	95.6	94.8	95.0	94.6	92.9	95.7	95.4
x1.2	68.1	96.5	94.9	94.8	94.9	94.9	95.0	94.4
x1.3	69.6	95.5	94.7	95.0	94.5	95.5	97.3	95.0
x2	26.3	64.2	96.5	96.9	95.9	94.5	95.4	96.0
x3	77.0	95.9	94.6	95.7	94.2	94.7	95.1	93.1
x4	81.2	96.7	95.5	95.2	92.9	94.5	95.6	95.3
x5	66.6	94.0	94.9	95.6	95.0	95.3	95.6	94.2
x1.1:x2	13.5	51.9	96.0	96.1	94.8	94.8	95.0	96.1
x1.3:x2	16.0	51.8	95.7	96.1	95.2	95.2	96.1	95.7
x3:x4:x5	29.6	84.8	96.1	95.0	94.6	95.0	95.0	95.5

Appendix B Table 17. Model 8 **(multinomial outcome y=2, mixed inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
x1	48.5	81.4	100.0	100.0	98.7	99.9	100.0	100.0
x2	3.3	13.5	49.7	77.2	26.6	47.0	86.7	92.8
x3	58.3	84.3	100.0	100.0	68.1	88.6	100.0	100.0
x4	63.5	85.4	100.0	100.0	71.8	88.9	100.0	100.0
x5	48.7	74.7	99.6	100.0	57.9	82.4	100.0	100.0
x6	4.9	4.3	5.1	5.2	5.1	4.7	4.8	4.9
x1:x2	11.7	20.7	67.3	89.0	38.5	66.2	96.8	99.8
x3:x4	5.5	4.8	5.1	6.0	4.9	5.2	4.9	6.0
x3:x5	5.8	5.3	4.7	6.1	4.8	5.3	4.2	5.6
x4:x5	5.6	4.2	4.3	4.8	5.0	5.3	5.3	6.0
x3:x4:x5	10.3	15.7	74.6	98.5	10.4	17.2	80.2	99.2

Appendix B Table 18. Model 8 **(multinomial outcome y=2, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	95.4	94.9	94.1	94.0	95.1	95.9	95.1	93.3
x1	95.5	95.9	95.3	95.0	95.6	93.9	94.7	94.8
x2	94.2	89.2	48.7	14.3	75.7	50.7	4.8	0.0
x3	94.7	94.8	93.8	94.3	95.2	95.1	94.9	94.0
x4	94.6	94.8	95.0	94.3	96.7	95.4	95.1	94.9
x5	95.2	94.6	94.4	95.1	95.4	94.2	94.8	93.9
x1:x2	84.0	63.7	9.9	0.0	37.1	13.6	0.0	0.0
x3:x4:x5	93.9	95.3	93.4	94.1	94.3	94.8	94.4	93.4

Appendix B Table 19. Model 8 **(multinomial outcome y=3, mixed inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
x1	50.6	81.3	100.0	100.0	98.2	100.0	100.0	100.0
x2	2.9	13.2	53.1	75.6	31.2	51.2	85.7	91.9
x3	58.7	83.5	100.0	100.0	69.6	89.1	100.0	100.0
x4	67.8	86.7	100.0	100.0	74.0	90.2	100.0	100.0
x5	52.0	75.8	100.0	100.0	65.1	83.3	100.0	100.0
x6	4.2	4.6	5.4	5.5	4.2	4.2	5.8	6.5
x1:x2	11.9	23.3	68.4	89.4	45.7	72.3	97.6	100.0
x3:x4	4.7	5.1	4.7	6.8	5.5	5.5	5.2	5.9
x3:x5	4.9	6.1	4.6	5.0	5.2	5.5	4.8	4.2
x4:x5	5.0	5.8	5.1	4.2	4.6	6.1	4.3	5.1
x3:x4:x5	10.4	16.8	79.3	99.5	10.5	21.9	86.3	99.9

Appendix B Table 20. Model 8 **(multinomial outcome y=3, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	95.2	94.9	93.6	95.5	95.7	95.0	94.9	93.8
x1	95.4	94.6	95.5	95.3	94.5	94.9	93.9	94.5
x2	94.0	89.1	46.6	15.3	72.6	48.7	4.2	0.0
x3	94.3	95.3	94.2	93.9	94.5	94.7	94.7	94.1
x4	94.2	94.9	94.3	95.2	94.8	94.3	95.0	94.4
x5	94.9	95.8	95.4	94.3	95.7	95.4	95.3	94.7
x1:x2	86.0	64.3	9.5	0.2	31.4	10.1	0.0	0.0
x3:x4:x5	93.9	94.1	95.2	96.0	94.2	94.4	95.7	95.1

Appendix B Table 21. Model 9 **(negative binomial outcome, categorical inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	5.7	6.1	5.3	4.9	4.5	5.1	4.1	5.2
x1.1	25.6	40.5	92.8	99.4	78.6	90.0	100.0	100.0
x1.2	24.7	41.3	91.5	99.6	78.6	90.3	99.9	100.0
x1.3	18.9	41.1	91.6	99.6	75.3	89.8	100.0	100.0
x2	15.4	21.4	77.6	95.6	51.3	76.1	99.0	100.0
x3	20.3	51.0	91.0	98.8	88.4	96.4	100.0	100.0
x4	28.2	55.2	92.3	99.1	90.8	95.3	100.0	100.0
x5	12.3	39.0	88.9	97.6	84.0	92.9	100.0	100.0
x6.1	4.4	6.3	5.1	4.8	4.1	5.2	6.1	5.4
x6.2	7.0	6.0	4.2	5.4	4.6	4.4	5.6	4.8
x1.1:x2	10.1	6.4	5.0	6.2	6.2	4.2	3.3	6.4
x1.1:x2	18.5	30.0	80.5	94.6	56.7	77.8	98.1	100.0
x1.3:x2	11.5	29.4	79.2	93.2	58.4	76.2	97.7	100.0
x3:x4	5.3	3.4	4.7	5.8	4.5	4.0	5.5	5.1
x3:x5	2.2	3.4	4.8	4.4	5.4	6.1	4.9	4.8
x4:x5	2.2	3.3	5.2	5.1	4.4	4.7	5.0	4.7
x3:x4:x5	1.8	4.2	19.2	49.1	15.3	29.8	86.8	98.6

Appendix B Table 22. Model 9 **(negative binomial outcome, categorical inputs)** regression coefficient confidence interval coverage (% of iterations) [a]

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	-	100.0	94.7	95.1	95.6	94.8	95.7	94.8
x1.1	-	100.0	95.3	95.6	95.6	96.3	96.0	95.6
x1.2	-	100.0	95.0	95.7	95.2	96.2	96.3	96.6
x1.3	-	100.0	95.5	96.0	94.3	94.9	96.1	95.6
x2	-	100.0	96.0	95.9	94.2	95.7	96.0	96.1
x3	-	100.0	94.6	93.8	95.9	94.2	95.8	95.5
x4	-	100.0	94.1	95.8	95.9	96.0	95.8	94.5
x5	-	100.0	94.7	94.8	95.5	94.9	95.4	96.0
x1.1:x2	-	100.0	94.1	95.6	94.8	96.2	94.5	94.1
x1.3:x2	-	100.0	95.2	95.4	94.5	95.0	94.1	94.4
x3:x4:x5	-	100.0	94.3	93.3	94.3	94.7	96.0	95.9

[a] at N=2000, confidence interval formation failed

Appendix B Table 23. Model 10 **(negative binomial outcome, mixed inputs)** regression coefficient significance (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	5.6	4.8	5.0	5.3	4.5	5.5	5.2	4.7
x1	90.4	97.4	100.0	100.0	98.1	99.7	100.0	100.0
x2	35.6	58.5	87.8	95.1	61.4	77.8	92.9	97.0
x3	81.8	91.8	100.0	100.0	85.7	94.0	100.0	100.0
x4	84.6	93.5	100.0	100.0	88.2	95.1	100.0	100.0
x5	75.5	90.9	99.8	100.0	79.2	92.9	99.9	100.0
x6	5.7	5.5	4.6	4.5	5.7	4.1	4.9	4.9
x1:x2	31.7	57.0	95.0	99.6	60.5	82.7	99.1	100.0
x3:x4	5.7	6.6	4.3	5.0	4.7	6.3	4.2	5.2
x3:x5	4.9	6.0	5.8	4.5	5.9	5.6	4.6	5.6
x4:x5	5.7	5.3	4.7	4.6	4.4	5.4	3.9	5.5
x3:x4:x5	15.9	23.6	79.6	95.5	16.7	27.4	85.5	97.3

Appendix B Table 24. Model 10 **(negative binomial outcome, mixed inputs)** regression coefficient confidence interval coverage (% of iterations)

	Over-specified				**Best-fitted**			
	N = 2000	N = 5000	N = 50000	N = 200000	N = 2000	N = 5000	N = 50000	N = 200000
Intercept	94.4	95.2	95.0	94.7	95.5	94.5	94.7	95.3
x1	94.8	95.2	95.3	95.7	95.5	95.7	96.2	94.6
x2	87.1	74.0	22.3	6.0	75.0	53.4	7.6	1.1
x3	93.6	94.6	93.6	95.4	95.2	94.7	94.2	95.2
x4	94.4	95.9	94.5	95.2	95.3	95.6	95.0	93.8
x5	93.7	94.8	95.1	95.3	94.0	95.5	95.6	95.8
x1:x2	46.4	24.5	2.1	0.0	26.6	9.9	0.5	0.0
x3:x4:x5	92.0	94.0	94.3	94.3	94.2	94.2	94.9	94.3

Appendix C. MAIHDA results for models with mixed inputs

Appendix C Table 1. Model 2 **(Continuous outcome, mixed inputs)** MAIHDA coefficient significance

	Intercept	x1	x2	x3	x4	x5	x6
Expected	0	100	100	100	100	100	0
N = 2000	71.1	95.9	85.8	89.5	89.1	86.3	2.1
N = 5000	80.2	96	89.4	89.4	90.4	89.6	1.8
N = 50000	85.4	98.5	92.1	90.5	89.4	90	4.4
N = 200000	88.2	98.9	91.8	89.6	90	90.3	5.5

Appendix C Table 2. Model 2 **(Continuous outcome, mixed inputs)** MAIHDA confidence interval coverage by definition 1 (typical additive effects)

	Intercept	x1	x2	x3	x4	x5
N = 2000	28.9	26.5	30	29.9	36.7	14.7
N = 5000	19.8	13	21.3	18.4	22.6	11.6
N = 50000	14.6	3.6	11.7	12.8	14	11.3
N = 200000	11.8	1.2	12.5	9.9	10	9.2

Appendix C Table 3. Model 4 **(Common binary outcome, mixed inputs)** MAIHDA coefficient significance

	Intercept	x1	x2	x3	x4	x5	x6
Expected	100	100	100	100	100	100	0
N = 2000	100.0	88.5	36.5	82.7	84.7	71.9	2.5
N = 5000	100.0	95.1	57.1	93.8	92.5	86.4	2.6
N = 50000	100.0	99.1	85.1	99.1	99.2	96.5	1.3
N = 200000	100.0	97.8	89.6	98.1	99.1	97.0	1.5

Appendix C Table 4. Model 4 **(Common binary outcome, mixed inputs)** MAIHDA confidence interval coverage by definition 1 (typical additive effects)

	Intercept	x1	x2	x3	x4	x5
N = 2000	96.9	90.2	79.7	94.8	96.0	93.0
N = 5000	97.0	70.8	61.4	95.9	94.7	89.4
N = 50000	88.6	11.6	10.1	78.8	80.9	49.7
N = 200000	53.8	0.6	2.4	43.9	48.8	24.1

Appendix C Table 5. Model 6 **(Rare binary outcome, mixed inputs)** MAIHDA coefficient significance

	Intercept	x1	x2	x3	x4	x5	x6
Expected	100	100	100	100	100	100	0
N = 2000	100.0	69.1	11.1	61.7	61.7	42.5	3.6
N = 5000	100.0	90.0	22.6	86.0	87.3	72.5	4.4
N = 50000	100.0	100.0	70.0	100.0	100.0	99.7	3.6
N = 200000	100.0	100.0	86.7	100.0	100.0	100.0	1.9

Appendix C Table 6. Model 6 **(Rare binary outcome, mixed inputs)** MAIHDA confidence interval coverage by definition 1 (typical additive effects)

	Intercept	x1	x2	x3	x4	x5
N = 2000	94.3	93.9	92.1	95.7	94.7	95.1
N = 5000	96.9	89.3	83.5	95.7	95.6	93.0
N = 50000	96.8	31.4	20.6	91.8	94.1	84.6
N = 200000	93.3	1.6	1.8	80.1	85.1	53.2

Appendix C Table 7. Model 10 **(Negative binomial outcome, mixed inputs)** MAIHDA coefficient significance

	Intercept	x1	x2	x3	x4	x5	x6
Expected	0	100	100	100	100	100	0
N = 2000	6.4	95.6	56.4	88.9	90.6	80.6	4.8
N = 5000	9.7	97.5	73.3	94.9	95.6	90.1	3.9
N = 50000	47.1	98.7	87.1	96.5	97.9	94.4	2.0
N = 200000	72.7	98.8	89.6	96.2	97.2	95.5	2.0

Appendix C Table 8. Model 10 **(Negative binomial outcome, mixed inputs)** MAIHDA confidence interval coverage by definition 1 (typical additive effects)

	Intercept	x1	x2	x3	x4	x5
N = 2000	93.6	78.7	69.3	92.6	94.3	83.9
N = 5000	90.3	55.6	45.1	91.1	90.8	71.1
N = 50000	52.9	5.1	6.0	45.4	52.0	24.3
N = 200000	27.3	1.0	3.7	19.9	21.4	13.9

Curriculum Vitae

Mayuri Mahendran

Education:

MSc. Student, Western University — Expected completion: April 2020
- Epidemiology and Biostatistics, Supervisor: Dr. Greta Bauer
- Thesis topic: Simulation study comparing statistical methods for quantitative intersectionality research, focussing on population health assessment

Bachelor of Medical Sciences, Western University — 2018
Honours Specialization in Epidemiology and Biostatistics, Minor in Microbiology and Immunology
- Fourth year thesis: Evaluating Canadian Community Health Survey data to assess socioeconomic inequities of fruit and vegetable consumption (Supervisor: Dr. Sisira Sarma)

Technical and Research Experience:

Graduate Research Assistant (Western University) — 2018 – 2020
- Contributed to data extraction and manuscript writing for a large systematic review of quantitative intersectionality methods
- Data analysis and manuscript writing for a publication comparing measures of visible minority status, and a commentary piece on quantitative intersectionality methods
- Codebook creation and data entry for a national survey (Trans PULSE Canada)

Clinical Intern - Lallemand Health Solutions (Montreal, QC) — 2016 –2017
- Conducted literature reviews on clinical research topics related to probiotics
- Developed clinical study protocols and other documentation in accordance to Health Canada regulatory standards
- Contacted third parties for the completion of clinical studies

Canadian League Against Epilepsy Undergraduate Summer Studentship — 2016
- Published a systematic review for unmet healthcare needs among patients with epilepsy

Software knowledge: • Advanced: SAS, STATA, and R • Intermediate: Visual Basic

Teaching and Leadership Experience:

Teaching Assistant (*Introduction to Epidemiology 2200*) 2019, 2020
- Taught one-hour weekly tutorials to undergraduate students, marked assessments and responded to student inquiries regarding course content

Western Epidemiology and Biostatistics Student Council 2017-2020
- Co-Chair (Sept. 2019 – Apr. 2020): Oversee operations of council, responsible for overall coordination of events and liaising between staff and faculty, advocated for student body at faculty meetings
- Special Events Coordinator (Sept. 2018 – August 2019): Organized and facilitated four-day workshop series, coordinating between presenters and attendees

Research Mentor 2018-2019
(Western U: Canadian Coalition for Global Health Research)
- Mentored 6 undergraduate students to conduct a health research project and present at the Western Student Research Conference (Title: *An exploratory study of the relationship between lack of access to healthcare and the prevalence of COPD, cardiovascular disease, and hypertension in Canada)*

Publications and Presentations:

Publications

Bauer GR, **Mahendran M**, Braimoh J, Alam S, Churchill S. Identifying Visible Minorities or Racialized Persons on Surveys: Can We Just Ask?. Submitted to Canadian Journal of Public Health. Under revisions.

Lizotte D, **Mahendran M,** Churchill SM, Bauer GR. Math versus meaning in MAIHDA: A commentary on multilevel statistical models for quantitative intersectionality. Social Science & Medicine. 2019 Aug 24:112500.

Mahendran M, Speechley K, Widjaja E. Systematic review of unmet healthcare needs in patients with epilepsy. Epilepsy & Behavior 2017; Oct (75):102-109.

Poster presentations

Mahendran M, Bauer G, Lizotte D. Evaluating quantitative methods for intersectionality research: a simulation study. Poster presented at: 2019 Canadian Society of Epidemiology and Biostatistics Biennial Conference; 2019 May 13-15; Ottawa, ON. *(poster presented by Mayuri Mahendran)*

Mahendran M, Sarma S. Trends in socioeconomic inequality of fruits and vegetables consumption in Canada. Poster presented at: Canadian Research Data Centre Network 2018 National Conference; 2018 Oct 18-19; Hamilton, ON. *(poster presented by Mayuri Mahendran)*

Awards:

CIHR Canadian Graduate Scholarship ($17500) – offered but declined 2019
Ontario Graduate Scholarship ($15000) 2019
Western Graduate Research Scholarship ($3000 annually) 2018-2020
Western Gold Medal recipient for Epidemiology and Biostatistics 2018
- For highest standing in graduating class

Dr. Karen Campbell Undergraduate Award ($500) 2018
- Departmental award for honours thesis project

Four Year Continuing Admission Scholarship ($10000) 2013 – 2015, 2017
Canadian League Against Epilepsy Undergraduate Summer Studentship ($5000) 2016
Scinapse Undergraduate Science Case Competition: Bronze Recipient 2015
Western's 125th Anniversary Alumni Award ($1500) 2014
Laurene Paterson Estate Scholarship ($1600) 2014