Characterizing Standard Variable Importance Measures and Growth Modeling of Bangladeshi Children over Two Years of Life

Heather Lynn Cook

Marshall, Illinois

B.S., Roanoke College, 2014

M.S., University of Virginia, 2016

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Statistics

University of Virginia

August 2019

Abstract

Creating interventions to avoid adverse events is an ongoing topic in numerous settings and thus it is often important to answer questions such as which treatments can be applied to avoid outcomes such as death or stunted growth. One may hope to answer these queries through the use of variable importance measures and through modeling the growth and development of individuals. Variable importance is an up and coming aspect of statistics that ranks variables in terms of some measure of importance which is often applied without the notion of either the exact meaning or how it can be compared with other regression or classification methods. Thus, characterizing standard variable importance measures could go a long way in the applicability and practicality of these ideas. In addition, confidence intervals and a lower threshold of importance was created and explored in order to advance the understanding and interpretability of such measures and methods. Simulations were conducted to show the behavior of such metrics with theoretical results stemming from a simple setting. In a specific population of Bangladeshi children from the PROVIDE study, growth models were explored where previous models have not correctly described these children's heights over the first two years of time, especially considering the plethora of covariates (900+). Developmental outcomes from 2 to 5 years of age were additionally modeled and explored. Throughout this research, the variable importance is described and explored in diverse manners while the children's heights and development is explained through various inclusive models.

Contents

Acknowledgement

Throughout the research, writing, and preparations of this dissertation, I have had a great deal of support from many aspects of my life. Firstly, I would like to express my great appreciation to Dr. Daniel Keenan and Dr. Douglas Lake, my advisers, for their constructive suggestions and wonderful enthusiasm in the planning and development of this research. I am very thankful for their willingness to meet with me weekly, sometimes even more than once a week, even though some progress on my part was rather slow. However, their support and expertise within statistics is invaluable.

A special thank you goes out to Dr. Jennie Ma and Dr. William Petri, Jr. for their valuable guidance along with Dr. Jianhui Zhou for his previous work and all three for their knowledge about the PROVIDE study cohort. I have really enjoyed working with the PROVIDE group and am entirely grateful for the wonderful opportunities I have had through the research assistantship from the published articles to just the pure experience of working with real data and applications. These are experiences I will apply within my classroom to help my future students understand the realities of working with actual observations. Another special thank you goes out to our department administrator, Karen Dalton, who is always so incredibly helpful and keeps us organized.

In addition, many friends, family, and others deserve thanks for their continued support in various ways: My parents first and foremost for raising me with a hard work ethic and the attitude of never giving up on my goals. My sister for pushing me when it was hard to find motivation. My friends for being so understanding at my lack of communication as I finished my studies. And last but not least, my husband for cleaning, cooking, and generally keeping me stress free at home.

I could not have completed this dissertation without any of these individuals and I will always be thankful and grateful.

1 Introduction

Infants born prematurely not only have a very low birth weight, but are at high risk of death and other morbidities such as bronchopulmonary dysplasia, intraventricular hemorrhage, and retinopathy of prematurity. Perinatal measurements available at birth such as weight, gestational age, and Apgar scores often determine an accurate baseline risk. However, vital sign measures such as heart rate and blood oxygen levels that are measured every second or two while in the Neonatal Intensive Care Unit (NICU) can significantly modify an infant's baseline risk. Features calculated from the vital sign time series that can be used to develop predictive risk models include metrics of average, variability, and skewness. Many of these measurements are highly correlated and the relationships with the undesirable outcomes are still being explored. Therefore, an important often asked question is which of these measurements should be monitored closely in order to lower the patient's risk of an adverse event?

It is becoming more well known how the first few years of life can impact the future health of an individual. A particular problem of stunted growth occurs too often throughout the world and contributes to mortality of individuals under five and overall developmental detriment. Being able to explain how children who are particularly susceptible to stunting, developmental delays, or malnutrition grow and which factors affect their health would allow for interventions and early detection which in turn would improve the lives of many. Within this study, hundreds of variables (> 900) involving environmental, maternal, blood and stool, growth, and developmental factors were measured on the PROVIDE cohort with 700 children from Bangladesh who were followed irregularly from birth through two years of life with 16 time points. A subset of these children were additionally followed up to seven years of life thus far. This data set and similar cohorts have been explored using functional principal component analysis (FPCA) with linear regression by Zhang et al., penalized linear regression by Lu et al., a combination of these two methods (working manuscript), and conditional ran-

dom forest by Donowitz et al. [Zhang et al., 2017] [Lu et al., 2017] [Donowitz et al., 2018] (coauthor). Of these analyses, only the working manuscript uses the actual heights of children whereas the others have explored the height-for-age Z score (HAZ) which is normalized for gender and age using the World Health Organization (WHO) Multicentre Growth Reference Study Child Growth Standards. Also, in the second article, only certain variables were considered of which some were measured at multiple time points with the aspect of repeated measures being unaccounted for within the analysis. Other articles focused on neurocognitive developmental outcomes such as in Moreau et al., Donowitz et al., and Jensen et al. or outcomes such as time to infection of cryptosporidiosis with survival analysis in two articles by Steiner et al. [Moreau et al., 2019] (coauthor) [Donowitz et al., 2018] (coauthor) [Jensen et al., 2019] [Steiner et al., 2018] (coauthor) [Steiner et al., 2019] (coauthor). Therefore, a more comprehensive model which includes numerous covariates over time needs to be explored and created in order for interventions to be discovered.

Both of these data sets give rise to similar questions in that we are asking which variables are most important for predicting complications in order to intervene. Extending our question to other populations and more general problems, a common practical question deals with selecting which variables are most important or predictive. Thus, the idea of measuring a variable's importance is very practical. The two data sets described above are both unique. For the NICU data, the variables are mostly numeric and there is only a small set of predictors while for the PROVIDE study, numerous variables were measured and have the potential to be important for the outcomes of interest.

Variable importance (VIMP) measures exist for linear regression and other such procedures, however most of these measures have issues with one being they are not applicable to separate methods one may wish to use to compare results. Logistic regression has been used in a couple cases from these data sets one including a VIMP measure for NICU infants by Sullivan et al. and one which may have been improved with the addition of a VIMP measure for patients with Clostridium difficile infection by Kulaylat et al. [Sullivan et al., 2018]

[Kulaylat et al., 2018] (coauthor). Another such example of an existing procedure for VIMP is using a conditional random forest along with conditional VIMP for the PROVIDE cohort [Donowitz et al., 2018] (coauthor) [Moreau et al., 2019] (coauthor). This procedure gives a very intuitive graphical display of the VIMP, however the scale is not comparable with other such methods and the calculations are somewhat complex. Thus, characterizing these standard VIMP measures would be beneficial to many individuals. The absence of full understanding of VIMP measures which is useful for various regression and classification methods drives us to explore these VIMP calculations. In order to explore VIMP measures, multiple analyses using the existing data along with creating new data for simulations is implemented. Some simulations where the truth about the VIMP is known for simple regression settings are completed while an exploration of creating confidence intervals and finding a cutoff for VIMP is explored using the NICU and the PROVIDE study data.

The first objective for this dissertation includes presenting and characterizing standard VIMP measures for different data sets and thus in various settings. The goal is to provide measures and their characterizations which allow for the identification of important variables which predict an outcome where interventions may be imposed to avoid adverse events. Secondly, the growth modeling process has a goal to build comprehensive models which explain the growth and development of children from Bangladesh who specifically have a lack of proper growth. This objective also has the idea to provide interventions to avoid stunting or neural deficits within this particular population by including numerous covariates to predict height and developmental responses at two to five years of age.

In Chapter 2, the data behind the motivation for VIMP is first described along with a data set with additional complexities. Then current VIMP measures with their criticisms are reported. Next, VIMP measures are characterized using the two data sets in the classification and regression settings including calculations for confidence intervals of VIMP and a cutoff/threshold to determine which variables are important. Lastly for Chapter 2, a theoretical query is postulated along with supporting simulation results.

Chapter 3 describes the growth modeling portion for the PROVIDE study cohort which is first described in detail. Multiple current growth models and a proposed growth model are outlined. Then, simple exploratory results on the data are presented followed by various statistical models including penalized linear regressions and random forests such that comprehensive models are suggested per outcome of interest.

The last chapter, Chapter 4, summarizes the findings and suggestions and then suggests future research paths from this current work. Chapters 2 and 3 were designed to be self contained in the hopes of publishing articles about these separate Chapters, each of which include their own unique aspects but are linked in terms of the overall goal to identify important variables leading to interventions.

2 Characterizing Variable Importance Measures

Variable importance (VIMP) is somewhat of a vague calculation for many who apply the methods which automatically output VIMP measures. However, VIMP is a very useful and often practical application. Often, VIMP is used to select predictors in order to create interventions for improving health or other aspects of life. Another use is for exploratory reasons where important predictors of the response may be identified and interpreted [Grömping, 2009]. The data fueling our motivation is first described followed by multiple current VIMP measures and their criticisms. Then, methods for confidence intervals and finding a threshold are discussed within the applications of available data. Finally, theoretical aspects about the probability for correctly selecting the one and only important variable in a simple setting are given supported with multiple simulations.

2.1 NICU Data

Preterm newborns are inherently more vulnerable to certain morbidities including death, intraventricular hemorrhage (IVH), bronchopulmonary dysplasia (BPD), late-onset septicemia (LOS), necrotizing enterocolitis (NEC), and retinopathy of prematurity (ROP). IVH is not necessarily fatal, however it may cause other complications since IVH is bleeding within the brain confined to areas of the brain which contain spinal fluid. BPD is damage to the lungs usually caused by ventilation and long term oxygen use, yet most patients will recover. LOS is sepsis which is inflammation throughout the body as a reaction to an infection which may result in multiple organ failures and death. NEC may also lead to death or an infection since it happens when a portion of the bowel dies. However, ROP does not usually lead to death but is rather abnormal blood vessels throughout the retina causing blindness or other eye problems. Death and IVH occur earlier within a patient's stay at the NICU. Thus, these would need early interventions whereas BPD, LOS, NEC, and ROP may occur during a later time within a patient's stay.

Previous work showed abnormal heart rate characteristics such as the decrease in HR variability predict death and other morbidities in premature infants [Sullivan et al., 2016]. However, these later morbidities (BPD, LOS, NEC, and ROP) have proved harder to predict from the following measures: weight at birth in grams (BW); sex; gestational age in weeks (GA); Apgar scores at one and five minutes; antenatal steroid indicator (number of steroid doses given before birth); the mean, standard deviation, skewness, or kurtosis of heart rate (HR) and blood oxygen levels (SPO_2); and the minimum and maximum of the cross-correlations between HR and SPO_2 [Sullivan et al., 2018]. The HR and SPO_2 were measured every few seconds over the week from available pulse oximetry data. The statistics of HR and SPO_2 were calculated after twelve hours and at the end of seven days. An Apgar score is a measure of the physical condition for a newborn which takes into account the following five aspects each with a perfect rating of two: heart rate, respiratory effort, muscle tone, reflexes, and skin color. The score one minute after birth tells how well the infant is tolerating the birthing procedure while the five minute score shows how well the newborn is doing after birth, or outside the womb [Kaneshiro, 2014]. Some of these variables are highly correlated as to be expected especially since premature infants usually weigh less; see Figure 1. These correlations are based on Spearman's correlation and have been split based on previous results. However, this adds a slight complication to the data structure along with multiple non-Normal distributions and some low incidence rates. It has been shown that the first 12 hours of data better predicts death and IVH while the first week of data better predicts the other outcomes due to when these events usually occur within a NICU stay [Sullivan et al., 2018] [Sullivan et al., 2016]. For this research, the focus was on the worst outcome of death. Thus, only the first 12 hours of data was used for calculating some of the predictors.

This data was collected at two separate sites, the University of Virginia (UVA) Children's Hospital from 2012-2015 and Morgan Stanley Children's Hospital of NewYork-Presbyterian Columbia University (CU) Medical Center from 2012-2015 to include a total of 778 infants

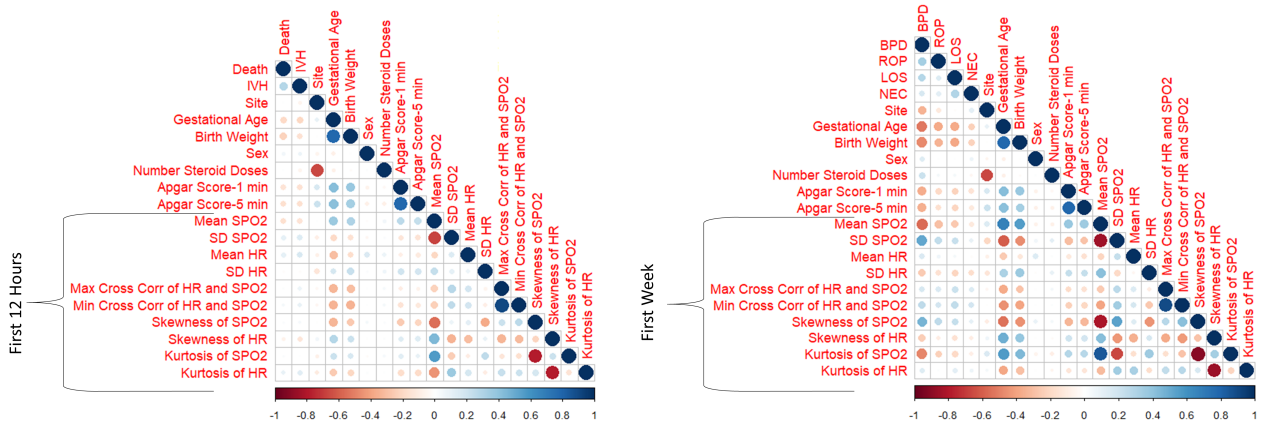

Figure 1: The Spearman correlations per the first 12 hours (left) and the first week (right) of data with the corresponding appropriate outcomes.

(443 from UVA and 335 from CU). Children with congenital heart defects, congenital anomalies, extreme prematurity which prompted planned comfort care only, or those with missing pulse oximetry within 12 hours of birth were excluded from the data set. This data set stands as our motivation for the VIMP exploration with the outcome of death and is used for the application aspects in 2.4.1 and 2.5.1.

2.2 PROVIDE Cohort Study Data

The PROVIDE birth cohort consisted of 700 infants born in Mirpur which is an urban slum in Dhaka, Bangladesh from May 2011 to November 2014. Children were recruited at birth and followed over a two-year period with in-home visits twice a week and irregularly scheduled clinical visits where blood or stool samples were occasionally taken. A more detailed description of the study design, recruitment, and follow-up were described previously [Kikpatrick et al., 2015]. This study was approved by the Ethical Review Board of the ICDDR,B (FWA 00001468) and the Institutional Review Boards of the University of Virginia (FWA 00006183) and the University of Vermont (FWA 00000727). A large set of biomarkers for nutrition and systemic inflammation were calculated from the available stool and blood samples along with numerous survey results, development, and growth measures including over 900 potential predictor variables. However, due to numerous missing values and in order to keep as many subject as possible, a subset of the predictors was selected for analysis including biomarkers, socioeconomic, and anthropometric measures. Thus, data is still from all sorts of sources and on numerous aspects of these children's lives. The primary outcome of interest is stunting by two years of age defined as a height-for-age Z score (HAZ) or length-for-age (LAZ) at two years below -2. HAZ specifically is a measure normalized for the child's age and gender from standards released from the World Health Organization (WHO) Multicentre Growth Reference Study Child Growth Standards. Stunting has been shown to be correlated with subsequent outcomes in later life such as diminished survival, weakened learning capacity, and lower annual incomes which leads to stunting being a primary interest. However, HAZ is a commonly used measurement for malnutrition due to it's ability to capture the cumulative effects through childhood and is our outcome of interest at two years of age [Dewey and Begum, 2011][Hoddinott et al., 2008]. The relationships between the 47 predictors are given in Figure 2 [Donowitz et al., 2018] (coauthor). This figure shows how the predictors may be clustered together based on Pearson correlations, especially how

systemic cytokines are related to each other (black cluster), how enrollment anthropometry is related to sanitation (red cluster), and how economic status is clustered with biomarkers for enteric inflammation (green cluster). It may be noted that for this particular data set, if a child had any missing values across the variables, they were excluded along with any subject having a value above five standard deviations in any of the predictors leading to 371 subjects to analyze. This specific subset of the PROVIDE study is used in 2.4.2 and 2.5.2 for applications aspects.

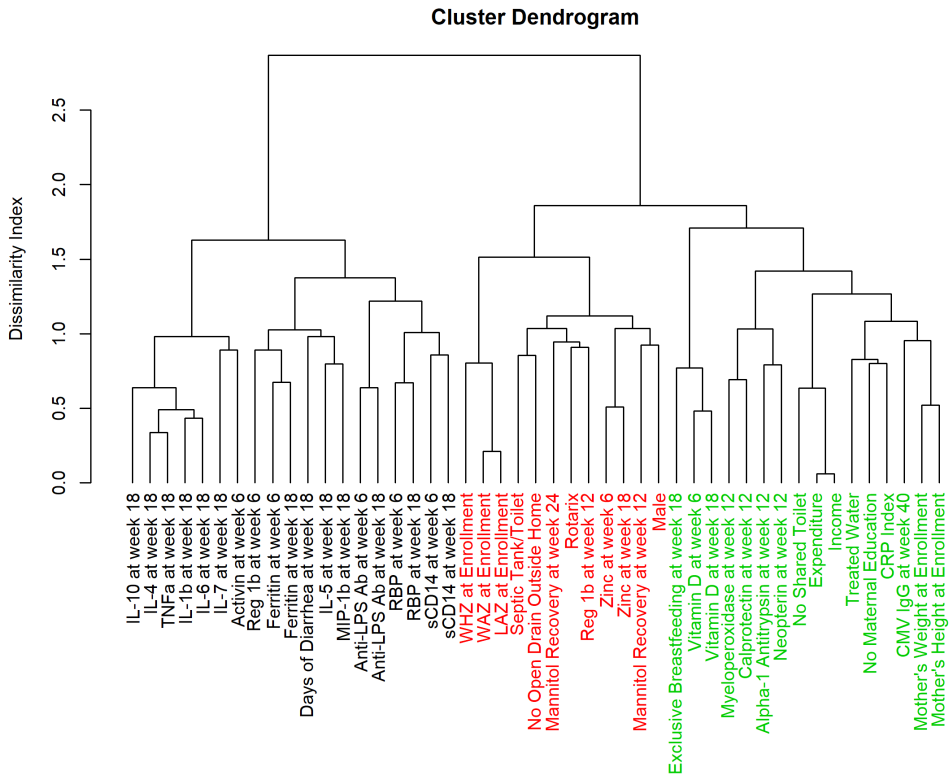

Figure 2: The Pearson correlations between the predictors for the PROVIDE study cohort were used to create this hierarchical cluster dendogram.

2.3 Current Variable Importance Measures with Criticisms

There are currently numerous different measures of a variable's importance, often depending on the specific setting and model one is working with. Simple methods such as linear and logistic regression have many different ways to calculate variable importance (VIMP) while more complex methods such as neural networks or random forests only have a few. The following outlines several ways one may choose to calculate VIMP along with certain criticisms.

2.3.1 Linear and Logistic Regression Setting

There have been various variable importance (VIMP) measures proposed for linear and logistic regression methods. These measures may be seen in Table 1. Since linear and logistic regression are widely used, there have been many different proposed methods of VIMP with some of the simpler ideas coming naturally. The absolute value or squares of the raw coefficients β_j, standardized coefficients $\beta_{j,st} = \beta_j \frac{s_j}{s_y}$, test statistics (t-values in linear regression and z-values in logistic regression), or p-values have all been used in numerous instances. However, each of these have their own criticisms. As is widely known, the raw coefficients β_j are not scale invariant and thus their interpretations and values depend on the initial scale of the predictors. Even though the standardized coefficients $\beta_{j,st}$ take care of this scale invariance, they are still not very useful when correlations appear between predictors [Grömping, 2015]. Each of these VIMP measures also are conditional on all other regressors in the model which, depending on the research question, may not be quite as useful as a marginal approach. A specific example in Figure 3 from the NICU data has shown how p-values from a logistic regression model have been used to rank the predictors from most significant to least significant per each adverse event and split by pulse oximetry or clinical variables. From this, we can see that for the clinical variables and the outcome of death, birth weight is ranked the highest whereas for the pulse oximetry measures and still for

the outcome of death, the mean blood oxygen level is ranked the highest indicating these are important variables for the prediction of death [Sullivan et al., 2018]. The last measure which may be applied to linear and logistic regression both is the sequential increase in R-squared where each regressor is entered into the model in a pre-specified order. This method can decompose the variance but is often not practical due to the dependence on the order the variables are entered [Grömping, 2015].

Coefficient rank in multivariate logistic regression models to predict each outcome using pulse oximetry measures (A) or clinical variables (B)

Outcome	Mean SPO_2	SD SPO_2	Mean HR	SD HR	Max XC HR-SpO_2	Min XC HR-SpO_2	Skewness SPO_2	Skewness HR	Kurtosis SPO_2	Kurtosis HR
Coefficient rank A										
Died	1[a]	6	2[a]	4	7	8	10	5	3	9
sIVH	10	2	1[a]	3	5	6	8	4	9	7
BPD	3[a]	2[a]	10	9	8	7	5[a]	4[a]	6[a]	1[a]
tROP	2[a]	10	8	6	4	5	9	3[a]	7	1[a]
LOS	8	4	10	1[a]	2	9	6	3	7	5
NEC	6	7	5	1[a]	10	2	4	8	3	9
Outcome	**GA**		**BW**		**Sex**	**Antenatal steroids**	**Apgar 1 min**	**Apgar 5 min**		**Site**
Coefficient rank B										
Died	5		1[a]		3	7	4	2[a]		6
sIVH	2[a]		7		4	1[a]	5	6		3[a]
BPD	3[a]		2[a]		4[a]	7	5	6		1[a]
tROP	1[a]		3[a]		7	6	4	5		2[a]
LOS	2[a]		1[a]		7	4	6	3		5
NEC	2[a]		6		3[a]	4	5	7		1[a]

Abbreviations: BPD, bronchopulmonary dysplasia; BW, birth weight; GA, gestational age; HR, heart rate; LOS, late-onset septicemia; NEC, necrotizing enterocolitis; SD, standard deviation; sIVH, severe intraventricular hemorrhage; SpO_2, oxygen saturation; tROP, treated retinopathy of prematurity; XC, cross-correlation.
[a] $p \leq 0.05$.

Figure 3: The predictors in the NICU data were ranked per each set of variables, clinical or pulse oximetry, via their corresponding p-value for the logistic regression analyses per each outcome [Sullivan et al., 2018].

Specifically for linear regression, multiple other VIMP measures have been assessed. The simple metrics include the absolute value or square of the raw correlations r_{YX_j} between each regressor X_j and the response Y, the absolute value or square of the semipartial correlations $r_{Y(X_j,other)}$, and the product of the standardized coefficients and their respective raw correlations $\beta_{j,st} r_{YX_j}$. The semipartial correlation between Y and X_1 when given the response Y

and two predictors X_1 and X_2 is

$$r_{Y(X_1,X_2)} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{X_1X_2}}}$$

which represents the correlation between Y and X_1 after the effect X_2 has on X_1 (but not the effect it has on Y) has been removed. Again, these have all been scrutinized for various reasons. For the raw correlations r_{YX_j}, these values show the marginal effects only and thus could be deemed important even though the coefficient in a linear regression is zero, meaning it would have no impact on the response given all other variables in the model. Likewise with the raw coefficients β_j or test statistics, the semipartial correlations $r_{Y(X_j,other)}$ are also conditional on all the other variables within the model which, again, may not answer the particular research question. The product of the standardized coefficients and the raw correlations, $\beta_{j,st} r_{YX_j}$, may decompose the R-squared or variance, but negative contributions might arise which is often criticized [Grömping, 2015]. Other metrics such as zero-order correlations (raw correlations) r_{YX_j}, partial correlations $r_{YX_j,other}$, Akaike weights w_i, and independent effects I_{X_j} were explored leading to the conclusion that no index for linear regression performed perfectly especially when correlations between predictors occurred [Murray and Conner, 2009]. The partial correlations assuming the response Y and two regressors X_1 and X_2 are calculated as

$$r_{YX_1,X_2} = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{\sqrt{1 - r^2_{YX_2}}\sqrt{1 - r^2_{X_1X_2}}}$$

and is thus the correlation for Y and X_1 after the effect of X_2 on both Y and X_1 has been removed. The Akaike weights is calculated given the data set and M candidate models. Thus, the Akaike weight for the ith model is

$$w_i = \frac{\exp\left(-\frac{1}{2}(AIC_i - AIC_{min})\right)}{\sum_{m=1}^{M} \exp\left(-\frac{1}{2}(AIC_m - AIC_{min})\right)}$$

where AIC_{min} is the Akaike information criterion (AIC) value for the model with the lowest AIC. The VIMP using these weights is the sum of the weights, w_i, across all models including the variable X_j. So, the higher this sum of the weights, the more important variable X_j is, however this VIMP measure is based on the number of models which contain X_j. Therefore, the number of computing models containing each variable must be balanced. The independent effect of a particular variable I_{X_j} indicates the average contribution of this variable to the variance for all response values over every possible model. This value is calculated through the comparison of the fit of all models with the predictor and the fit of all possible nested models without this predictor. With the response Y and P predictor variables, the independent effect of X_j is

$$I_{X_j} = \sum_{i=0}^{P-1} \frac{\sum \left(r^2_{Y,X_j X_h} - r^2_{Y,X_h} \right) / \binom{P-1}{i}}{P}$$

where X_h is some subset of i regressors where X_j is excluded. This is similar to dominance analysis but using hierarchical partitioning. Murray and Conner recommend using the zero-order correlations first for predictors with near zero correlations, then using the independent effects to rank the predictors [Murray and Conner, 2009]. This recommendation is flawed in the case of correlations between predictors since the squared correlations will no longer sum up to be the R-squared value and thus do not reflect their true contributions [Murray and Conner, 2009]. Even though these metrics are often used, they all have been scrutinized and two glaring issues are that these measures are limited to specific regression methods and do not share the same scales of VIMP.

Additional but more complex VIMP measures have been proposed for linear regression. These additional VIMP metrics are based on the decomposition of the variance (See Table 1). These methods are referred to as LMG (Lindeman, Merenda, and Gold), PMVD (Proportional Marginal Variance Decomposition), Gibson/CAR scores, and Fab-

bris/Genizi/Johnson. The variance of a linear regression model can be written as

$$\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{XX} \boldsymbol{\beta} = \sum_{j=1}^{P} \sum_{k=1}^{P} \beta_j \beta_k \sigma_{jk}$$

where $\boldsymbol{\beta}^T$ excludes the intercept, $\boldsymbol{\Sigma}_{XX}$ represents the true but unknown covariance matrix $(P \times P)$ of the predictors where σ_{jk} are the elements and the matrix can be rewritten as $diag_j(\sqrt{\sigma_{jj}})\mathbf{P}_{XX}diag_j(\sqrt{\sigma_{jj}})$, and where $\mathbf{P}_{XX}$ is the theoretical correlation matrix ($\mathbf{P}$ is capital Rho). For the following explanations, the data involved is assumed to be centered such that the empirical covariance matrices are $\mathbf{S}_{XX} = \mathbf{X}^T\mathbf{X}/(n-1)$ and $\mathbf{S}_{XY} = \mathbf{X}^T\mathbf{Y}/(n-1)$.

LMG and PMVD are both computationally expensive algorithms which are related to game theory. The regressors are the players of the game and the variances that can be explained by a set of predictors is the worth for that particular set of predictors while the achievable variance explained overall is the total gain or worth which can be allotted among all regressors fairly. LMG and PMVD are both averages of the sequentially explained variances over all the possible orderings of the regressors. LMG is the unweighed version and tends to the marginal side where PMVD is the weighted version which tends to the conditional side of VIMP. Given disjoint sets of predictors S and M, the explained variance $(evar)$ and the sequentially added variance $(svar)$ are calculated as

$$evar(S) = var(Y) - var(Y|X_j, j \in S)$$

$$svar(M|S) = evar(M \cup S) - evar(S).$$

These formulas will allow us to define LMG and PMVD. Thus, the following shows the equations for the first predictor for simplicity. Here $S_1(\pi)$ is the set of predecessors (previous predictors) for variable 1 for permutation π.

$$LMG(1) = \frac{1}{P!} \sum_{\pi permutation} svar(\{1\}|S_1(\pi))$$

This shows that LMG is an unweighted average for every ordering of the sequential contribution from regressor 1. Since PMVD is also an average over all orderings of the sequential contribution for regressor 1, although a weighted average, we may write

$$PMVD(1) = \sum_{\pi permutation} P(\pi) svar(\{1\}|S_1(\pi))$$

where the weights

$$P(\pi) = L(\pi)/\sum_{\pi} L(\pi)$$

and where

$$L(\pi) = \prod_{i=1}^{P-1} svar(\{\pi_{i+1}, \ldots, \pi_P\}|\{\pi_1, \ldots, \pi_i\})^{-1} \quad (1)$$

$$= \prod_{i=1}^{P-1} (evar(\{1, \ldots, P\}) - evar(\{\pi_1, \ldots, \pi_i\}))^{-1} \quad (2)$$

These methods fail however when the number of regressors is large, especially due to their expensive computations [Grömping, 2015] [Grömping, 2009].

In addition to the already assumed centered data, the following variance decomposition methods require the data to be standardized. So, the empirical correlation matrices for the normalized data are $\mathbf{R}_{XX} = \mathbf{X}^T\mathbf{X}/(n-1)$ and $\mathbf{R}_{XY} = \mathbf{X}^T\mathbf{Y}/(n-1)$. Keeping this in mind, the next variance decomposition method is Gibson/CAR scores which uses the squared coefficients, c_j^2 with $j = 1, \ldots, P$, from the predictors when the normalized outcome variable is regressed on an orthogonalized matrix $\mathbf{Z}$. To find the best matrix $\mathbf{Z}$, we assume full column rank for $\mathbf{X}$, and then the orthogonalization from singular value decomposition begins with $\mathbf{X} = \mathbf{UDV}^T$ where $\mathbf{U}$ is a $n \times P$ matrix with orthonormal columns, $\mathbf{V}$ is an orthogonal matrix ($P \times P$), and $\mathbf{D}$ is a diagonal $P \times P$ matrix. $\mathbf{Z} = \mathbf{UV}^T$ gives the set of P orthonormal vectors that is most similar to the $\mathbf{X}$ variable's columns. Thus,

$$\mathbf{R}_{XX} = \mathbf{X}^T\mathbf{X} = \mathbf{VDU}^T\mathbf{UDV}^T = \mathbf{VD}^2\mathbf{V}^T$$

such that $\mathbf{R}_{XX}^{-1/2} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T$. This therefore leads to

$$\mathbf{Z} = \mathbf{X}\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}^{-1}\mathbf{V}^T = \mathbf{U}\mathbf{V}^T.$$

To find the squared coefficients from regressing $\mathbf{Y}$ on $\mathbf{Z}$ one can simply square the components of $\mathbf{R}_{XX}^{-1/2}\mathbf{R}_{XY}$. Even though the orthogonalized matrix $\mathbf{Z}$ is a surrogate for the normalized $\mathbf{X}$ values, when correlations are present between predictors, this matrix may not be the best choice. Thus, the Fabbris/Genizi/Johnson method modified this approach by calculating the squared coefficients from regressing the original predictors $\mathbf{X}$ on the orthogonalized matrix $\mathbf{Z}$ then obtaining the R-squared contribution for an original predictor variable via a weighted sum of the squared coefficients similar to the Gibson/CAR scores method. This Fabbris/Genizi/Johnson method is regarded as the better of the two by some whereas others state the measure as theoretically flawed due to the choice of relative weights [Grömping, 2015]. One important note is that variance decomposition in linear models where correlations between predictors are present is still being researched and thus these methods are not applicable in that case.

Often, it is discussed whether VIMP measures should be marginal or conditional or balance the two. For the VIMP measures given thus far, many of them are one or the other. Table 1 distinguishes which VIMP measures take either the marginal or conditional approach and which try to balance these aspects. The conditional types of VIMP are good when one wants to select a small number of predictors for accurate predictions. Marginal types are better for interpretations or explanations when selecting the important predictors. However, one always must understand the relationships within the data, meaning correlation between predictors can have a large effect on some VIMP measures, especially the marginal ones.

Table 1: Linear and Logistic Regression VIMP Measures.

Linear/Logistic	**VIMP Measure**	**Type**
Both	Absolute value or square of raw coefficients	Cond
Both	Absolute value or square of standardized coefficients	Cond
Both	Absolute value or square of test statistics	Cond
Both	Absolute value or square of p-values	Cond
Both	Sequential increase in the R-squared	Cond
Both	Akaike weights	Cond
Linear	Absolute value or square of raw correlations	Marg
Linear	Absolute value or square of partial correlations	Marg
Linear	Absolute value of square of semipartial correlations	Cond
Linear	Independent effects	Marg
Linear	Product of the standardized coefficient and raw correlation	Cond
Linear	LMG: Variance decomposition	Marg
Linear	PMVD: Variance decomposition	Cond
Linear	Gibson/CAR Scores: Variance decomposition	Both
Linear	Fabbris/Genizi/Johnson: Variance decomposition	Marg

2.3.2 Random Forest and Conditional Random Forest Setting

In order to understand the calculations for VIMP within a tree-based model, one must first understand a basic decision tree. In Figure 4 panel A, the terminology of a basic decision tree is given where each split is at a decision node and the end of a branch is a terminal node. An example is given in Figure 4 panel B. For this example, given an individual, we follow the tree from top to bottom through the decision nodes ending at a terminal node which will give a classification, in this case, of high or low mileage. Let's say we have a 2012 Toyota Camry to classify. Following the decision tree in Figure 4 panel B, the Toyota is not heavy and has a horsepower greater than 86 leading to a low mileage classification. A classification or regression tree searches through each predictor at each split to find a value from one variable which will divide the data into two groups based on some splitting criteria. The splitting continues until some stopping criteria is met and thus a tree is grown. These trees usually do not have many restrictions and may grow very large (imagine a more complex setting with numerous predictors unlike Figure 4). However, they may be controlled by

certain parameters such as the minimum number of observations that is allowed within a terminal node. A strength of these trees is that they can handle several different types of predictors, although they can suffer from model instability meaning each tree is dependent on the sample used to create it therefore giving a single tree low bias but high variability [Kuhn and Johnson, 2016]. To reduce the variability, methods such as bagged trees which averages individual trees together are implemented.

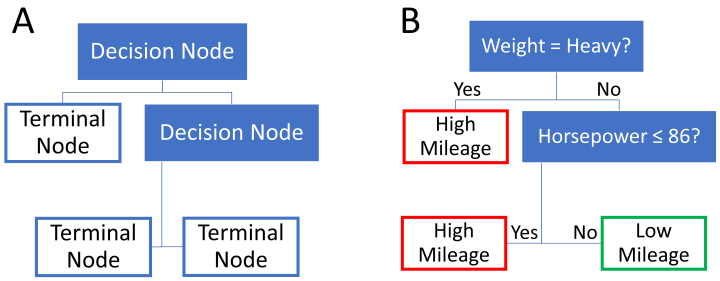

Figure 4: (A) Basic decision tree terminology and (B) an example to classify mileage of vehicles.

Bagging, aka bootstrap aggregation, is an ensemble method that aggregates the decision trees generated from bootstrapped samples as explained in Figure 5. Each of the trees is grown deep giving each tree low bias but high variability. The variability is reduced when averaged, but there are no tuning parameters which can lead overfitting the data. For a binary outcome, each tree can be thought of as casting a vote for which category that tree thinks the new observation should belong. The total number of votes for each category is divided by the total number of trees to produce the predicted probability for the new observation [Kuhn and Johnson, 2016]. These predicted probabilities can then be used to classify the new observation based on a decision threshold, which can be naively chosen as 0.5. One general downfall to aggregating trees is the loss of interpretability. A specific downfall for bagged trees is that the trees are not completely independent of one another. This is due to all of the predictors being considered at each split for every single tree in the ensemble

which leads to tree correlation. Tree correlation may prevent the method from optimally reducing the variance of the predictions since each tree can have comparable structures due to the underlying relationships.

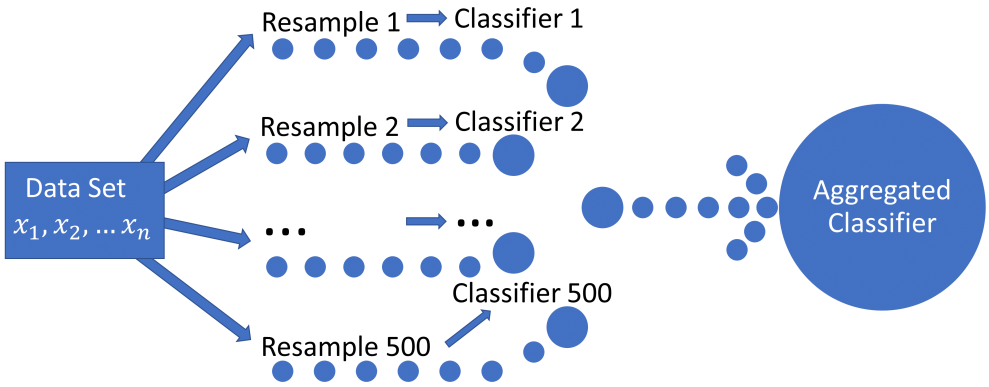

Figure 5: Flow chart of the general steps bagging, random forests, and conditional random forests follow.

An improvement from bagged trees is random forest (RF) where the trees are built on bootstrapped samples similar to bagged trees, however, each time a split in a tree is considered, a random sample of predictors is chosen as split candidates which decorrelates the trees. The split then only uses one of the sampled predictors at that decision node. At each split, or decision node, a fresh set of predictors is randomly chosen from which to select. Each tree is grown to the maximal depth and contributes equally to the final model, thus if the number of randomly sampled predictors equals the total number of predictors, then RF becomes bagged trees. Due to the lessening of tree correlation, the number of trees created does not attribute much to overfitting [Kuhn and Johnson, 2016]. It is important to note that when a tree is built, a portion of the observations are left out. These observations are often called out-of-bag (OOB) observations and are used to calculate VIMP as described below.

The steps for creating a RF model similar to Figure 5 are:

1. Select the number of trees to be built, T
2. For each tree, t, complete the following:
   (a) Select a bootstrap sample from the original data
   (b) Build a tree, t, for this particular sample but for each split, s, of the tree:
      i. Randomly select k predictors from the original list of predictors of size P
      ii. Select the best predictor among those k in order to split the data
   (c) Build the tree until the stopping criteria is met (could be built to maximal depth: one observations per terminal node)
3. Collectively, these T trees create the forest

An issue arises with the RF method though when predictors are highly correlated. The importance of correlated predictors may be overestimated making these predictors appear more important than uncorrelated ones [Strobl et al., 2008] [Strobl et al., 2009] [Boulestix et al., 2012]. The conditional random forest (CRF) method takes into account these correlations and other aspects to reflect the impact of a single variable conditioned on associated predictors in predicting the outcome. Thus, CRF uses an unbiased splitting criteria to avoid the issues that arise with highly correlated variables and is described in more detail below. A tuning parameter for both RF and CRF is the number of predictors to be randomly chosen at each decision node. It is important to tune this parameter since a small number of predictors chosen at each split can lead to choosing variables that are suboptimal and can lead to a loss of information [Boulestix et al., 2012][Strobl et al., 2008]. In any situation, a main goal is to have informative predictors identified in order to get the best predictions.

The VIMP measures for RF has been explored fairly extensively with different representations for the VIMP as seen in Table 3. Originally, the Gini index was postulated for the classification case. For the classification case, the decrease in a heterogeneity index is used

as a splitting criteria. For the jth variable at node s of the t^{th} tree, d_{js} is the decrease in a heterogeneity index (in our case it will be the Gini index). This predictor X_j is used in the split at node s if the decrease in the index for that variable is larger than the decrease in the index for any other predictor, $d_{js} > d_{ks}$ for $k = 1, \ldots, P$ where $j \neq k$. Then the VIMP for that variable X_j in the t^{th} tree is the sum of the decrease in heterogeneity when that variable is used in the split:

$$\widehat{\text{VIMP}}_{X_j}(t) = \sum_s d_{js} I_{js}$$

where I_{js} is the indicator that takes a value of one if the variable X_j is used in node s. So, the overall VIMP for X_j is the average over all the trees:

$$\widehat{\text{VIMP}}_{X_j} = \frac{1}{T} \sum_{t=1}^{T} \widehat{\text{VIMP}}_{X_j}(t)$$

For the calculation of the Gini index in the case of a binary outcome, a given sample is taken and for a given split of a particular variable at a specific node, the following are calculated: the number of samples at the node N, the number of observations to the left and right nodes respectively N_L and N_R, and the number of cases with the response of 1 and 0 denoted by N_1 and N_0. Then the empirical Gini index is computed using $\hat{p} = N_1/N$ in the following:

$$\hat{G} = 2\hat{p}(1 - \hat{p})$$

Then, the Gini gain, or the impurity reduction, at the particular node produced by the particular splitting cutpoint is:

$$d_{js} = \widehat{\Delta G} = \hat{G} - \left(\frac{N_L}{N} \hat{G}_L + \frac{N_R}{N} \hat{G}_R \right) = \text{VIMP}_{\text{Gini}}$$

where $\hat{G}_L$ and $\hat{G}_R$ are respectively the Gini indexes for the left and right nodes. Then for a particular tree the VIMP is the sum of the Gini gain when the particular variable is used for splitting. Thus, over an ensemble of trees, the VIMP for a certain predictor is the average over

all trees of this sum as shown above. However, if the true Gini index is $G = 2p(1-p)$ then the estimate $\hat{G}$ has a bias of $Bias(\hat{G}) = -G/N$. Another source of bias for this estimate comes from how predictors with multiple categories or continuous variables are favored since there are multiple possible partitions on which to split the data [Sandri and Zuccolotto, 2008]. This bias means the variables with more candidate splits are more likely to be chosen and at least one of the candidate splits may yield a good splitting criterion just by random chance. Similarly, there is a bias if certain categories have more cases than other categories, even if the predictors all have the same number of categories (unbalanced categories). Therefore, the Gini gain is expected to perform best when the covariates are all continuous without ties and uncorrelated or when the predictors are categorical but all have the same number of categories with similar sizes [Boulestix et al., 2012]. A bias correction for the Gini gain was proposed by Sandri and Zuccolotto (2008) where a set of uninformative noise variables are added to the original set of predictors. The VIMP of these uninformative variables can under certain conditions approximate the bias of the Gini impurity which is often unknown [Sandri and Zuccolotto, 2008].

Another process outlined in Figure 6 for RF with classification is a mean decrease in accuracy which takes advantage of the out-of-bag (OOB) samples as a test set of data as included in Table 3. Using the respective tree the OOB samples were not used to build, the predicted classes are found for these OOB observations and the number of correct predictions are summed to get $v_{t,OOB}$. Then, for a particular predictor, permute the values of that predictor only in the OOB observations and use the tree to again predict the classes for the OOB sample with that permuted variable. With these new predicted classes, sum the number of correctly predicted classes to get $v_{t,permutedOOB}$. Then, these two sums are subtracted, original sum minus the permuted sum, to obtain D_t and the VIMP is the average of these differences over the number of trees for that particular variable:

$$\text{VIMP}_{\text{Accuracy}} = \frac{1}{T}\sum_{t=1}^{T} D_t.$$

The process to calculate D_t is outlined as:

1. For each bootstrap sample and thus each tree built with that sample:
   (a) Identify the OOB observations
   (b) Using the tree, t, built from this bootstrapped sample, predict the class membership for the OOB observations
   (c) Sum the number of times tree t correctly predicts the class to get the number of votes for the correct class: $v_{t,OOB}$ (The error rate could also be calculated in this step.)
   (d) For each variable X_j with the OOB observations:
      i. Shuffle the values of X_j (permutation)
      ii. Use tree t to predict the class for these permuted values of X_j within the OOB data
      iii. Sum the number of times tree t correctly predicts the class to get the number of votes for the correct class: $v_{t,permutedOOB}$ (The error rate could also be calculated in this step.)
      iv. Calculate the number of votes for the correct class for the OOB samples minus the number of votes for the correct class for the permuted samples: $D_t = v_{t,OOB} - v_{t,permutedOOB}$ [Archer and Kimes, 2008]

The main idea behind this type of VIMP measure is that when the values are shuffled, the prediction accuracy will substantially decrease if that variable was indeed important for predicting the response [Strobl et al., 2009]. For this VIMP measure, Archer and Kimes state that the VIMP for a true predictor may be ranked within the top, but it may not have the maximum VIMP [Archer and Kimes, 2008]. The relationship of correlated variables was also found to affect VIMP. Specifically, even if a predictor was not associated with the response but the predictor was correlated to another predictor which was predictive of the response,

this collinear variable also received a high VIMP value [Archer and Kimes, 2008]. This makes distinguishing between possible causal predictors and these collinear predictors very difficult since they may not directly affect the response, but are deemed more important than an uncorrelated predictor with also no direct effect on the response. This issue has been described by many and there have been some given solutions with conditional VIMP being the main one [Boulestix et al., 2012]. Similar performance was found when the Gini index was used instead [Archer and Kimes, 2008]. Unlike the Gini gain though, the bias of the algorithm favoring predictors with many categories or continuous variables does not affect this permutation based VIMP. In the case of dealing with SNPs, the Gini index was found to favor the uncorrelated SNPs over highly correlated ones which gives rise to use the permutation based VIMP for similar situations [Boulestix et al., 2012]. Thus, permutation based VIMP may be better than the Gini impurity, but both are affected by correlation between predictors.

Another method is available for RF classification which outperforms the traditional metrics, specifically the permutation based VIMP measure, in the case of unbalanced data. This metric is simply replacing the error rate, or $v_{t,OOB}$ and $v_{t,permutedOOB}$, with the area under the curve (AUC) as stated in Table 3. Then we would have

$$\text{VIMP}_{AUC} = \frac{1}{T}\sum_{t=1}^{T}(AUC_{t,OOB} - AUC_{t,permutedOOB})$$

The AUC here is calculated as follows. Imagine a classification tree has been built and we have an OOB observation with a response value of 0 and another OOB observation with the response value of 1. A good tree is expected to give an observation with the true response value of 1 a higher class probability for that class, where the value is 1, than for the observation which belongs to the 0 class. The AUC is then the proportion of pairs for which this is the case within a particular tree, t. In other words, this metric is an estimation for the probability that a randomly chosen observation from the class 1 is given

a higher class 1 probability than a randomly chosen observation from class 0. Compared to the standard permutation based VIMP measure outlined above, this AUC based metric outperforms when the classes of the outcome are unbalanced. However, similar performance is seen when the class sizes are balanced. This is due to the fact that as the level of imbalance increases, the standard permutation VIMP measure loses it's ability to distinguish between the predictor variables which are actually associated with the outcome and those which are not [Janitza et al., 2013]. Therefore, this procedure may be favored when the class sizes of the outcome are unbalanced but is interchangeable when the outcome has balanced classes.

For the RF regression case, a permutation based mean square error (MSE) is given again using the OOB samples. Similar to the case above where the AUC or mean decrease in accuracy are calculated, this method also involves permuting one variable at a time for the OOB observations and recalculating the MSE for that particular tree. Then this new MSE is subtracted from the original MSE with no permutations in which the average difference in MSE over all trees is taken as the VIMP for the permuted variable as is outlined in Figure 6 [Grömping, 2009]. Therefore,

$$\text{VIMP}_{MSE} = \frac{1}{T}\sum_{t=1}^{T}(MSE_{t,OOB} - MSE_{t,permutedOOB}).$$

This metric, like the others for random forest is still strongly affected by correlations between predictors. This is due to the fact that when the variable's values are permuted, any associations with other variables are all lost. So, a variable that has no effect of its own on the outcome, but is correlated with a predictor which does have an effect on the response, will have an artificially high VIMP measure. Thus, the high VIMP may indicate the relationship between the variable and the response or it may represent the relationship between it and another predictor. This may show a variable which has no main effect on the response as more important than another predictor which isn't associated with any other predictors [Strobl et al., 2008] [Strobl et al., 2009]. Thus, in order to use any of the above

measures for random forest, one must know the structure of the data well, especially in terms of associations.

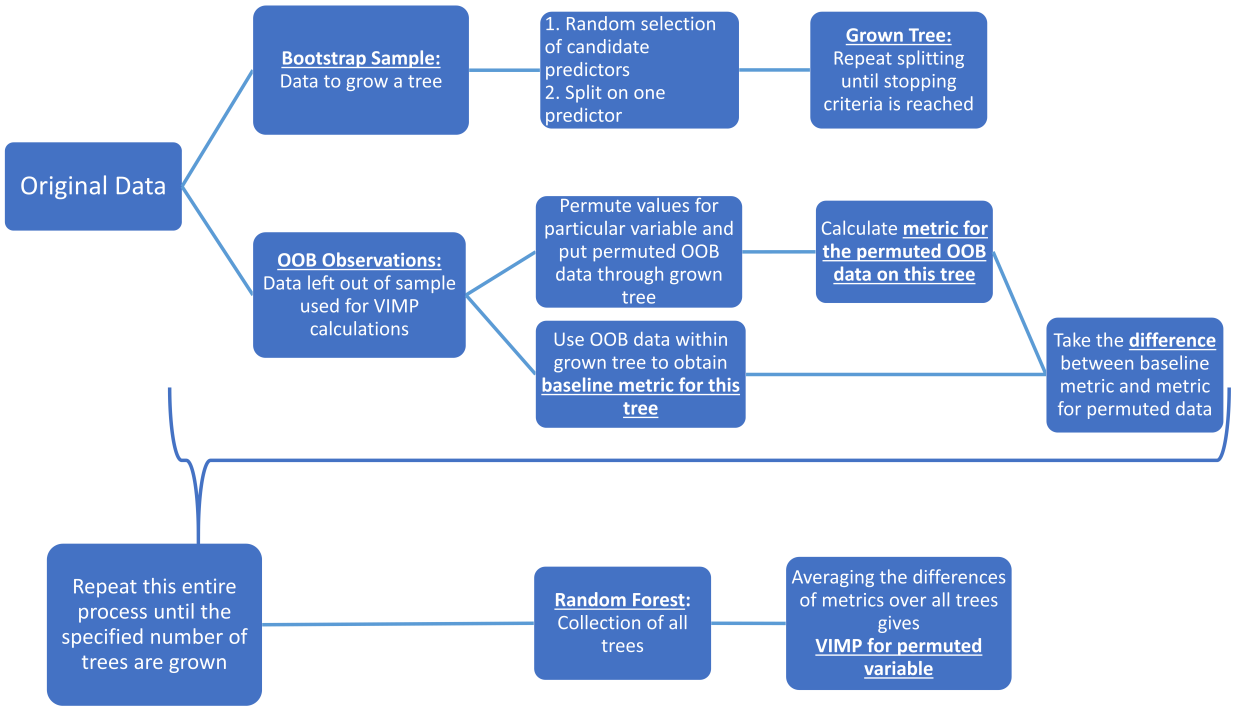

Figure 6: The general process to calculate VIMP measures for RF or CRF using the OOB data.

In order to combat the issues arising from collinearity, the method of conditional random forest (CRF) was proposed. This method is based on an unbiased splitting criteria which is based on conditional hypothesis testing. Essentially, at each split and for each candidate predictor, the association between this predictor and response is tested globally where a p-value is outputted. This p-value represents the probability of obtaining that high of correlation or more of that predictor with the response given the marginal distribution for the response and that of the predictor. Thus, this p-value is conditional and contests issue of correlated predictors being favored over uncorrelated ones each with no direct effect on the response [Boulestix et al., 2012]. For the VIMP measure within CRF, a particular variable X_j is still permuted as in regular RF VIMP calculations, however the values of X_j are only

permuted within particular groups of observations based on a conditioning grid from the predictors, Z, which are correlated with X_j as seen in Table 2 where $x_{\pi_j(1),j}$ indicates the first observation for X_j after permutation and $x_{\pi_{j|Z=a}(1),j}$ represents the first observation for X_j conditioned on the group of values from Z which equal a. This will then preserve the correlation structure between predictors and is similar to the idea of partial correlations from linear regression. When conditioning, the idea is straightforward for categorical variables, i.e. condition on the categories. However, for continuous variables, conditioning is a bit more complicated since these variables need to be discretized. Luckily, the tree built for that particular sample comes in handy and gives a partition of these features which may be used. The set of predictors being conditioned on contains all variables which the particular current variable is correlated with. A lower threshold is given with a default of 0.2. So, all predictors whose correlation with the particular predictor satisfies one minus their p-value greater than the threshold ($1-$p-value>0.2) will be used for the conditioning process. A higher value for the threshold indicates only strongly correlated variables will be used and thus, the method takes less computation time [Strobl et al., 2009]. An example of this method on a subset of the PROVIDE data is shown in Figure 7. These results show how important the mother's health is for their child's height-for-age and gender z score (HAZ) and the change over two years of HAZ. CRF was chosen for this data due to the mix of variables and correlations between predictors [Donowitz et al., 2018] (coauthor). This type of graphic is very attractive to many especially due to its intuitive understanding.

Another aspect for RF or CRF is to make sure the parameter *mtry*, the number of candidate predictors for a split, is tuned before creating and using the final model. This parameter is known to affect the results of variability for conditional VIMP while for RF the correlated variables' VIMP may be overestimated especially when the parameter is small. Additionally, smaller variation is seen with conditional VIMP than with regular permutation based VIMP which may lead to easier identifiability of important variables [Strobl et al., 2008]. Auret and Aldrich explored RF and CRF along with a few other methods and confirmed that VIMP

Table 2: Permutation scheme for the regular marginal VIMP calculation (left) versus the permutation for conditional VIMP (right). The shaded cells indicate where permutation occurs with the different transparencies indicating separate permutations.

Y	X_j	Z
y_1	$x_{\pi_j(1),j}$	z_1
$\vdots$	$\vdots$	z_2
y_n	$x_{\pi_j(n),j}$	$\vdots$

Y	X_j	Z
y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
y_{21}	$x_{\pi_{j\|Z=b}(21),j}$	$z_{21} = b$
$\vdots$	$\vdots$	$\vdots$

measures from CRF avoid the complications brought on by correlation between predictors [Auret and Aldrich, 2011]. Thus, the CRF method proves best for data when the predictors are of multiple types, including continuous and categorical even with various numbers of categories or categorical sizes, and specifically for correlated predictors.

2.3.3 Additional Methods and Settings

A VIMP measure for parametric nonlinear modeling in Table 3, referred to as the Chevan and Sutherland method, is given by averaging over the ordering similar to LMG or PMVD, however the metric used can be a goodness of fit measure. The goodness of fit measure one wishes to use is found for the full model versus the null model, i.e. the difference of the deviances, which is denoted by G_D. The independent contributions, I_j, are calculated as the unweighted average over the different orderings of the order dependent additions of the predictor X_j to the goodness of fit measure G_D. Then the goodness of fit for the full model is the sum of these independent contributions. The overall contribution is the goodness of fit metric for the model with only the particular predictor X_j versus the null model. If the individual contribution I_j is subtracted from the overall contribution value R_j, then the

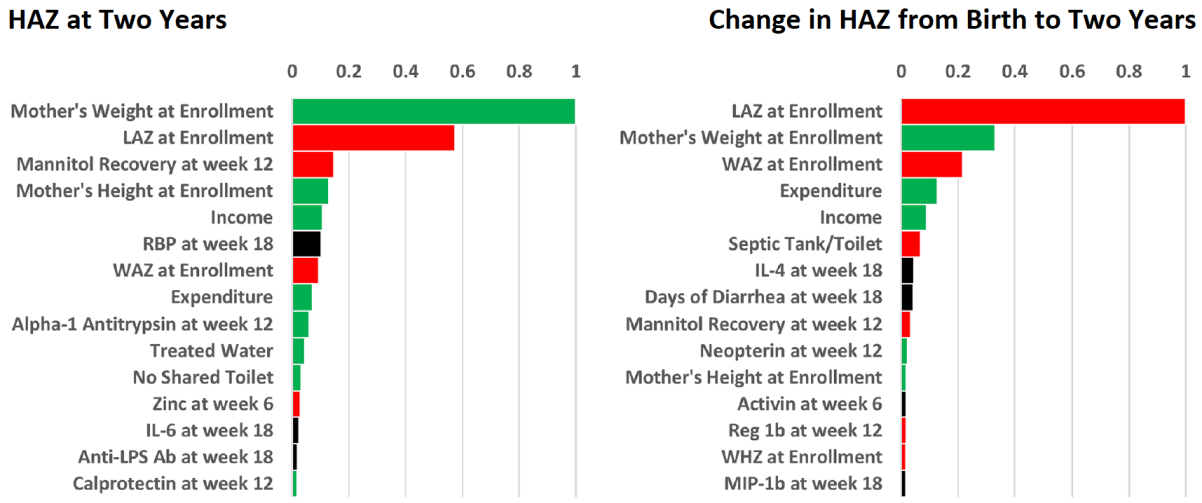

Figure 7: The VIMP measures are from a conditional random forest (CRF) model with the conditional VIMP calculated then scaled for the PROVIDE data. This scaled conditional VIMP is the original conditional VIMP divided by the largest value of conditional VIMP (that of mother's weight at enrollment for HAZ at two years and that of LAZ (HAZ) at enrollment for the infant when predicting the change in HAZ over the two years). Only the top 15 variables are shown [Donowitz et al., 2018].

joint contribution is created as $J_j = R_j - I_j$. The R package which implements the Chevan and Sutherland method which is essentially hierarchical partitioning only works for up to 9 regressors correctly and completely stops for more than 12 predictors [Grömping, 2015].

A common method which is often deemed a 'black-box' are neural networks (NNet) which attempt to mimic the learning pattern of humans' natural biological neural networks. The response is modeled through hidden units (often called the neurons) which are unobservable linear combinations of the original predictors as in Figure 8 for the regression case. These linear combinations are not constrained in any manner. The hidden units are then transformed by a nonlinear (sigmoidal) function. In the logistic case, the hidden units would be

$$h_k(\mathbf{X}) = g\left(\beta_{0k} + \sum_{i=1}^{P} x_j \beta_{jk}\right)$$

where $g(u) = \frac{1}{1+e^{-u}}$ and P indicates the total number of predictors. The β values are similar to that of regression coefficients in that β_{jk} represents the effect for the j^{th} variable on the k^{th}

hidden unit. The model consists of several hidden units and due to the lack of constraints on the linear combinations, it's probable that each coefficient within a particular unit represents some piece of information. After the hidden units have been defined they need to be related to the response which may be completed through another set of linear combinations such as

$$f(\mathbf{X}) = \gamma_0 + \sum_{k=1}^{H} \gamma_k h_k$$

where H is the total number of hidden units. So, for this type of network there are $H(P + 1) + H + 1$ parameters being estimated in total which as one may imagine grows quickly as P increases. These parameters are often optimized via the minimum of the sum of squared residuals when this setting is treated as a nonlinear regression model. However, due to the complexity, the back-propagation algorithm is used to find the optimal values of the parameters. This method is very efficient, but has the caveat of possibly not giving the global solutions. In the case of classification, the last layer (the outcome layer in Figure 8) will have multiple nodes to accommodate the possible categories or classes. Thus, an additional nonlinear transformation will be used for the combination of hidden units. More transformations are needed to make the predictions per class like probabilities (between zero and one and which sum to one) including the softmax transformation. In this case though, the analogous optimization comes from the error across classes and samples as

$$\sum_{l=1}^{C} \sum_{i=1}^{n} (y_{il} - f_{il}^{*}(x))^2$$

where y_{il} is the indicator for a particular class l, C is the total number of classes, and f_{il}^{*} is the model prediction of the lth class and the ith sample after the softmax transformation [Kuhn and Johnson, 2016].

There are several available VIMP measures for this particular method of NNet as outlined in Table 3. Specifically, connection weights, Garson's algorithm, partial derivatives, input perturbation, sensitivity analysis, and many stepwise selection algorithms were explored in

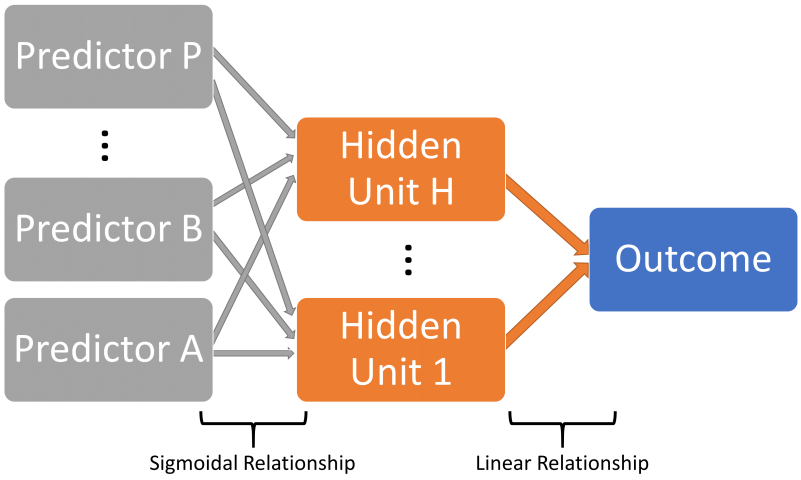

Figure 8: The basic structure of a neural network.

Olden et al. [Olden et al., 2004]. The connection weights refer to "the product of the raw input-hidden and hidden-output connection weights between each input neuron and output neuron" where the sums of these products are calculated across all of the hidden neurons [Olden et al., 2004]. Thus, $\text{VIMP}_{\text{connection weights}} = \sum_{k=1}^{H} \beta_{jk}\gamma_k$. Garson's algorithm involves a partition into aspects associated with each input neuron of the hidden-output connection weights by implementing the absolute values of the connection weights. This method is often the most common within ecological data but had the worst performance within Olden et al. while the connection weights proposed had the top performance [Olden et al., 2004]. Partial derivatives may be calculated from the output of the artificial neural network with respect to the input neurons while input perturbation examines the change in the MSE of the network. The input perturbation adds a specific amount of noise to each input neuron while all other input neurons are held as observed. The change in MSE is then calculated and this demonstrates the relative VIMP for that particular predictor. Sensitivity analysis is where each input predictor across 12 data values are varied while delimiting 11 equal intervals for its whole range and holding other predictors constant at the values of their five-number summaries. Across the five-number summary values, the median prediction is calculated where the relative importance is then the magnitude of its range of the predicted values.

Four stepwise methods were also explored for Olden et al. [Olden et al., 2004]. These include the forward selection and backward selection where the change in MSE was considered as the VIMP measure and the neural networks were rebuilt at each step. The first other stepwise method is very similar to backwards selection, however instead of just removing the input neuron the associated weight is also removed sequentially without rebuilding the network. The change in MSE for each predictor removal is still the VIMP metric. The last stepwise method involves sequentially replacing input neurons with their respective mean value where the VIMP measure is still the change in MSE [Olden et al., 2004]. After simulations, Olden et al. show that the last two stepwise selection methods perform similarly but outperform both forward and backward selection. The results also state that the first metric, connection weights, was able to consistently identify the correct ranking of all variables whereas most other measures could only identify the top few or none [Olden et al., 2004]. All except the stepwise methods have been more recently explored by Oña and Garrido where due to the instability or high variability of these metrics a set of neural networks with the same architecture were used instead of a single neural network [Oña and Garrido, 2014]. The findings show stability of the ranking from all the importance metrics when a set of neural networks are used, but the partial derivatives show the highest variability which leads them to be the least recommended [Oña and Garrido, 2014]. Therefore, the connection weights, Garson's algorithm, input perturbation, and sensitivity analysis all seem to be viable options for NNet.

Another couple ways to calculate variable importance were presented by Parr et al. and are in Table 3 [Parr et al., 2018]. The first method is named drop-column importance. First, a baseline performance measure is obtained from the full model with all predictors, then a predictor is removed entirely and the model is recomputed with a new performance metric calculated and the random seed being controlled (such as if one were using random forest). The VIMP is then measured by the difference between the baseline and the new performance measure. Even though this would give a great estimate of VIMP each time, one can imagine

that with numerous variables or observations, this method may become computationally expensive. However, this method is faster than cross validation for VIMP within RF. Like many VIMP measures, this one is also affected by correlations within the predictors. For example, if a decision tree were built where a duplicate of one predictor was included, the duplicate and the original would be each chosen about 50% of the time gaining equal but low importance per duplicated predictor. Likewise if variables are correlated, the VIMP measure would be shared between these predictors by the amount they are associated. Therefore, the idea that correlated predictors should be assessed together arises. If the correlated variables are permuted together as one feature instead of individually, the correlation structure is not broken and the difference in accuracy or MSE (or other criteria) may be assessed for the set of predictors and become the group's VIMP. The sets of correlated predictors may also overlap since they are being treated as separate features [Parr et al., 2018]. Another downfall for any RF or CRF method is that even though these trees take into account interactions, the VIMP does not. If two predictors that interact with each other have an effect on the response together but no direct effects alone, then they will most likely not receive a high VIMP [Boulestix et al., 2012]. Thus, depending on the structure of the data, one may try the drop-column to create VIMP, but if correlations are present between predictors or if interacting predictors are of interest, permuting groups of variables may be more meaningful.

2.3.4 VIMP Measures Available in R

One large downfall to having numerous different metrics is that no one metric can be applied for various methods. This setback is pronounced within the multitude of methods and calculations of VIMP within R packages. Table 4 shows a list of the several different packages for assorted classification or regression methods (which is by no means an exhaustive list) along with whether or not a model specific VIMP measure exists. The R package 'caret' employs various other packages including all those in Table 4 where these models can be trained if there are tuning parameters. This package also includes some VIMP metrics which

Table 3: VIMP Measures for Nonlinear Models.

Method	**VIMP Measure**	**Type**
Parametric Nonlinear	Chevan and Sutherland: goodness of fit measures	Both
RF - Regression	Permutation based VIMP using the OOB samples	Marg
RF - Classification	Gini impurity	Marg
RF - Classification	Mean decrease in accuracy	Marg
RF - Classification	AUC-based permutation VIMP measure	Marg
CRF	Conditional VIMP	Cond
NNet	Connection Weight: input-hidden and hidden-output	Cond
NNet	Garson's: partitions hidden-output connection weights	Cond
NNet	Partial Derivatives	Cond
NNet	Input perturbation: Change in MSE	Marg
NNet	Profile by Gevrey et al.: Sensitivity analysis	Cond
NNet	Change in MSE for addition or removal of variable	Cond
Any	Drop-column	Marg
Any	Permute groups of predictors	Cond

are model independent. For classification with only two classes, the area under the ROC curve is computed for each predictor and used as the VIMP measure. When there are more than two classes, the area under the ROC curve is still calculated, but now is calculate for each pair of classes. Thus, the VIMP for a specific class is the average of these relevant pairwise areas. In terms of regression, a relative measure of VIMP is calculated through the relationship of each regressor and the response. There are two model fitting techniques, one is simply a linear model with the VIMP metric being the absolute value of the test statistic for the slope. The second is the loess smoother being fitted where the VIMP measure the R-squared value calculated for the smoothed fit vs the fit with only an intercept [Kuhn, 2018]. This review stokes the fire for the need of a more standard VIMP measure which can be calculated across various methods.

Table 4: Available Classification or Regression Methods with Model Specific VIMP Measures.

Method	**R Package**	**VIMP?**
Linear Regression	base R or glm	
Generalized Linear Model	glm	X
Penalized Regression	glmnet	X
Generalized Additive Model	gam	X
Multivariate Adaptive Regression Spline	earth	X
Nonlinear Mixed Effects	nlme	
Classification And Regression Trees (CART)	rpart	X
Bagged AdaBoost	adabag	X
Bagged CART	ipred	X
Bagged Model	caret	
CRF	party	X
RF	ranger	X
RF	Rborist	X
RF	randomForest	X
RF	extraTrees	
Bayesian Additive Regression Trees	bartMachine	X
Naive Bayes	naivebayes or klaR	
C5.0	C5.0	X
Stochastic Gradient Boosting	gbm	X
ROC-Based Classifier	rocc	
Linear Discriminant Analysis	MASS	
Quadratic Discriminant Analysis	MASS	
k-Nearest Neighbors	kknn	
Support Vector Machines	kernlab	
NNet	nnet	X
Model Averaged NNet	nnet	

2.4 Exploring Confidence Interval Calculations for VIMP with Applications

Currently, few have explored the variability of VIMP metrics such as confidence intervals. The LMG and PMVD methods do have such calculations, though PMVD is not accessible for US residents due to a patent [Grömping, 2015]. Also due to the computationally expensive methods that are LMG and PMVD, these results are not further discussed. However, one applicable source is from Ishwaran and Lu for random forest discussed a bit later [Ishwaran and Lu, 2018].

Adding confidence intervals onto ranked graphs of VIMP such as in Figure 9 add greatly to the interpretability of such measures. From this example of the NICU data in Figure 9 one can see that when the bars of the confidence intervals overlap, one predictor is not necessarily more important than the other. From this particular graph, we may say that birth weight and gestational age are the two most important predictors, however, we cannot say that one is more important than the other due to the overlapping intervals.

In order to create confidence intervals for VIMP measures per variable, a form of random sampling such as bootstrapping may be performed. The procedure takes numerous samples for which VIMP measures per variable is calculated. These VIMP values will then collectively allow us to assign confidence intervals for VIMP per predictor. The basic idea for bootstrapping involves taking a random sample of size n from the original data with replacement. This random sampling is then repeated say 1000 times. Then, the VIMP measures' distributions is examined through a histogram and other statistics so the $(1-\alpha/2)100$ and $(\alpha/2)100$ percentiles can be calculated representing a confidence interval for the VIMP measures.

This bootstrapping idea works great within the simpler settings including logistic regression as can be seen by Figure 10 which shows how close the bootstrapped confidence intervals may be to the actual confidence intervals for the coefficients of the NICU variables. How-

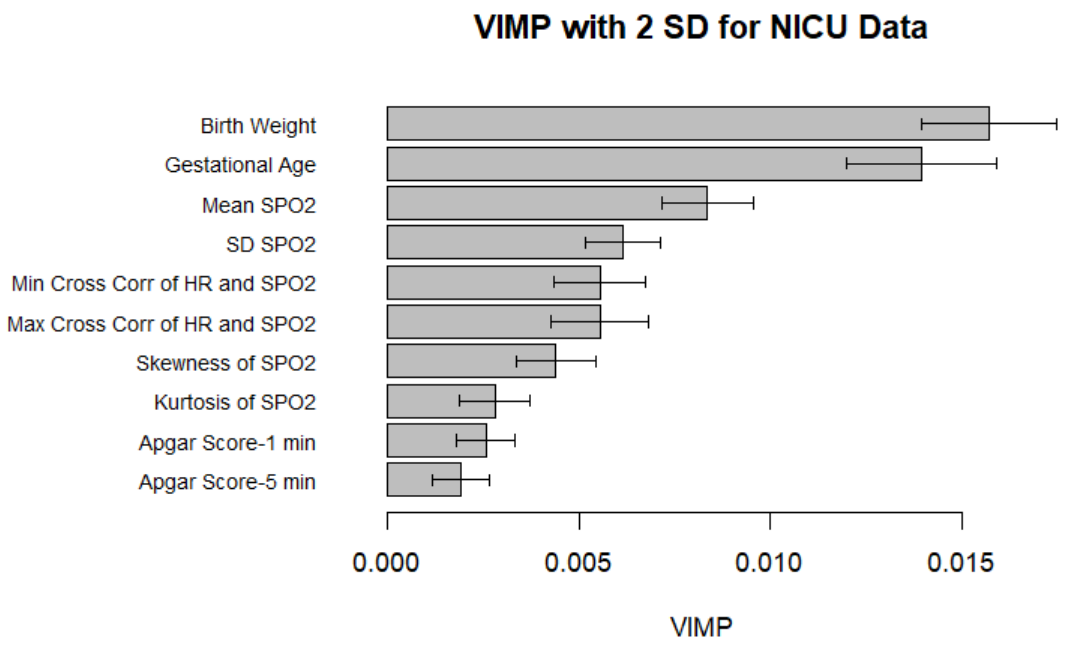

Figure 9: Mean decrease in accuracy VIMP for the top 10 predictors from random forest with 500 trees on the NICU data with intervals showing two standard deviations and the length of the bars being the mean decrease in accuracy over the 500 trees.

ever, when thinking about bootstrapping within the random forest setting, an issue arises when the same subject may be selected more than once, if sampling with replacement, as discussed within Ishwaran and Lu [Ishwaran and Lu, 2018]. The issue is that if a bootstrap sample with replacement is used to create a random forest (RF), the procedure has a chance to select one subject for the growth of a tree while the same subject may also be used in the OOB (out-of-bag) samples which then would not make the OOB set independent of the tree growing set. Thus, Ishwaran and Lu give a few solutions which may be implemented in the *randomForestSRC* R package and *subsample()* R function after building a random forest [Ishwaran and Lu, 2018]. These individuals focus on subsampling techniques which helps approximate the distribution of such statistics and measures.

Their first approach includes bootstrapping, however instead of the typical chance of 0.368 a particular case has of being OOB, the probability a subject has of being truly OOB is only 0.164. This means that only those cases which are actually OOB, those not repeated

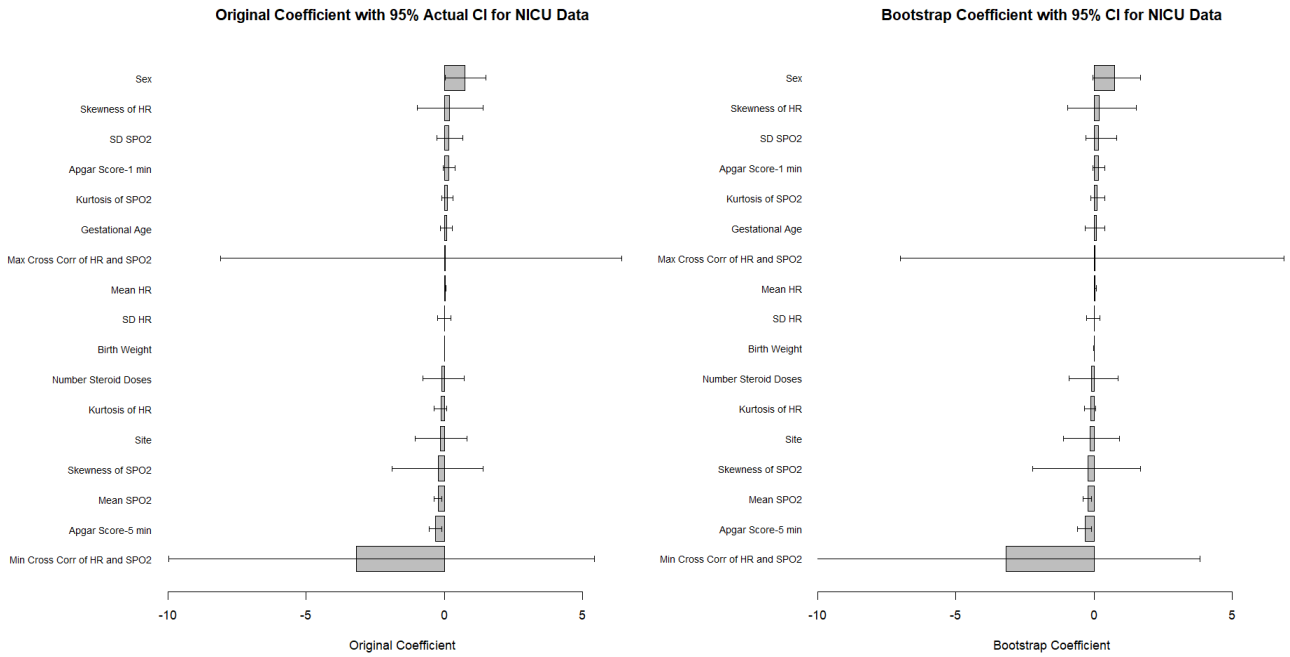

Figure 10: The original confidence intervals for the estimated coefficients (left) and confidence intervals created via bootstrapping (right) in logistic regression for the NICU data.

between the OOB and the tree growing set, are used to estimate the variance of the VIMP measures for RF. To explain why the probability drops to 0.164, let $I_{n,T}^{(j)}$ be the VIMP for the RF and the bootstrap estimator of the variance be $Var(I_{n,T}^{(j)})$ for the j^{th} predictor with T trees. Let $\mathbb{P}_n$ be the empirical measure for $\mathcal{L}$ which is the entire sample of data. Thus, $\mathcal{L}^*$ is the bootstrap sample from $\mathbb{P}_n$. We then must draw a bootstrap sample $\mathcal{L}^*(\Theta^*)$ for the random forest from this bootstrap sample $\mathcal{L}^*$ where Θ^* is the set of growing rules for that particular bootstrap sample (from the other bootstrap sample) making $\mathcal{L}^*(\Theta^*)$ the double bootstrap sample. If a specific case is duplicated, it's not guaranteed that all of these cases will be in the double bootstrap sample or all in the OOB data. To work out the probability, let the number of occurrences for case i in $\mathcal{L}$ (the first bootstrap sample) be denoted by n_i. With

$$P(i \text{ is truly OOB in } \mathcal{L}^*(\Theta^*)) = \sum_{l=1}^{n} P(i \text{ is truly OOB in } \mathcal{L}^*(\Theta^*)|n_i = l)P(n_i = l)$$

we have

$$(n_1, \ldots, n_n) \sim Multinomial(n, (1/n, \ldots, 1/n))$$
$$n_i \sim Binomial(n, 1/n) \asymp Poisson(1)$$

Therefore

$$\begin{aligned} P(i \text{ is truly OOB in } \mathcal{L}^*(\Theta^*)) &= \sum_{l=1}^{n} \left(\frac{n-l}{n}\right)^n P(n_i = l) \\ &\asymp \sum_{l=1}^{n} \left(\frac{n-l}{n}\right)^n \left(\frac{e^{-1}1^l}{l!}\right) \\ &= e^{-1} \sum_{l=1}^{n} \left(1 - \frac{l}{n}\right)^n \frac{1}{l!} \\ &\asymp e^{-1} \sum_{l=1}^{n} \frac{e^{-l}}{l!} \\ &\asymp 0.1635 \end{aligned}$$

Therefore, this first procedure takes a bootstrap sample then passes it to the RF method which takes a bootstrap sample from the bootstrap sample given. The OOB data are defined to be those not within the tree growing data and unique amongst themselves where then the VIMP is calculated for the tree grown. These steps of the RF are repeated several times to grow T trees in which the VIMP will be averaged for the forest. This entire procedure is then repeated for B bootstrap samples and the variance of the VIMP will be estimated [Ishwaran and Lu, 2018]. Outlined, the process is:

1. Draw a bootstrap sample.

2. Implement random forest where a bootstrap sample is drawn and a tree is grown from that sample.

3. The tree VIMP is calculated using the OOB data values that are unique cases.

4. Steps 2 and 3 are repeated T times to grow T trees in the forest.

5. Over the whole forest, the average tree VIMP is calculated which gives the whole forest's VIMP.

6. Steps 1 to 5 are repeated B times so the variance of these averaged VIMP values may be estimated.

Secondly, Ishwaran and Lu use subsampling and the delete-d jackknife approaches since the double bootstrapping approach leads to the OOB set being much smaller than normal [Ishwaran and Lu, 2018]. Also since the 0.164 (double) bootstrapping method can become computationally expensive fairly quickly as the sample size increases, the subsampling will reduce the computation time since it is more efficient. This method uses small but iid subsets of the data over which VIMP is calculated. These samples are selected without replacement which also negates the complication of ties in the OOB set and the subsample used to grow the tree. The delete-d estimator works with subsets of data with size $r = n - d$ and can be related to the subsampling estimator as the bias corrected version. So, the subsampling method with b being the size of the subset is as such:

1. Draw a subsample set of size b.

2. Calculate the forest VIMP using the subsample set.

3. Repeat steps 1 and 2 B times to estimate the variance of the forest VIMP values.

The delete-d jackknife method where $d = n - b$ is the subsampling method above but the variance estimator is replaced with the bias corrected version.

In a regression setting simulated by Ishwaran and Lu, it was found that the bias of the subsampling estimator was higher for predictors' with larger VIMP (the more important variables), specifically underestimating the VIMP [Ishwaran and Lu, 2018]. The delete-d improved this bias for the larger VIMP valued predictors, however the 0.164 bootstrap method outperformed the others. The downfall to the double bootstrapping method is

the computational cost. Thus, it is recommended that the delete-d estimator should be used when bias is an issue or that the subsampling rate could be increased to improve the subsampling estimator which may lead to the subsampling methods outperforming the double bootstrapping method [Ishwaran and Lu, 2018].

For the specific confidence intervals of these methods, nonparametric and parametric confidence intervals may be calculated although normality may be justified and thus the parametric confidence intervals are more stable. These procedures have been shown to produce too long of intervals when the VIMP is small which is not necessarily a problem since we would rather overestimate than underestimate. The subsampling procedure creates intervals too short when the VIMP is large due to the underestimation of the variance previously stated. This issue of underestimation may be improved by increasing the subsampling rate which then makes the subsampling method generally better than the delete-d estimator [Ishwaran and Lu, 2018]. These methods for confidence intervals were calculated with our data for comparison in 2.4.1 and 2.4.2.

Since bootstrapping gives rise to issues and becomes computationally expensive with large n, we explored VIMP measures in different manners. One thought was to explore how the VIMP acts as the number of bootstrap samples increases for logistic regression and as the number of trees increases for RF. Due to the complex nature of RF, the number of trees increasing was approached in two different manners. One strictly included building random forests with increasing numbers of trees while the other included building numerous single tree forests. These methods were applied to the NICU and PROVIDE data in 2.4.1 and 2.4.2 which encompasses the classification and regression settings respectively. Additionally, since the VIMP per tree can be accessed, another method is to create a RF trained for optimal prediction with the number of trees chosen as such and then bootstrap these per tree VIMP measures to approximate the variation for the specific trained RF. This was also implemented in 2.4.1 and 2.4.2 for both the NICU and PROVIDE data respectively.

2.4.1 VIMP Confidence Intervals for the NICU Data

In the case of the NICU data, we are trying to predict or classify mortality. The simplest most used method for this type of data is often logistic regression especially since the number of predictors is fairly low. If one is interested in VIMP in this setting, it is common to use the estimated coefficients, test statistics, or the p-values which as previously described all have strengths and weaknesses. Since the p-value and test statistic will give similar results and the estimated coefficients need to be standardized, we choose the test statistic. However, since we are generally only interested in the magnitude of the test statistic, we will take the absolute value of the test statistic as our measure of VIMP. The square of the test statistic was also explored as the VIMP metric in which the same or very similar results were given and the results have thus been omitted. In order to explore when the mean absolute test statistic over the number of bootstrap samples converges, 1000 bootstrap samples in total were taken. For each bootstrap sample, logistic regression was completed with the test statistic being stored. So, per increase of the number of bootstrap samples, the mean and mean rank of the absolute test statistics were computed to create the first and last panel respectively in Figure 11. The standard deviation of the VIMP values per increase in the number of bootstrap samples was created as in the middle panel of Figure 11. These graphs, which were truncated at 300 bootstrap samples due to lack of change after this value, show how much variation there is within the VIMP over the number of bootstrap samples, especially when the number of samples is small. These results also show that logistic regression in and of itself may lead to variable VIMP results since the standard deviations converge to various values for the different predictors. However, the bootstrapping method is used to give a final result of VIMP intervals within the logistic regression setting for the NICU data for comparison in Figure 20. Additionally, the percentage of times the absolute value of the test statistic obtained the correct rank of the predictors (correct rank was set as the rank of the original test statistics) was assessed per increase in the number of bootstrap samples.

Figure 12 states these results in which it seems increasing the number of bootstrap samples does not allow for better ranking of these predictors.

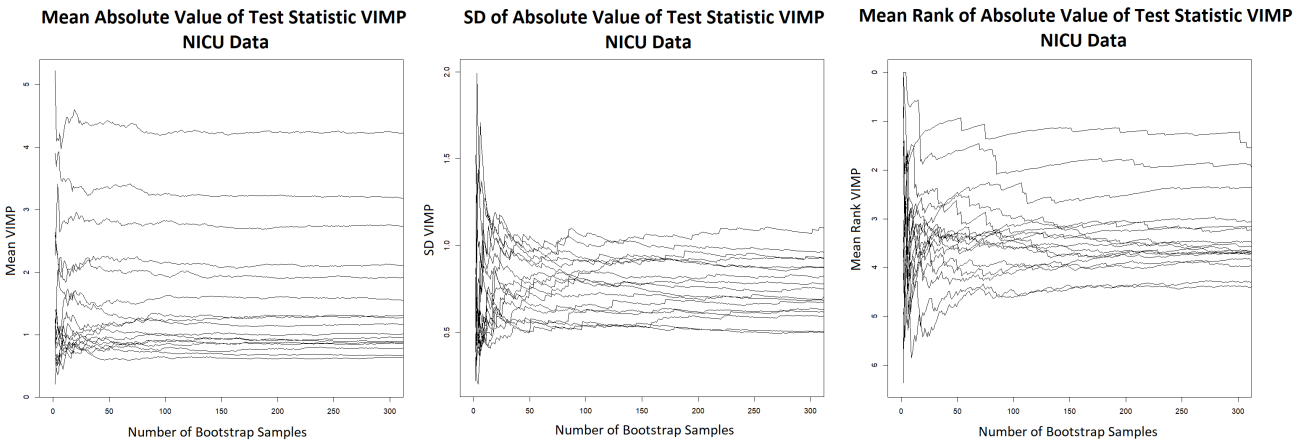

Figure 11: Mean (left), standard deviation (middle), and mean rank (right) of the absolute value of the test statistic over increasing the number of bootstrap samples up to 300 for the NICU data.

The question of how many trees it takes for the VIMP to become stable within the random forest (RF) setting needs to be answered in order for confidence intervals to be explored. A simple approach may be to calculate several single tree forests which one may believe to be similar to a large forest of trees. However, the results defy this first intuition about how the VIMP behaves. Using the *randomForest* R package, 50k single tree random forests were built with the default *mtry* parameter (number of variables to randomly sample during each split). Each single treed forest's Gini index VIMP and mean decrease in accuracy VIMP was calculated. Then over all the 50k forests, the mean and standard deviation of the VIMP values were computed per predictor. These values were used to create intervals (plus and minus one standard deviation) about the mean as shown in Figure 13 and Figure 14 for the Gini VIMP and the permutation based mean decrease in accuracy VIMP respectively. One may note the large amount of overlap between the predictors' intervals and thus the substantial amount of variation amongst the single tree forests' VIMP values.

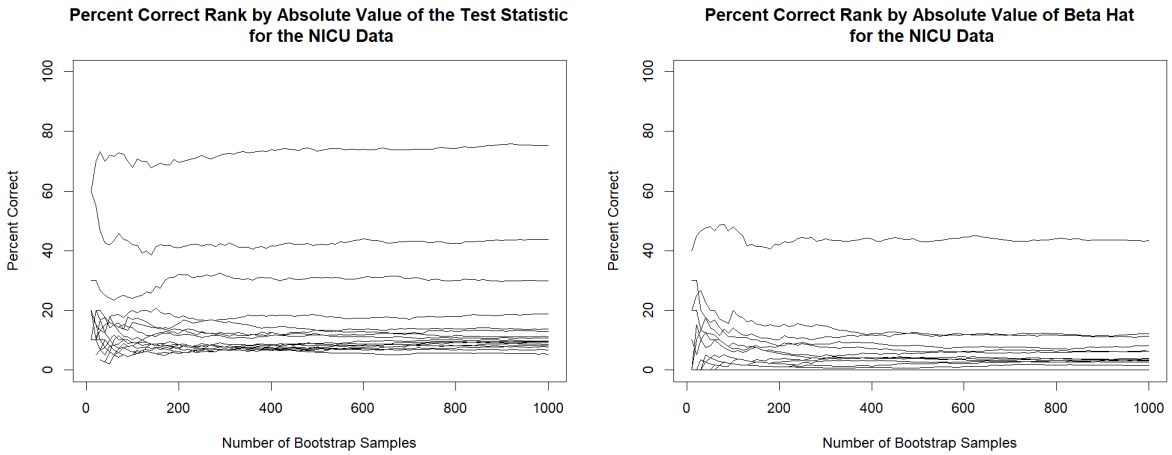

Figure 12: Percent correct rank of the absolute value of the test statistic (left) and absolute value of the estimated coefficient (right) over increasing the number of bootstrap samples up to 1000. Correct rank is considered the rank of the original absolute test statistic with the full NICU data.

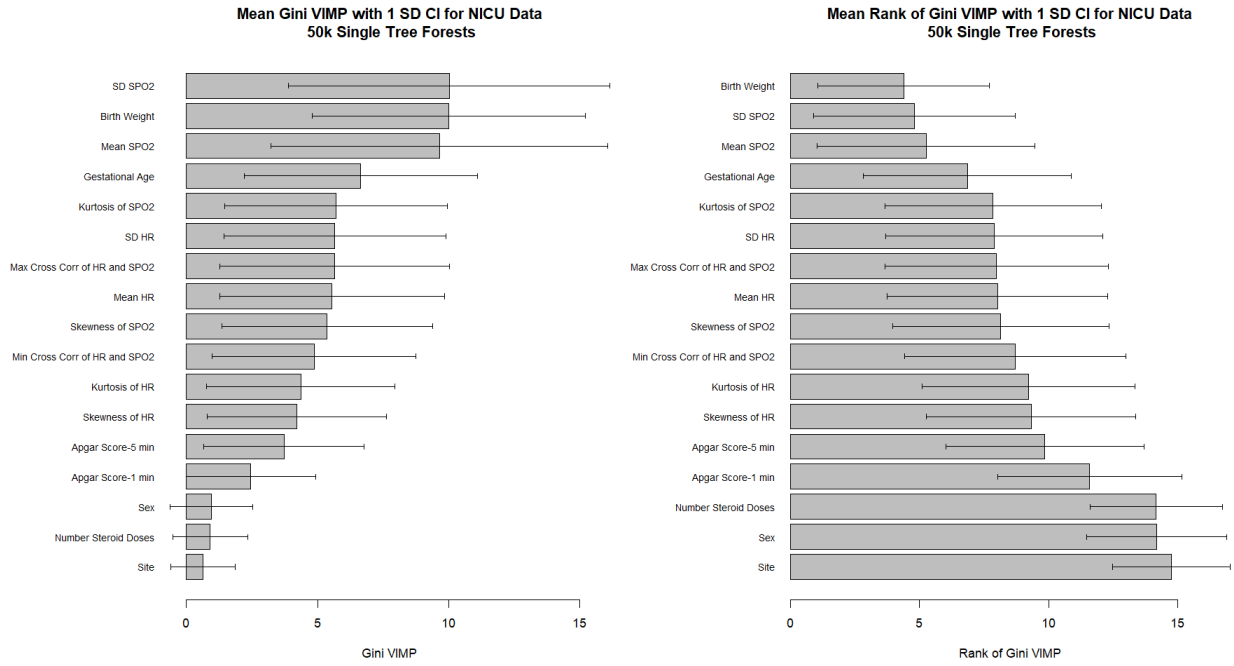

Figure 13: Gini Index VIMP and rank of Gini index VIMP average with one standard deviation for 50k single tree forests with NICU data.

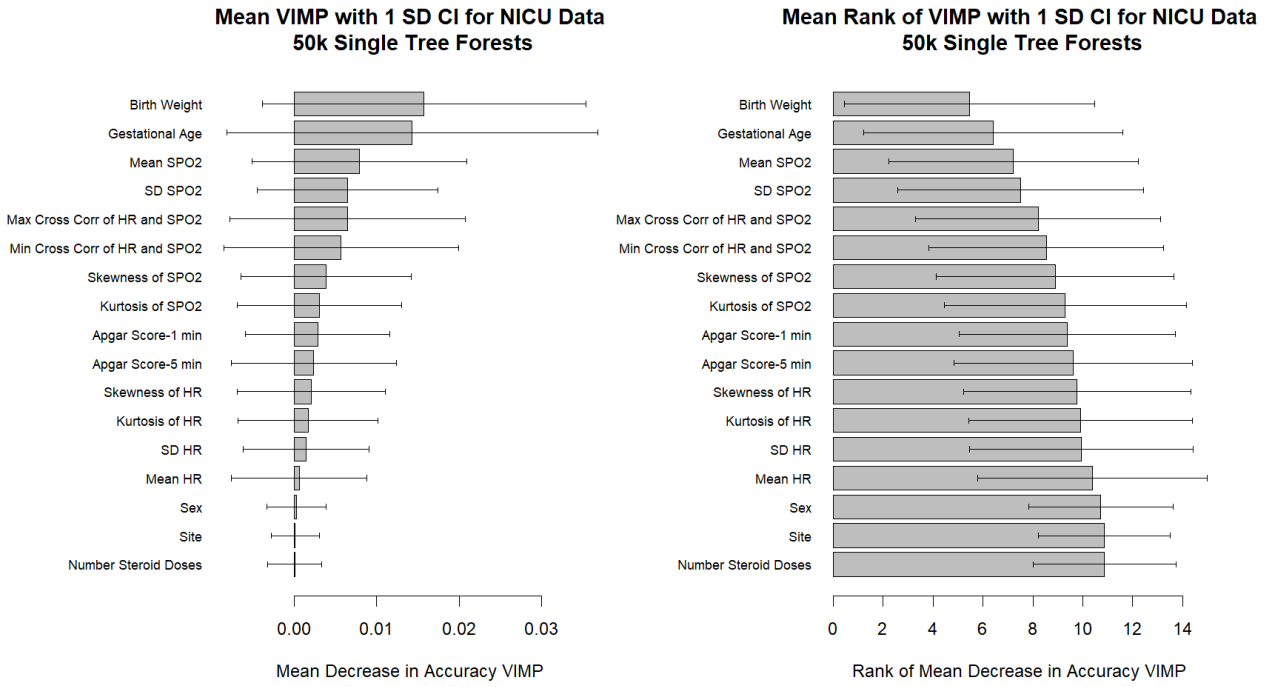

Figure 14: Mean decrease in accuracy VIMP and rank with one standard deviation for 50k single tree forests with NICU data.

As a next step, instead of creating single tree forests, the number of trees in the RFs was increased. In particular, 100 RFs were created per value of the number of trees in the forests such that the average of the mean decrease in accuracy, the average standard deviation of the mean decrease in accuracy, and the average Gini index over the 100 forests per number of trees could be calculated. In addition, the mean ranks of the two VIMP measures were also calculated per number of trees. For clarification, per each value of the number of trees and each random forest, there is a mean, standard deviation, and rank for the decrease in accuracy per each predictor along with the Gini index and its rank per predictor. Then, since there are 100 RFs per number of trees, the average of these measures are calculated. Thus, Figure 15 shows how nicely the mean decrease in accuracy and its rank stabilizes along with how the standard deviations will go to zero as the number of trees go to infinity. This effect is due to the fact that this is the standard deviation of the decrease in accuracy over the trees in a forest. So, as the number of trees increases, the more information there is and thus as the VIMP values stabilize, there is no longer much variation (both the mean and standard deviation involve a n in the denominator; basic statistical properties thus apply). For the VIMP measures and their ranks in Figures 15 and 16, we can see that even at the default of 500 trees in a forest, the results are already fairly stable. This is good news for first time users with a simpler data set! It must also be noted that if the goal is to predict, then less trees are needed for good results. However, if the goal is to obtain a stable ordering of VIMP, increasing the number of trees for the forest until stability of the ranks is achieved such as what was completed here should be the standard. This is especially true if the correct order of the variables is needed in contrast to if just the top few important variables are needed regardless of their order. Similar to the logistic regression case, the percentage of times the correct rank was achieved per number of trees over all the 100 random forests was calculated. The correct rank though was considered as the mean rankings from the 100 random forests with 50k trees. Here, the percentage of correct ranks improves as the number of trees in the forest increases. This solidifies the statement of increasing the number of trees in order to

obtain the correct ranking of the predictors, especially in the case of having a smaller set of predictors which are mostly continuous.

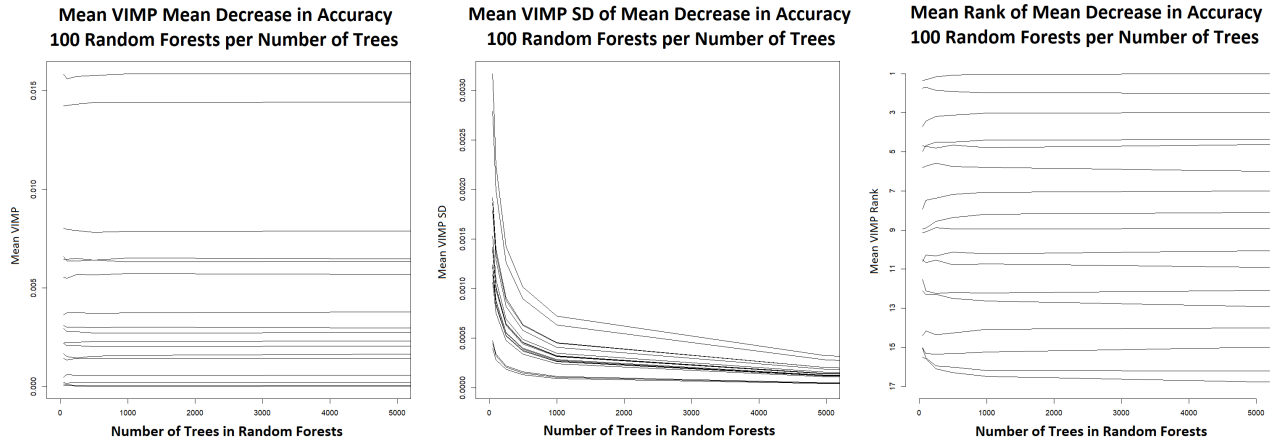

Figure 15: Average VIMP (left), average VIMP standard deviation (middle), and average VIMP rank (right) of VIMP for 100 random forests with mean decrease in accuracy VIMP over increasing number of trees for the NICU data.

One may also explore the individual predictors' values and distributions as for birth weight in Figure 18 showing how the VIMP metrics and ranks for 100 random forests per number of trees converges as the number of trees increases. In the top left panel, each boxplot represents the distribution of the mean decrease in accuracy for the variable birth weight over the 100 forests (there are 100 values per boxplot, each value coming from one random forest's VIMP measure for birth weight). The mean decrease in accuracy and its standard deviation along with the ranks all do well at the default of 500, but improvements may of course be made if one is particularly worried about having the variables in the correct order of importance. With the Gini index, we can see that the rank of this measure is a bit more variable, even with a monstrous amount of trees! This may speak towards the known bias with this metric especially since we have continuous variables with differing distributions and with some predictors having more possible splitting values.

Comparing logistic regression VIMP to RF VIMP measures for the NICU data, the VIMP measures from RF tend to converge faster and better than that of logistic regression. This is

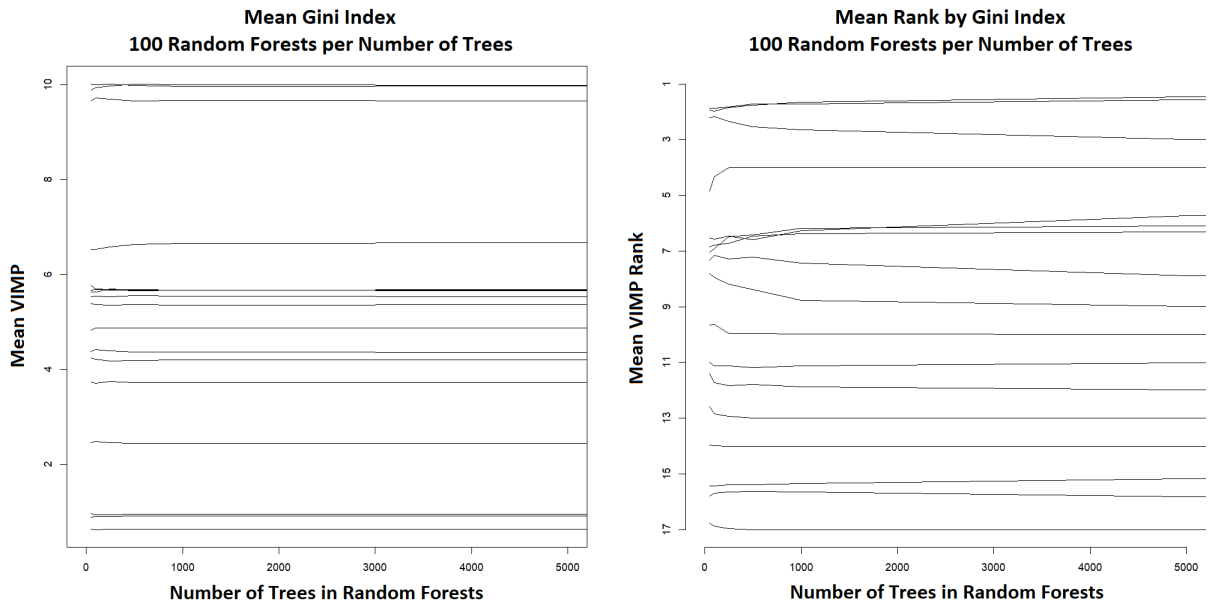

Figure 16: Mean VIMP (left) and mean VIMP rank (right) of the Gini index for 100 random forests over increasing number of trees for the NICU data.

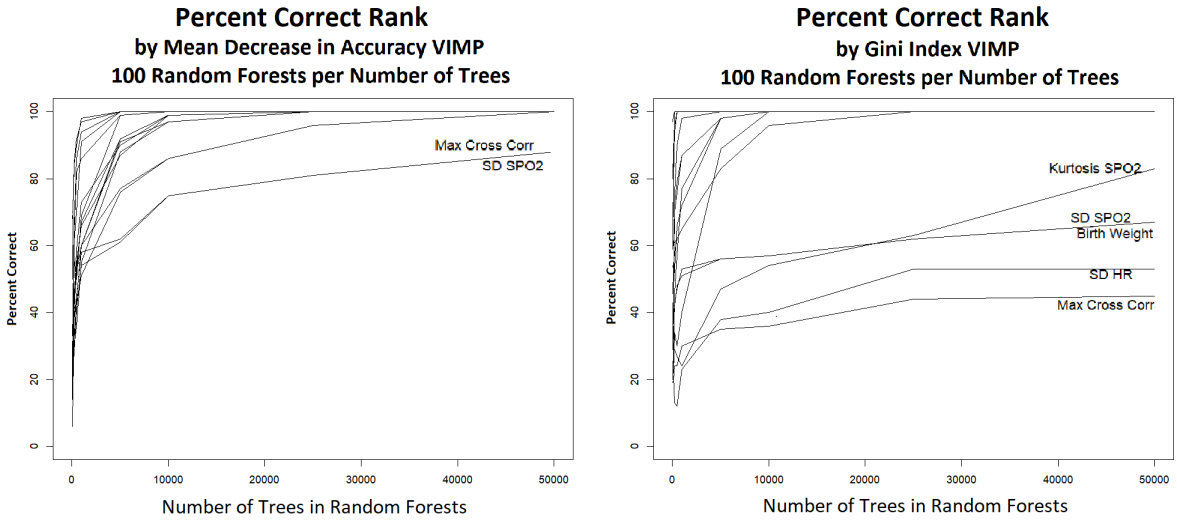

Figure 17: Percent correct rank by the mean decrease in accuracy VIMP with 100 random forests per number of trees for the NICU variables. Correct rank was considered the mean rank for the 100 forests with 50k trees.

mainly due to increasing the number of trees which also diminishes the standard deviation of the decrease in accuracy. The Gini index though is not an average but also performs better when more information is gained through increasing the number of trees in a RF model.

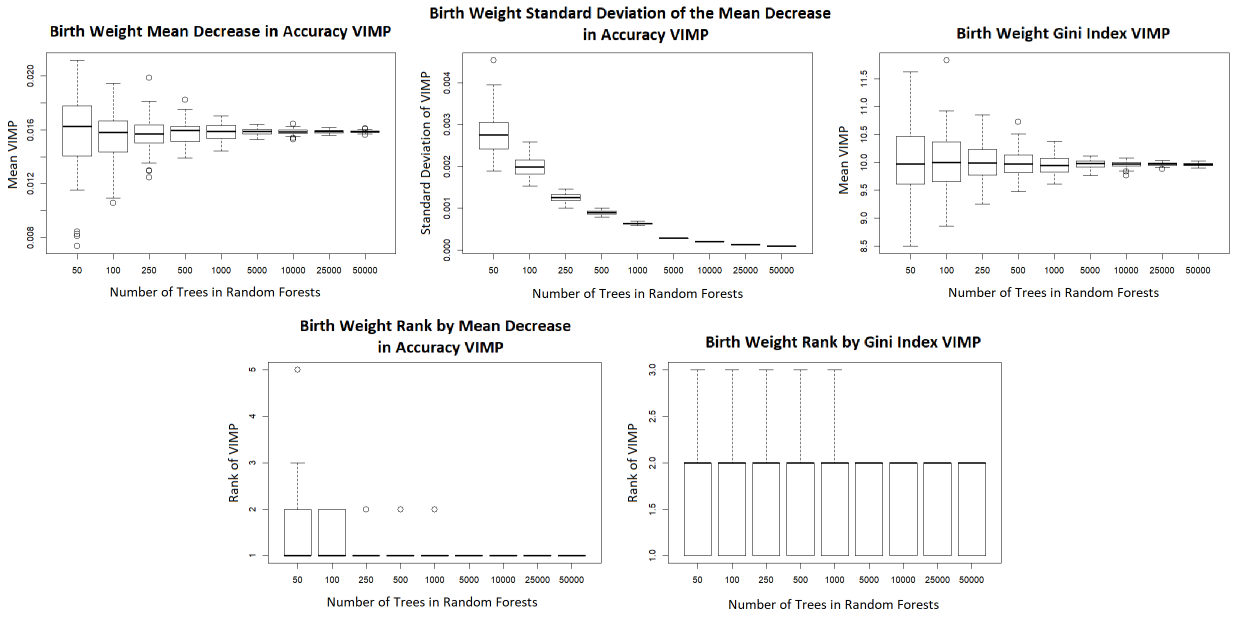

Figure 18: The VIMP (mean decrease in accuracy and Gini index) and rank of the VIMP for birth weight predicting mortality in the NICU data over 100 random forests per number of trees.

Thus, RF methods would be preferred over logistic regression, especially when it comes to variation of VIMP measures and the ranking of the predictors.

Since the RF VIMP should be preferred over logistic regression, one could explore the number of trees needed in order for results to become stable within a certain data set. With the NICU data, we have seen that the VIMP estimates are not bad even at 500 trees, however if interested in more stable results, increasing the number of trees is best. We can see the effect the number of trees has on the interpretation when the mean VIMP is plotted as the length of the bars and the intervals are plus and minus two standard deviations from that mean as in Figure 19. These graphs show that even at 500 trees, the top two predictors are birth weight and the gestational age while there is still much overlap between the rest of the predictors' intervals. At 5k trees, we see less overlap and the order is the same as in the 50k trees results. This leads to say that birth weight and gestational age are still the top two important variables, but we still cannot say which is more important than the other. Since

there is less overlap for the rest of the variables with 5k trees than in the 500 tree case, more solid interpretations and ranks can be made. At 50k trees, we see very tight margins for the intervals although this many trees can take quite a bit longer than a forest with only 1k or even 5k trees making it seem as though 50k trees is overkill. From this large number of trees though, we could now say that birth weight is more important than gestational age, however the max cross correlation between HR and SPO_2 and the standard deviation of SPO_2 along with a few other pairs are still not separable in terms of the VIMP intervals.

For comparison, the bootstrapped confidence intervals for the absolute value of the test statistic and the rank of those within the logistic regression setting have been calculated over the 1000 samples. Figure 20 shows these intervals and the large variation that goes with. We definitely cannot state which variables are more important than others due to the amount of overlap of the intervals between predictors. Again, this showcases the high variability involved with the logistic VIMP leading again to the preference of RF over logistic regression.

The published article for the outcome of mortality in Figure 3 chose birth weight, Apgar at 5 minutes, and sex as the most important from the clinical variables and mean $[SPO]_2$, mean heart rate, and kurtosis of SPO_2 as the top pulse oximetry variables [Sullivan et al., 2018]. Using a large RF to obtain stable rankings leads to birth weight, gestational age, mean SPO_2, the max cross correlation between heart rate and SPO_2, the standard deviation of SPO_2, and the min cross correlation between heart rate and SPO_2 being the top six important variables shown in Figure 19 for the RF with 5k or 50k trees. Thus, the RF does not exactly agree with the logistic regression implemented using the p-values of the test statistics as the VIMP. However, when the bootstrapped absolute value of the test statistic VIMP was implemented, all but the kurtosis of SPO_2 was ranked in the top six as in Figure 20. Even though the results from the bootstrap samples and the published article mostly agree, there is great variability as shown in this research for the logistic regression method leading to

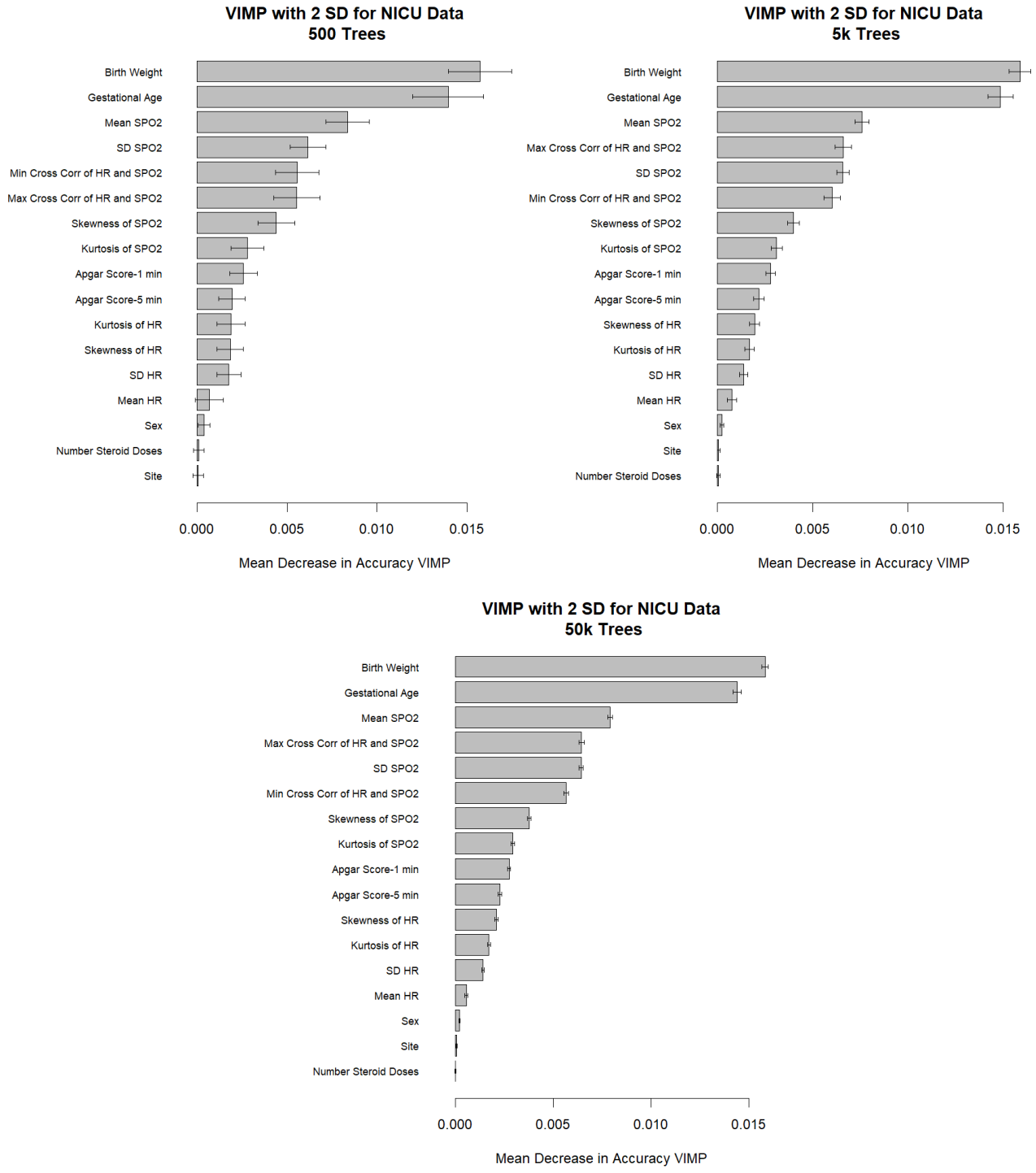

Figure 19: Mean decrease in accuracy with two standard deviations over a 500 tree random forest (top left), a 5k tree random forest (top right), and a 50k tree random forest (bottom) for the NICU data.

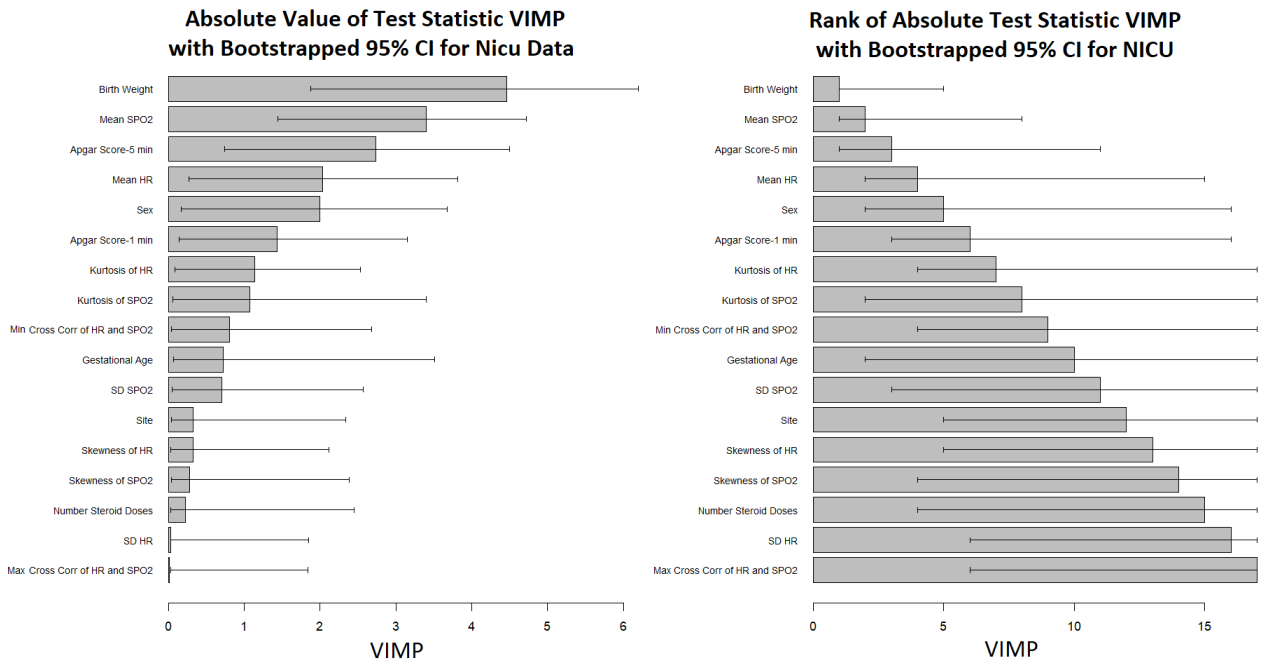

Figure 20: Absolute value of test statistic VIMP (left) and rank of VIMP (right) with bootstrapped confidence intervals (CI) from logistic regression for the NICU data.

suggest the RF method with a large amount of trees should be used to obtain the variables' stable rankings.

One may wish to build confidence intervals for their particular forest which is used for prediction. First, a RF must be trained to select the appropriate number of trees to use for optimal prediction performance. Then, the VIMP of each tree must be extracted to gain the set of per tree VIMP measures which will be used to create the confidence intervals via bootstrapping. The process is:

1. Train a random forest for the optimal prediction to find the number of trees.

2. Create the random forest model and store the per tree VIMP values per predictor.

3. Using the per tree VIMP values, take a random sample with replacement and calculate the mean VIMP per predictor.

4. Repeat step 3 several times, say 1000 times.

5. Take the 2.5^{th} and 97.5^{th} percentiles of these means to create 95% confidence intervals per predictor for the VIMP.

This thus gives confidence intervals about the original VIMP values which for the permutation based VIMP are averages over the entire forest. Additionally, one may repeat this process but with the ranks of the VIMP instead making the results more comparable to other methods. Figure 21 exemplifies the results for the NICU data and our method of VIMP confidence intervals. These results state again that the birth weight and gestational age are the top two important predictors in which their ranks' confidence intervals fully overlap showing that they may not be separated in terms of ordering. We can also state that the rest of the predictors are harder to order with much of their intervals overlapping, however it seems much easier to examine the ordering via the ranks in the right panel of Figure 21.

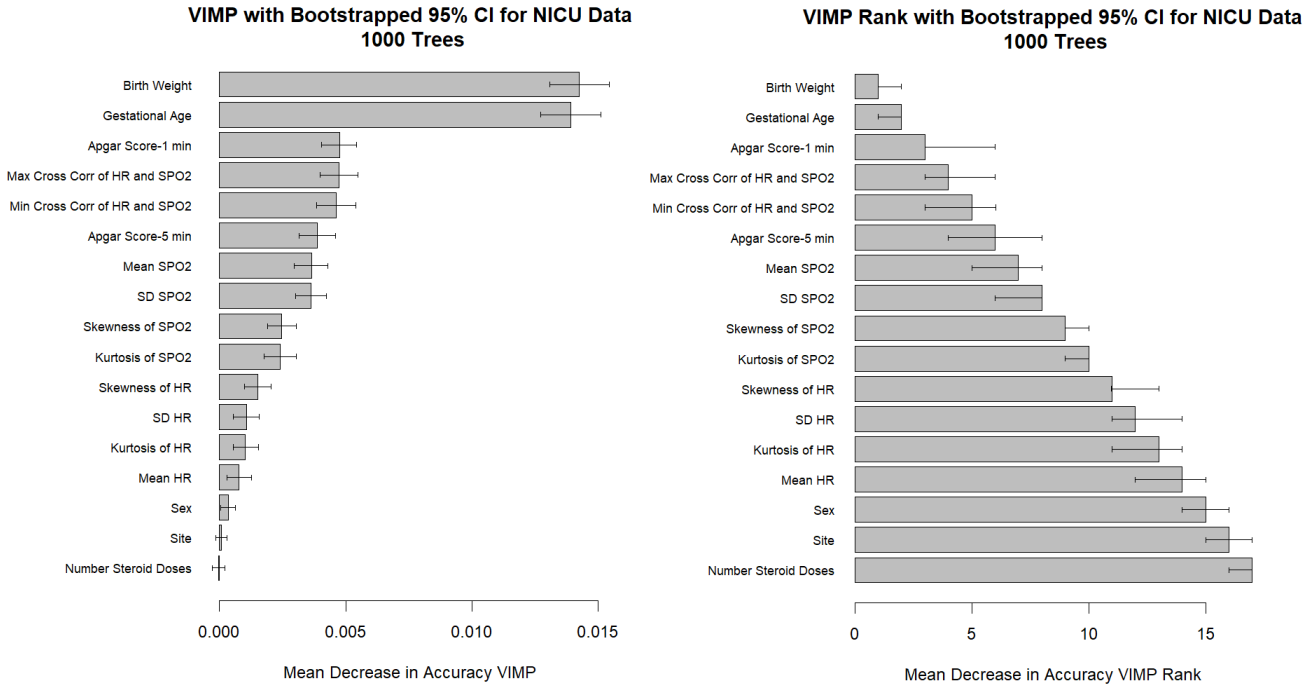

Figure 21: Mean decrease in accuracy VIMP (left) and rank of VIMP (right) with bootstrapped 95% confidence intervals (CI) from the trained random forest's per tree VIMP values for the NICU data.

For comparison when using a trained RF, the methods described and developed by Ishwaran and Lu were used on the NICU data [Ishwaran and Lu, 2018]. A RF model was trained as would be usual for prediction and used then for these calculations. The subsampling, delete-d, and double bootstrap methods all with the parametric confidence intervals were plotted in Figure 22. These results show that the variance for the mean decrease in accuracy VIMP measures in the double bootstrap procedure is largest of these three methods. It also seems that the variance for the delete-d jackknife is a bit larger than the regular subsampling method's variance, although if one is only interested in which variables are important, then the subsampling and delete-d methods agree. These methods also agree with previous results in that the top two predictors should be birth weight and gestational age along with that one is not more important than the other. In contrast to our RF results for the NICU data, these methods would rank the predictors differently even though their intervals overlap and thus one is not truly more important than another. Therefore, it seems that we can only state that the top two important variables are definitely birth weight and gestational age, however, these methods are not ideal for ranking the predictors. These methods from Ishwaran and Lu versus our bootstrapping of per tree VIMP measures have similar computation times and use the same R package, however Ishwaran and Lu's methods do not allow for other plotting methods, manipulation of the resulting values, or confidence intervals for the ranks of the VIMP. So, without the flexibility of working with the results from Ishwaran and Lu's methods' outputs, one cannot create their own graphics or have directly comparable interpretations in terms of the ranks for VIMP.

2.4.2 VIMP Confidence Intervals for the PROVIDE Data

The convergence of the mean percent increase in MSE for the PROVIDE data was also explored such that a number of trees for RF with stable VIMP values and rankings could be achieved. Figure 23 and top left panel shows the top few important variables clearly while the rest are clumped near the lower spectrum of VIMP. This bottom panel in the figure

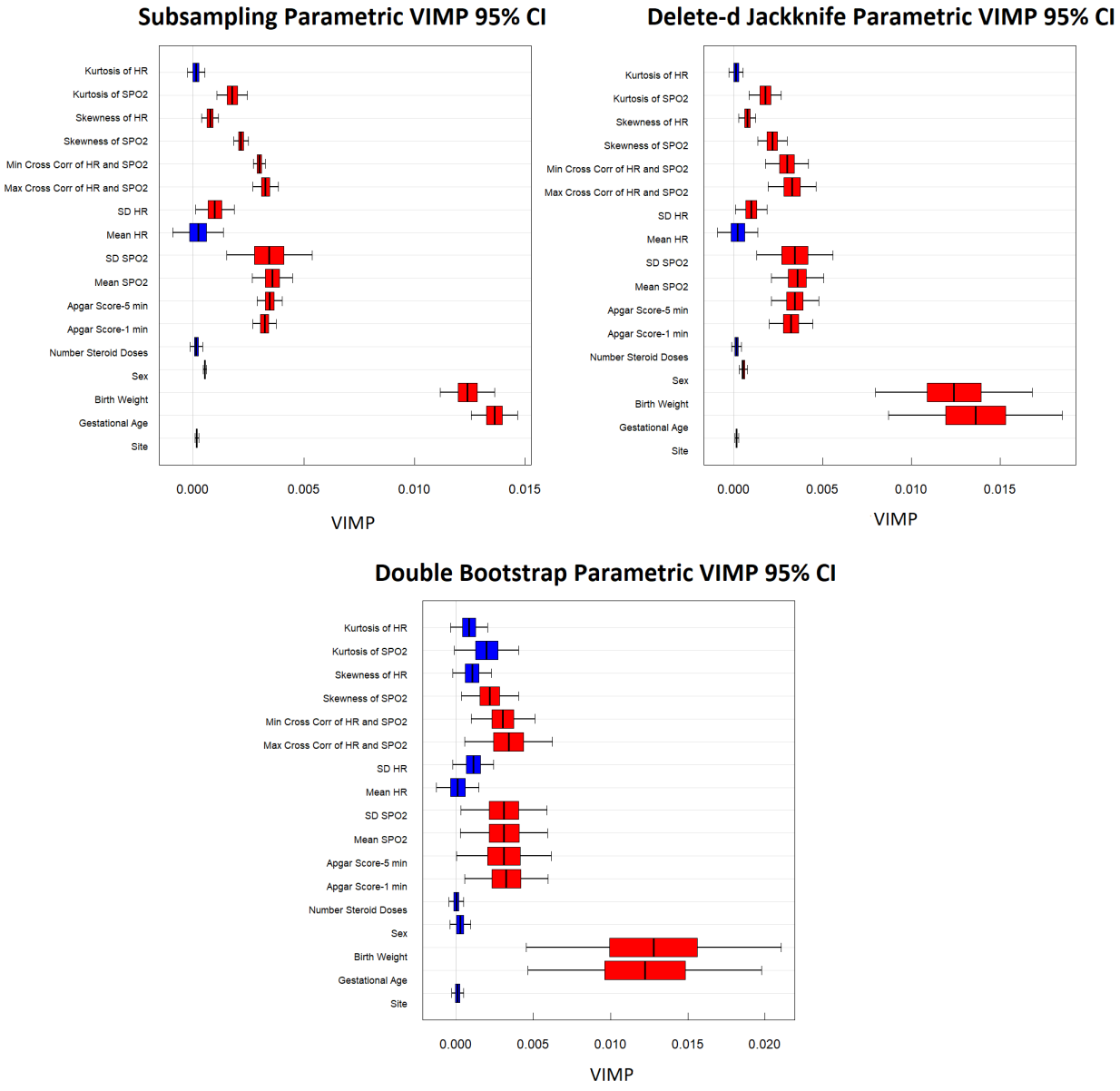

Figure 22: Subsampling parametric (top left), delete-d jackknife parametric (top right), and double bootstrap parametric (bottom) confidence intervals with methods from article for NICU Data [Ishwaran and Lu, 2018].

also shows how the top important predictors have stable ranks, but the rest are harder to separate but do eventually have stable orderings. In general though, the results of VIMP convergence are similar as for the NICU data in that the results become stable fairly quickly and the standard deviations will eventually go to zero as the number of trees increases. The percentage of correct ranks per number of trees out of the 100 random forests was also calculated with the correct ranking being the mean ranking over the 100 RFs with 50k trees. Contrary to the NICU data, the results for the PROVIDE data show the difficulty of getting the correct rankings and that the number of trees could be increased even further to obtain even more stable rankings for the rest of the predictors. These results show the complexity of this data especially since these results used the mean increase in MSE and not the conditional VIMP which would more appropriately consider the relationships between the predictors. Also, due to computation time and the inability to access the per tree VIMP, conditional random forests and the conditional VIMP was not computed.

Using these results though with regular RF, Figure 25 shows RFs with 500 (top left), 5k (top right), and 50k (bottom) trees. Mother's weight can be deemed the most important variable for the 5k and 50k tree forests, but for the 500 tree forest mother's weight overlaps with the infant's HAZ (LAZ) at birth predictor stating that we could not say which is more important. While the 5k and 50k tree forests agree on the top five predictors, the rest are not in the exact same order showing that for the 5k forest one variable (after the fifth position) is not necessarily more important than another, especially since the intervals are overlapping. Thus, we cannot rightfully rank the predictors after the top 5. These differences in the PROVIDE results from the NICU results come from the PROVIDE data having a complex structure with more and various types of predictors than the NICU data. This complex structure thus leads to more variability within the VIMP, specifically in terms of the rankings.

Mother's weight, HAZ (LAZ) at birth, mannitol recovery at week 12, mother's height, and income were ranked as the top 5 predictors for predicting HAZ at two years in a previ-

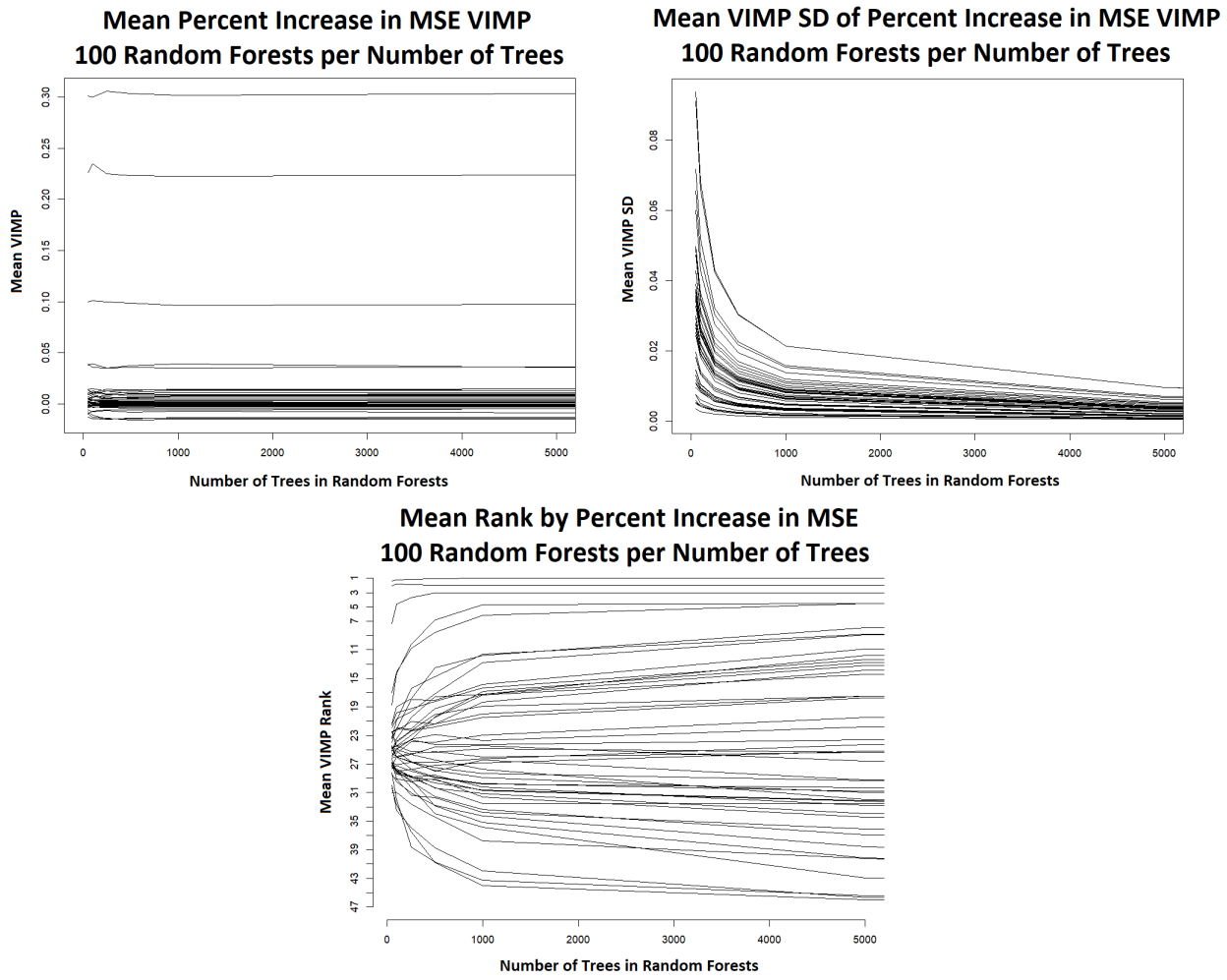

Figure 23: Mean (top left), standard deviation (top right), and mean rank (bottom) of mean percent increase in MSE over 100 random forests per number of trees for the PROVIDE data.

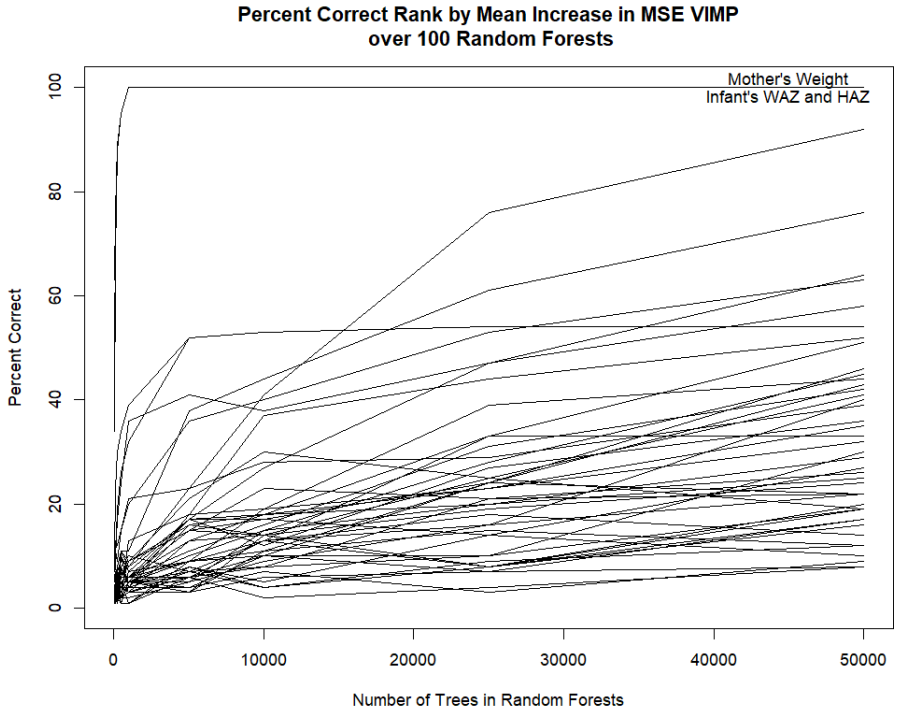

Figure 24: Percent correct rank by the mean percent increase in MSE VIMP with 100 random forests per number of trees for the PROVIDE variables. Correct rank was considered the mean rank for the 100 50k tree forests.

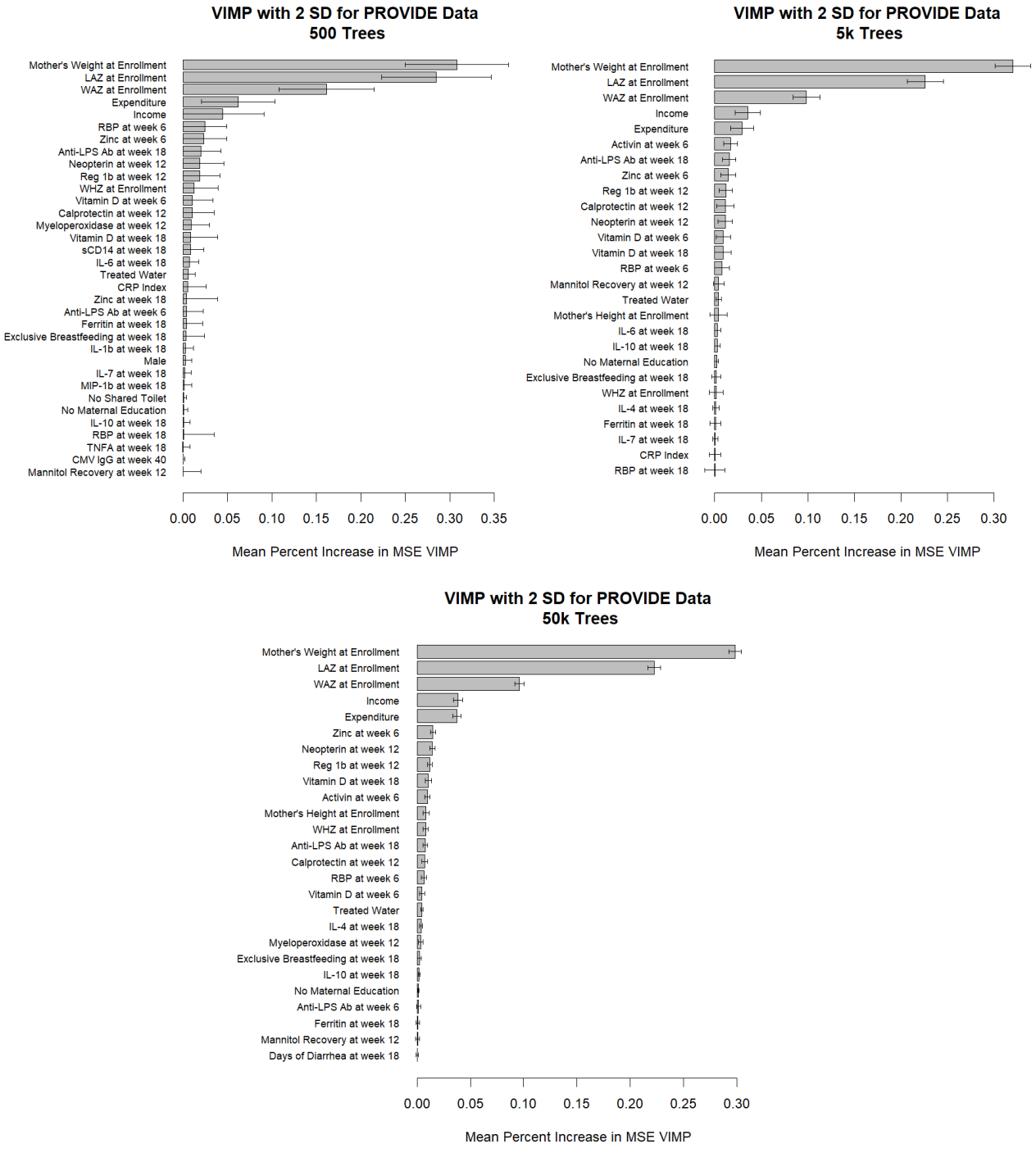

Figure 25: Mean percent increase in MSE with two standard deviations over a 500 tree random forest (top left), a 5k tree random forest (top right), and a 50k tree random forest (bottom) for the PROVIDE data.

ously published article using conditional RF and conditional VIMP [Donowitz et al., 2018] (coauthor). This research here shows that when using regular RF and the mean percent increase in MSE VIMP, mother's weight, HAZ (LAZ) at birth, WAZ at birth, income, and expenditure were the top 5 predictors as in Figure 25 for 5k and 50k trees. These are the top 5 with 500 trees as well, but with income and expenditure switching order. Thus, three out of the top 5 in each method were the same, in particular mother's weight was ranked the most important followed by HAZ (LAZ) at birth. One must note that the conditional RF and thus conditional VIMP was not explored in this research to create confidence intervals and such due to the computation time and memory required for increasing the number of trees. The conditional RF also uses a separate R package in which not all of the same information can be accessed and thus lead to comparable results. One must also note the common criticism for RF which is to avoid the regular VIMP metrics when the predictors are correlated which is the case here. Therefore, these results should be taken with the reminder of the effect correlation may have for this data set.

The method described above in 2.4.1 to implement bootstrapping from the per tree VIMP values and thus creating confidence intervals for a particularly trained RF about the VIMP is implemented here for the PROVIDE data. The results in Figure 26 show the mother's weight, HAZ (LAZ) at enrollment, income, WAZ at enrollment, and expenditure being the top 5 predictors where mother's weight is the most important and HAZ (LAZ) at enrollment is the second most important since their confidence intervals do not overlap with any other variables' intervals. The next three all have overlapping confidence intervals with expenditure additionally overlapping with IL-4 at week 18 and ever so slightly with mother's height. Thus, income, WAZ at enrollment, and expenditure are in the running for ranks 3, 4, and 5 with expenditure also being in the running for ranks 6 and 7. Looking at the ranks, mother's weight and HAZ (LAZ) at enrollment had constant rankings at 1 and 2 respectively while income ranged from ranks 3 to 4. WAZ at enrollment and expenditure's ranks ranged from 4 to 5. These top 5 predictors' confidence intervals did not overlap with

any other predictors' confidence intervals meaning they are indeed the top 5 even though there is some variability within the specific ordering of these variables for this particularly trained RF.

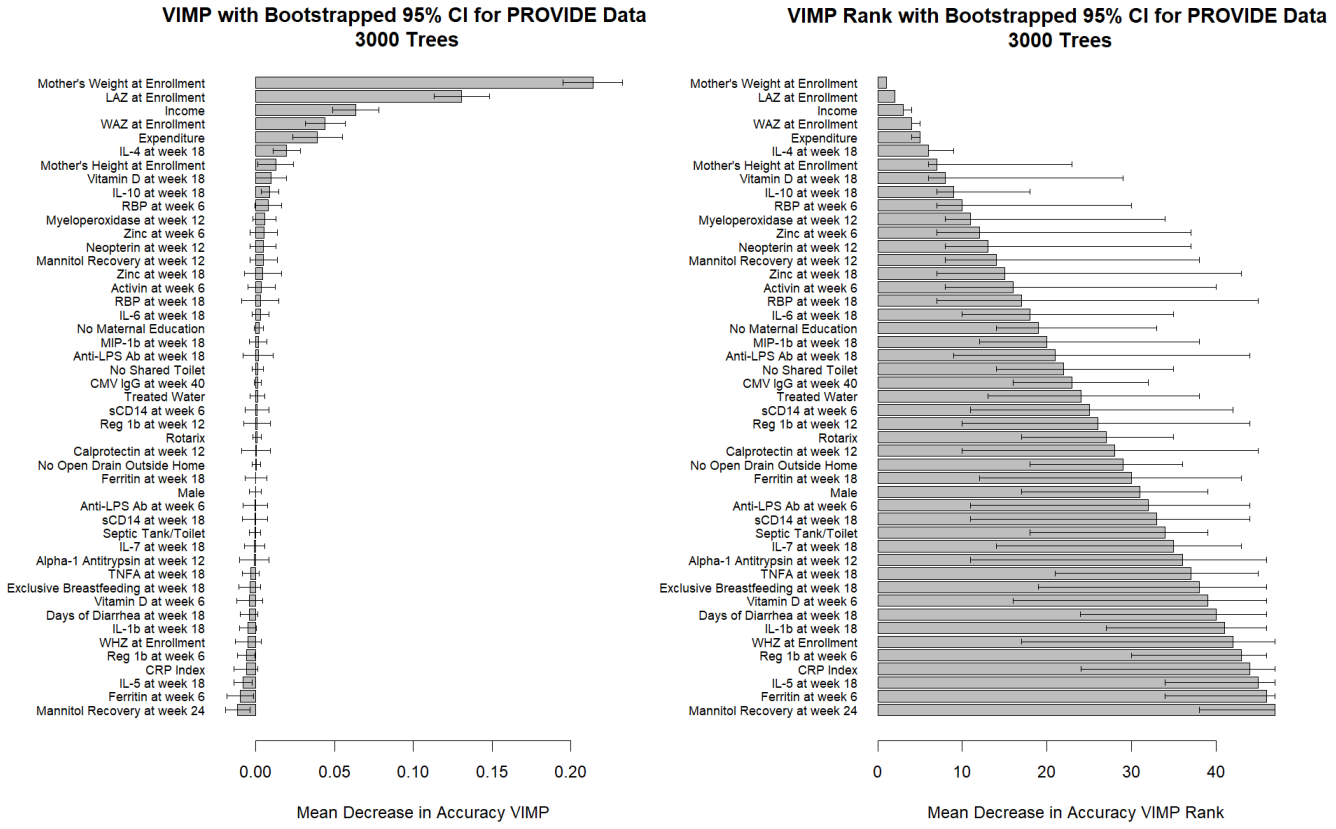

Figure 26: Mean increase in MSE VIMP (left) and rank of VIMP (right) with bootstrapped 95% confidence intervals (CI) from the trained random forest's per tree VIMP values for the PROVIDE data.

The PROVIDE data was also used in conjunction with Ishwaran and Lu's procedures [Ishwaran and Lu, 2018]. These methods produce Figure 27 which shows by the double bootstrap that only mother's weight and the infant's HAZ (LAZ) at birth are important due to high variability while from the subsampling and the delete-d methods income, expenditure, and the infant's WAZ at birth are added to that list of important variables. These methods, especially the subsampling method, corroborate with the our RF results and methods in that the mother's weight followed by the HAZ (LAZ) at birth are the top two important while

the other three are definitely in the top five. The specific rankings are a bit different, but in general the confidence intervals give similar results as our method in Figure 26. Timing wise, the double bootstrap method takes quite a bit longer than the other two and even longer than creating a RF with 50k trees. Thus, out of these three methods and for this data set, either the subsampling or delete-d methods should be prefer especially if the double bootstrap is performing poorly without having to increase the sampling rate for the subsampling method.

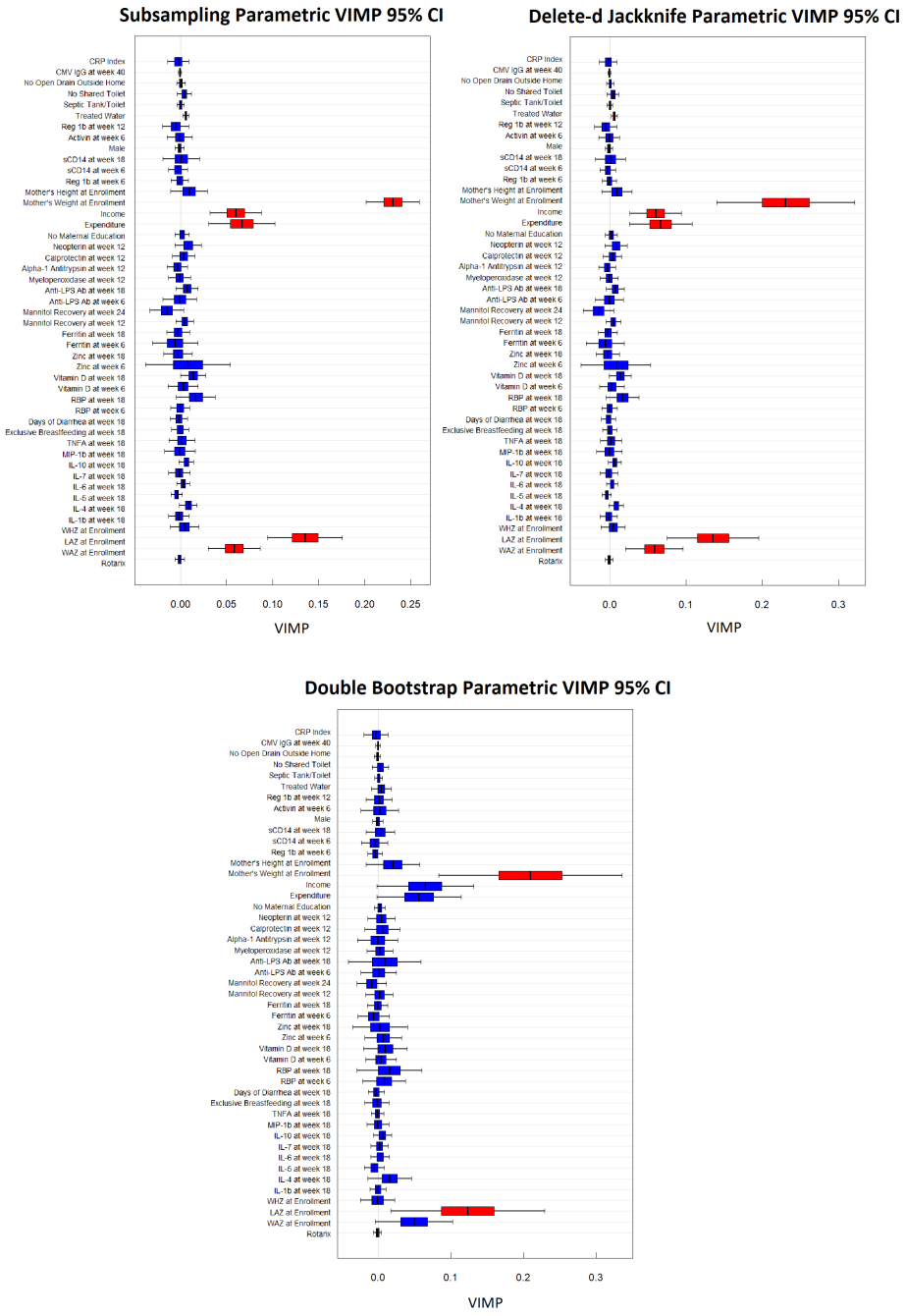

Figure 27: Subsampling parametric (top left), delete-d jackknife parametric (top right), and double bootstrap parametric (bottom) confidence intervals with methods from article for PROVIDE Data [Ishwaran and Lu, 2018]. Mother's weight, income, expenditure, HAZ at enrollment, and WAZ at enrollment were selected in the top graphs while only income and HAZ at enrollment were selected for the double bootstrap method.

2.5 Creating a Threshold to Select Important Variables through Applications

It is a very nice result to be able to rank variables in order of importance, however, how can we tell exactly which variables to select? Specifically for calculations of VIMP within the RF setting, a straightforward approach may be to select any variables with a positive VIMP value as a value equal to or below zero clearly indicates a variable's lack of predictive power towards the outcome [Ishwaran and Lu, 2018]. However, what do values slightly above zero indicate? Should they still be considered important or useful? Confidence intervals are one way to answer these questions, but another is to create a threshold or cutoff VIMP value which may help to select the important predictors.

In order to state which variables within a data set are actually important for the prediction of the outcome, we propose then apply the following method for creating a cutoff value in order to choose important predictors. The steps are as follows:

1. Shuffle the observed values for all predictor variables.
2. For each shuffle, calculate the VIMP and take the maximum, minimum, median, and mean over all variables.
3. Repeat steps 1 and 2 several times, say 100 times.
4. Take the maximum of the results to find potential cutoffs of which variables should be considered important.

The maximum of the maximums indicates the value of VIMP which is the highest such value corresponding to essentially nonsense. Meaning that any predictor with a VIMP value above this threshold should definitely be deemed an important predictor. However, this particular maximum may be higher than any of the variables' VIMP values as in the top right panel in Figure 28. Thus, the minimum, median, and mean were also considered. Similarly for

the maximum of the minimums, this cutoff may be below all of the variables' VIMP values and thus all predictors may be deemed important shown by the top left panel in Figure 28. This is where the maximum of the medians or of the means may be a tradeoff for these more extreme cutoffs.

2.5.1 VIMP Threshold for the NICU Data

Under the NICU data, the four cutoffs were calculated as described above per each VIMP measure as shown in Figure 28. In the logistic regression setting the absolute value of the test statistic was used which is the bottom panel of Figure 28. The max of the means and medians are very close to each other with both suggesting there are six important variables: birth weight, mean SPO_2, Apgar at 5 minutes, mean heart rate, sex (male= 1), and Apgar at 1 minute. The cutoff based on the minimums chooses an additional 5 variables leaving only 6 deemed unimportant. The threshold based on the maximums chooses one important variable. For the RF method, the mean decrease in accuracy VIMP and the Gini index were the VIMP measures. The means and medians cutoffs correspond fairly well for the mean decrease in accuracy VIMP but there are some differences for the Gini VIMP. For the mean decrease in accuracy VIMP, all the variables are selected by the mins cutoff and all but 3 are selected for the means and medians while the maximum of the maximums cutoff selects 9 predictors as important. With the Gini index, the threshold from the maximums chooses no important variables while the minimums cutoff chooses all but three. The cutoff from the means selects 9 and the cutoff from the medians chooses only 3 predictors with them being the birth weight, the standard deviation of SPO_2, and the mean of SPO_2. In short, all three plots give different ordering to the variables, except the birth weight being the most important in all and the mean SPO_2 being in the top 3 for each method. Each of the cutoff values also indicate various variables as being important and thus no one threshold seems to stand as the best.

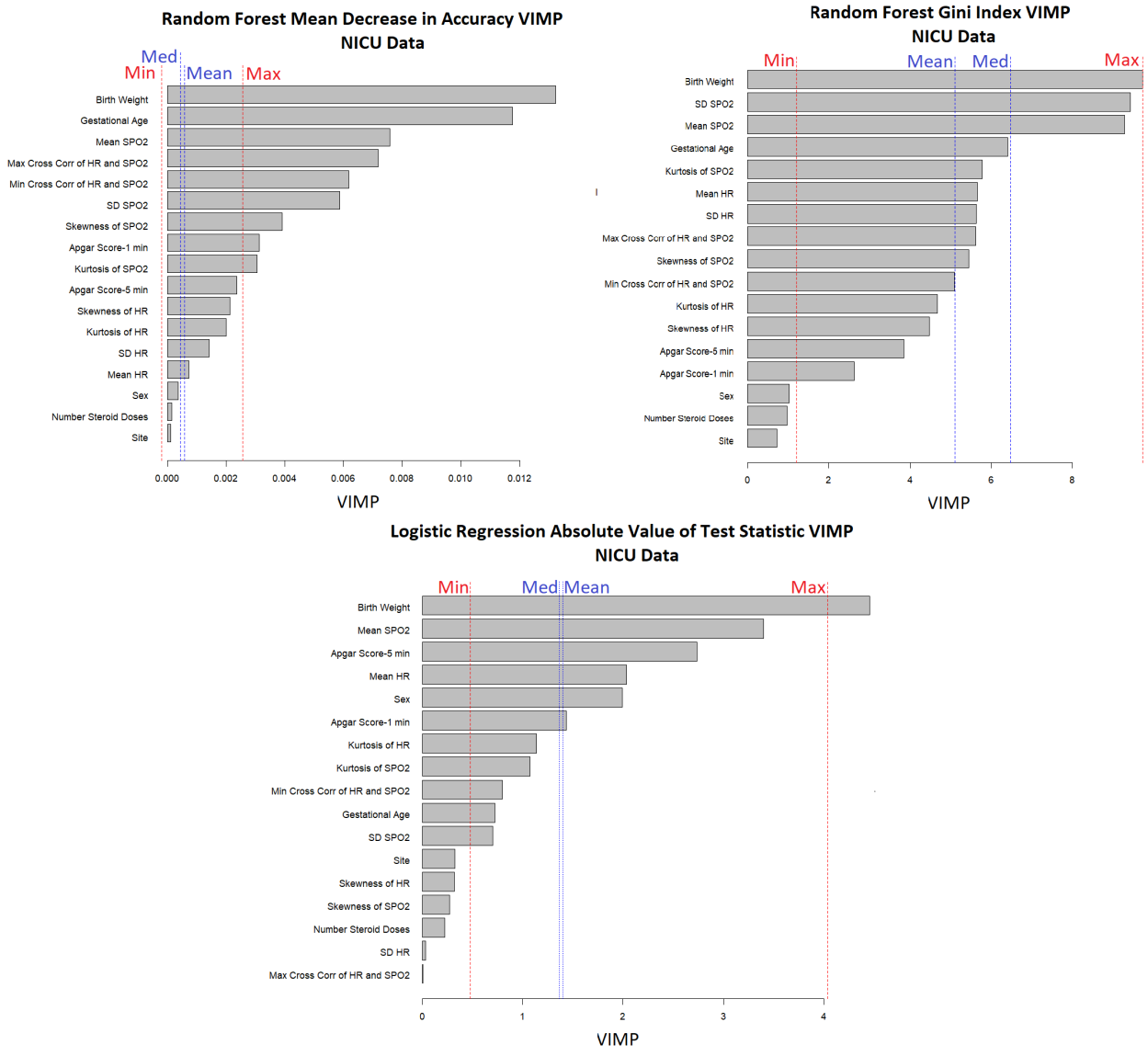

Figure 28: Creating cutoffs to select NICU variables using random forest's mean decrease in accuracy VIMP (top left) and Gini index VIMP (top right) and the absolute value of test statistic for logistic regression. Cutoffs are the maximum of the minimums (Min), medians (Med), means (Mean), and maximums (Max) for the VIMP of all variables over several random shuffles of the data.

2.5.2 VIMP Threshold for the PROVIDE Data

In order to find a cut point for selecting important variables with the PROVIDE data, conditional and regular random forest methods were implemented. With the conditional random forest (CRF), the conditional VIMP and the mean percent increase in MSE VIMP were calculated. For random forest (RF), only the mean percent increase in MSE VIMP was computed. Again, all four cutoffs were applied to each situation giving Figure 29. For the CRF with conditional VIMP the maximum of the maximums cutoff selects no predictors while the maximum of the minimums cutoff selects all but five. The medians threshold is less than the means threshold by enough to select several more predictors (around 8 additional). The cutoff by the means selects only 6 variables: mother's weight, HAZ (LAZ) at birth, income, RBP at week 18 (retinol binding protein), calprotectin at week 12, and the indicator for having a septic tank/toilet for the home. For the mean percent increase in MSE VIMP calculation within CRF, the minimums chooses all but two while the maximums selects only mother's weight as an important predictor. The medians threshold is very close to zero but slightly below indicating it's no better than choosing the variables with positive VIMP. The max of the means cutoff selects five variables: mother's weight, HAZ (LAZ) at birth, income, IL-4 at week 18, and expenditure. With the RF method, the minimums chooses all of the variables and the maximums selects no variables both being completely unhelpful to select variables. The maximum of the means selects 8 predictors with the top four being mother's weight, HAZ (LAZ) at birth, income, and expenditure while the max of the medians selects several more, but is only slightly above zero. For the most part, the maximum of the means has performed the best for the PROVIDE data to find a balance between selecting too few and too many predictors. As for the three different VIMP calculations, the top few variables mostly agree across the different calculations. It's safe to say that mother's weight, HAZ (LAZ) at birth, and income are the top three important predictors of HAZ at two years for this data set. However, similar results as with the NICU data and these threshold calculations arise in that there is no one best cutoff metric that stands out.

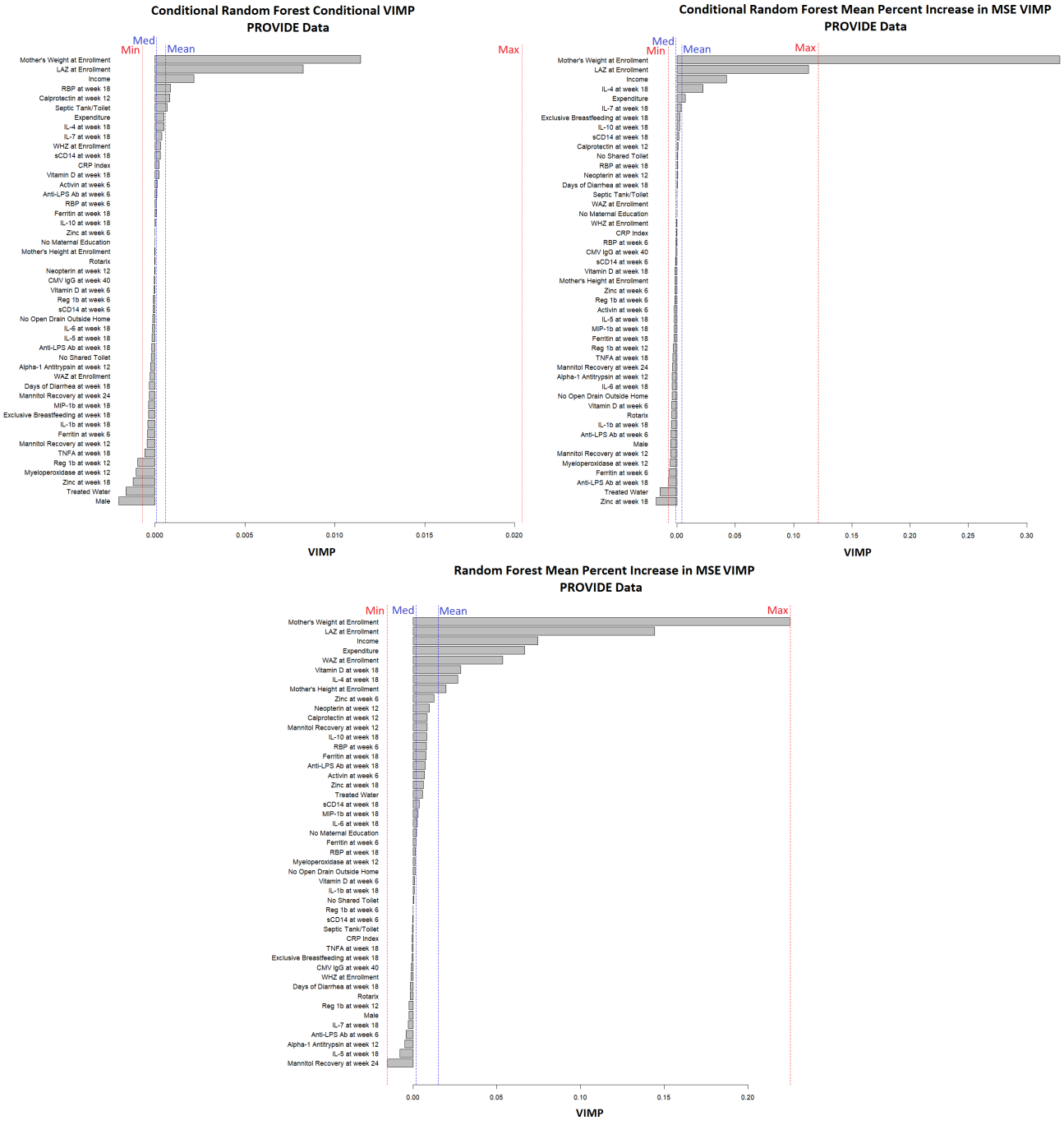

Figure 29: Creating cutoffs to select PROVIDE variables using conditional random forest's conditional VIMP (top left) and mean decrease in accuracy VIMP (top right) and the mean decrease in accuracy for regular random forest. Cutoffs are the maximum of the minimums (Min), medians (Med), means (Mean), and maximums (Max) for the VIMP of all variables over several random shuffles of the data.

2.6 Theoretical Aspects for the Probability of Obtaining the Important Variable

The following ideas are suggested and supplemented by the simulations in 2.6.1 which show the estimated probabilities in various situations along with the maximum absolute error between the estimated coefficients and their true values. The simulations are all described in 2.6.1 in detail along with their implications.

Theorem 2.1. *Suppose we have p_n predictor variables $X_1, \ldots, X_{p_n}$ from a sample of size n which have been standardized. The continuous response variable Y is regressed on these variables. We assume the correct model where $\beta_1, \beta_2, \ldots, \beta_{p_n}$ are the true coefficients and there is one important variable such that $\beta_1 > \max(\beta_2, \ldots, \beta_{p_n})$. From the regression we obtain the estimates $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{p_n}$ which are the VIMP measures.*

Let $e_j = \hat{\beta}_j - \beta_j$ for all predictors $j = 1, \ldots, p_n$. If $\max(|e_j|) \to 0$, then

$$P\left(\hat{\beta}_1 > \max(\hat{\beta}_2, \ldots, \hat{\beta}_{p_n})\right) \to 1.$$

That is, as the maximum distance between the estimated coefficients, the VIMP measures, and their respective true coefficients decreases, the probability of the VIMP for the important variable being greater than the maximum of all other VIMP values for the rest of the predictors goes to one.

For the case with two predictors we have the following: Let $\hat{\beta}_1 \sim N(\beta_1, \sigma_1^2)$ and $\hat{\beta}_2 \sim N(\beta_2, \sigma_2^2)$ where ρ is the correlation between these random variables and $\sigma_1^2 = \sigma_2^2 = 1$ since the data is standardized. Then, the distribution of $Z = \hat{\beta}_1 - \hat{\beta}_2$ is:

$$Z \sim N\left(\beta_1 - \beta_2, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\right) = N(\beta_1 - \beta_2, 2 - 2\rho)$$

So, $P(\hat{\beta}_1 > \hat{\beta}_2) = P(\hat{\beta}_1 - \hat{\beta}_2 > 0) = P(Z > 0) = \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}\sqrt{2-2\rho}} \exp\left(\frac{z^2-(\beta_1-\beta_2)}{2(2-2\rho)}\right) dz$

Theorem 2.2. *(Conjecture) Following from the previous theorem, let* $e_j = \hat{\beta}_j - \beta_j$ *for all* $j = 1, \ldots, p_n$ *and* p_n *be the number of features for sample size* n*. If* $\frac{\log(p_n)}{n} \to 0$ *as* $n \to \infty$*, then*

$$\max(|e_j|) \to 0$$

and thus

$$P\left(\hat{\beta}_1 > \max(\hat{\beta}_2, \ldots, \hat{\beta}_{p_n})\right) \to 1.$$

That is, if the number of features for sample size n *over the sample size converges to zero, then the maximum distance between the VIMP measures and their true values will decrease to zero and thus the probability of the VIMP for the important variable being greater than the maximum of all other VIMP values for the rest of the predictors goes to one.*

A counterexample to this second theorem (conjecture) has been constructed. The simulation explores the probability of correctly obtaining the one and only important variable with no correlation between any predictors. Thus, $\boldsymbol{\beta} = [1, 1, 0, \ldots, 0]^T$ gives the true coefficients. This simulation sets $n = clog(p_n)$ where c is a constant and p_n is the number of features for sample size n. Thus, various values of p_n and c were set in order to calculate the corresponding sample size n with results shown in Figure 30. The goal of this particular simulation was to figure out when $P\left(\hat{\beta}_1 > \max(\hat{\beta}_2, \ldots, \hat{\beta}_p)\right) \nrightarrow 1$. So, p_n variables each with standard deviation equal to one were simulated, the response $Y = \boldsymbol{\beta}\mathbf{X} + \epsilon$ where $\epsilon \sim N(0, 1)$ was calculated, and the coefficients were estimated through linear regression. This was repeated 1000 times to calculate the percent of times the important variable's coefficient is larger than the maximum of the rest of the coefficients. In other words, the $P\left(\hat{\beta}_1 > \max(\hat{\beta}_2, \ldots, \hat{\beta}_p)\right)$ was estimated. Thus, for different values of c and over the $log(p_n)$ or p_n, Figure 30 shows the percentage of times the important variable's coefficient was larger than the maximum of the rest and the log of the percentage of times the important variable's coefficient was not larger than any of the other variables'. As a result, the probability of correctly obtaining the important variable does not always go to one even in the simplest case of no correlation. Thus, if the number of predictors is growing faster than the sample size, one can no longer correctly identify the important variable using the coefficients.

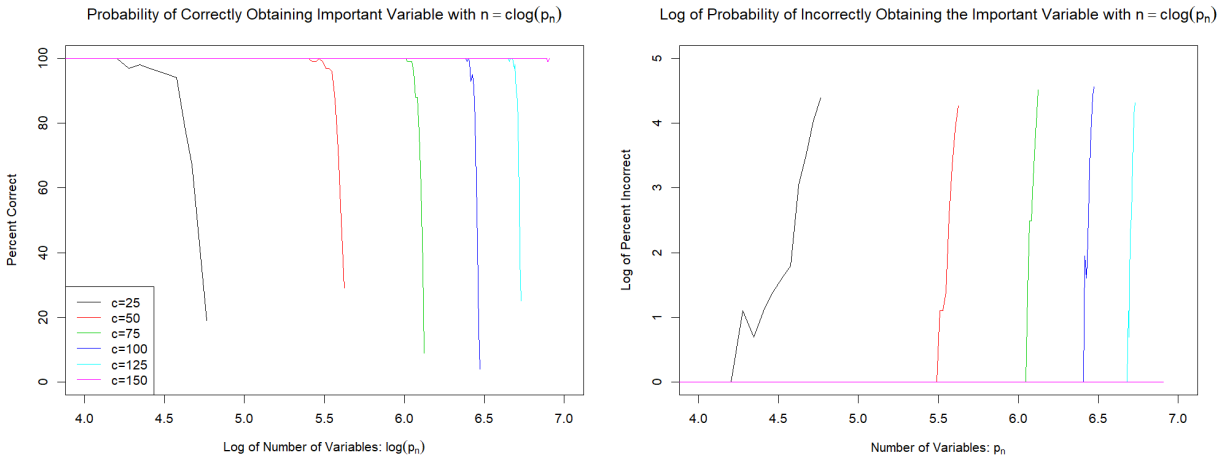

Figure 30: The probability of correctly obtaining the only important variable when $n = clog(p_n)$ where c is a constant, p_n is the number of variables for sample size n over the log of the number of variables when no correlation is involved (left). The probability of incorrectly obtaining the only important variable versus the number of variables for linear regression (right).

2.6.1 Simulations for the Probability of Obtaining the Important Variable(s)

The main question for the simulations is: What is the probability of correctly identifying the most important variable as the number of variables and sample size increases? The framework is of the theorems above where the data is assumed standardized prior to estimating the coefficients in linear regression. The importance of a variable is estimated by the coefficient. In order to explore this, multiple simulations were created to find the percent of times the correct important variable or correct order was obtained and the maximum absolute difference between the predictors' estimated coefficients and the true coefficients was computed.

For the simplest case, only two variables were considered with one variable being deemed important meaning that the true coefficient of this variable was one and the true coefficient of the unimportant variable was zero. The correlation between these variables was varied and different sample sizes were assessed. The method was ran 1000 times to calculate the probability of having the estimated coefficient for the important variable be larger than

the estimated coefficient of the unimportant variable. The process for each combination of correlation ρ and sample size n is as follows:

1. Simulate two variables each with standard deviation one and correlation ρ. Specifically, $\mathbf{X} \sim N(0, \Sigma)$ where $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$.
2. Set the true coefficients as $\boldsymbol{\beta} = [1, 1, 0]^T$.
3. Calculate the response as $Y = \boldsymbol{\beta}\mathbf{X} + \epsilon$ where $\epsilon \sim N(0, 1)$.
4. Calculate the regression model to obtain the estimated coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$.
5. Repeat steps 1 through 4 1000 times.
6. Calculate the percentage of times $\hat{\beta}_1 > \hat{\beta}_2$ over the 1000 trials to estimate $P\left(\hat{\beta}_1 > \hat{\beta}_2\right)$.

These percentages are plotted over the correlations for different values of n in Figure 31 along with the log of the probability of incorrectly obtaining the important variable. These graphs show that with a small sample size and especially with higher correlations, the chance of correctly obtaining the important variable goes down quickly for values of correlation close to one. That is, the task of choosing the important variable by its coefficient becomes increasingly difficult as the correlation between the two variables increases and the variables become essentially the same. In order to explore the relationship between sample size and correlation, the percentages were plotted in Figure 32 against $n(1 - \rho)$ per sample size n and correlation ρ. This shows that for such a small sample size, $n = 10$, the probability of obtaining the important variable is lower than for sample sizes $n \geq 25$. Thus, having a decent sized sample along with lower correlations, about $\rho < 0.5$, should allow for easy obtainment of the important variable with only two predictors.

The natural next step includes three variables with one important variable and two unimportant variables. This means that $\boldsymbol{\beta} = [1, 1, 0, 0]^T$ are the true coefficients. First though, the two unimportant variables were considered to be the only correlated variables.

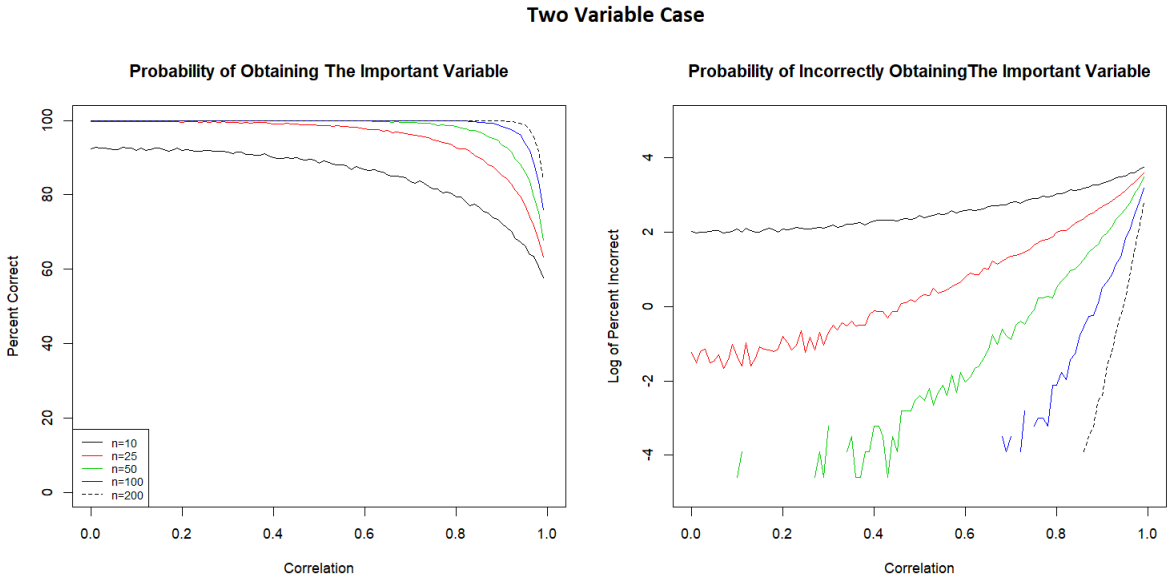

Figure 31: Probability of correctly (left) and the log of incorrectly (right) obtaining the important variable via the estimated coefficient over various correlations and for different sample sizes with only two variables.

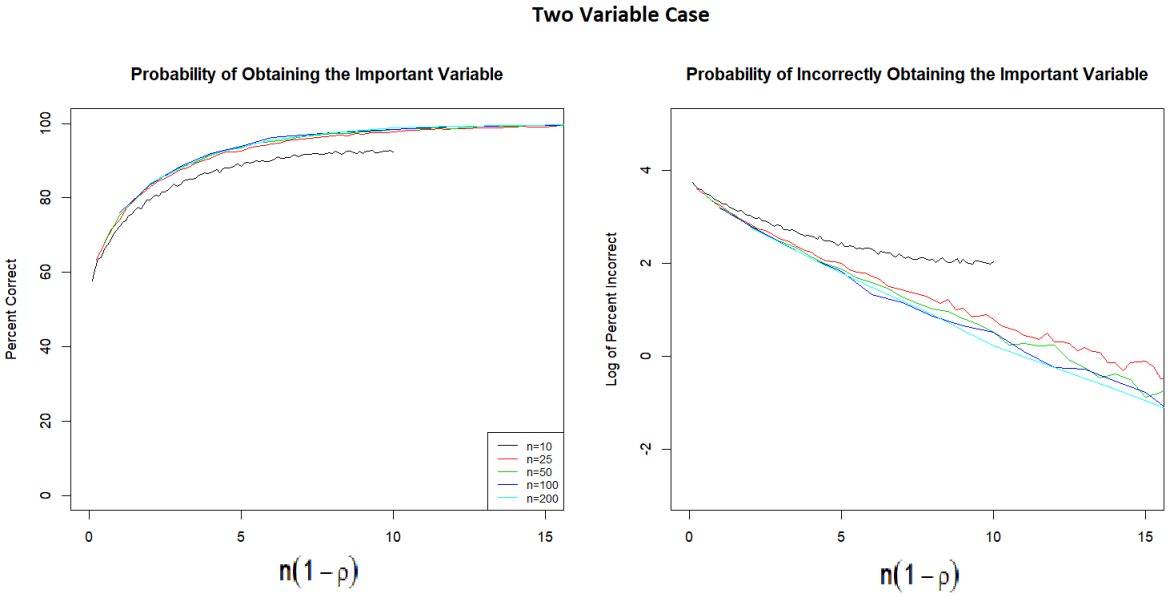

Figure 32: Probability of correctly (left) and the log of incorrectly (right) obtaining the important variable via the estimated coefficients for different sample sizes and over various values of the sample size multiplied by one minus the correlation, $n(1-\rho)$, with only two variables.

Thus, the correlation structure for these variables is $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$. The process described for the two variable case is the same with the appropriate changes to the true coefficients and correlation structure for the three variable case described here. Figure 33 shows the percentage of times the important variable's estimated coefficient was greater than the maximum of the two unimportant variables' coefficients which estimates $P\left(\hat{\beta}_1 > \max(\hat{\beta}_2, \hat{\beta}_3)\right)$. One minus this probability was also plotted in the right panel of Figure 33. Similar to the two variable case, these plots show the difficulty of obtaining the important variable with a small sample size and high correlations. Additionally, the mean difference in the important variable's coefficient and the maximum of the two unimportant variables' coefficients, $mean(\hat{\beta}_1 - \max(\hat{\beta}_2, \hat{\beta}_3))$, was calculated per correlation and sample size combination as in Figure 34. These results exemplify the difficulty of obtaining the important variable as the correlation increases, even though the correlation is only between the unimportant variables in this case. Since the true distance is one and we would need at most an error of 0.5 (half the true distance) to obtain the important variable in which this figure solidifies the results from the percentage plots where higher correlations create difficulty in correctly selecting the important predictor since the estimated coefficients stray away from their true values as the correlation increases. This makes sense because the closer the second and third variables are to each other, the harder it may be to give good estimates of the coefficients especially when one remembers effects of collinearity and the variables essentially becoming one predictor due to the correlation being close to one.

The previous simulation is somewhat simple, and thus the three variable case was further explored where all three variables are important, specifically $\beta_1 > \beta_2 > \beta_3$, and thus $\boldsymbol{\beta} = [1, 3, 2, 1]^T$. This setting was explored with various correlation structures, however not all possible combinations of correlations can be explored due to the correlation matrix becoming

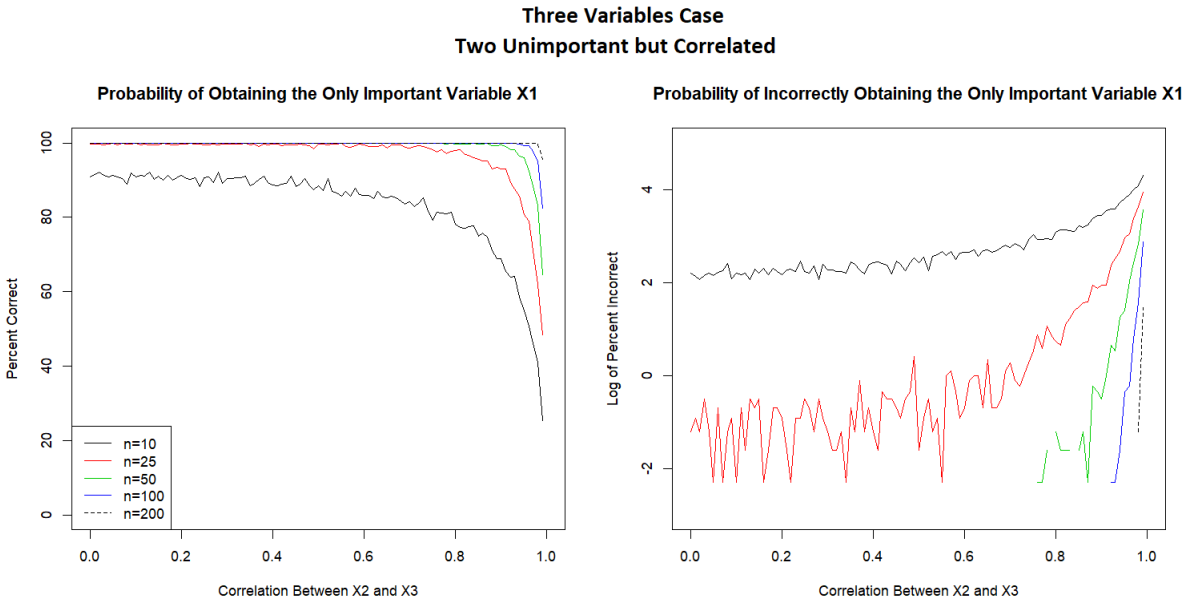

Figure 33: Probability of correctly (left) and the log of the probability of incorrectly (right) obtaining the only important variable via the estimated coefficients over various correlations between the two unimportant variables and for different sample sizes. (Three variable case)

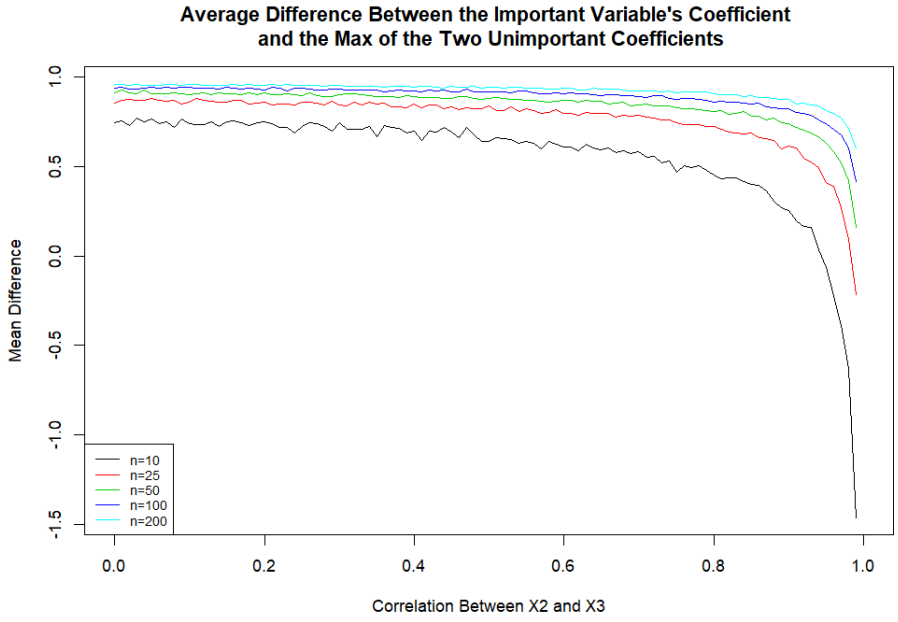

Figure 34: Mean difference between the important variable's estimated coefficient and the maximum of the two unimportant variables' coefficients for different sample sizes and over various correlations between the two unimportant variables. (Three variable case)

not positive definite with certain correlation structures between the three variables. However, the following were explored and simulated:

- Each pair of variables correlated: X_1 & X_2, X_1 & X_3, X_2 & X_3 creating three separate settings in Figure 35.
- All three variables are equicorrelated: They all have the same correlation to each other displayed in Figure 36.
- All three variables are correlated, but with varied values. (This setting has issues with positive definite matrices.)
- Sets of pairs of variables being correlated adding three settings. One example: X_1 & X_2 and X_1 & X_3 are correlated but not X_2 & X_3. (This setting also has issues with positive definite matrices.) Results are shown in Figures 37 and 38.

For each setting, the process is still the same, but with the new true coefficients and with a specific correlation structure per setting in the list above. The plots show the percentage of times the correct order was obtained, in particular $P\left(\hat{\beta}_1 > \hat{\beta}_2 > \hat{\beta}_3\right)$. Due to the complex nature of the results and the issue with not having a positive definite matrix for certain correlation structures, plots for the setting with all three variables being correlated with various values are not included. However, there is still much to learn from the remaining figures. In the first case where only a pair of variables are correlated at one time, we can see from Figure 35 that it is most difficult to obtain the correct order of importance for the variables when the first two or the last two variables are correlated. This shows that it's easier to discern the order when the most and least important variables are correlated. When all three variables are equicorrelated, we again see the distinct pattern of difficulty obtaining the correct order when the correlations increase. However, all three variables are important here not just one, which leads to greater difficulty of identifying the correct order of the three variables with smaller samples than if only one variable were important. In the

last setting of sets of pairs of correlated predictors, the results for the first and second then the first and third being correlated (X_1 & X_2 and X_1 & X_3 are correlated but not X_2 & X_3) are displayed in Figures 37 and 38. These figures display the percentage of times the correct and incorrect order was obtained for various correlations between the two pairs. From this, we can see that when the correlation between X_1 and X_3 is changed from 0.1 to 0.5 and the correlation between X_1 and X_2 is varied, the probability of obtaining the correct order would be zero for correlations between X_1 and X_2 above 0.9. Similar conclusions can be made when the correlation between X_1 and X_3 is varied while the correlation between X_1 and X_2 is changed from 0.1 to 0.5. Also, for various sample sizes, the rate of the percentages seems to decrease faster when the correlation is held at 0.1 for X_1 and X_3 than when the correlation is held at 0.1 for X_1 and X_2. This is similar to the previous results in that the more separation there is between X_1 and X_2 and X_2 and X_3, the easier it is to obtain the correct order.

To infer more about when the probability of obtaining the correct important variable does not go to one, the maximum absolute error between the estimated coefficients and the true coefficients is explored. In this setting, there are p_n predictors with all the same level of importance such that $\boldsymbol{\beta} = [1, 1, 1, \ldots, 1]^T$. The sample size n was varied along with the number of predictors p_n and the correlation ρ between all predictors (equicorrelation). Similar to before, the data was simulated, the response calculated, and the coefficients estimated. The next step was to find the maximum absolute error defined as $\max(|e_j|) = \max\left(|\hat{\beta}_j - \beta_j|\right)$ for $j = 1, \ldots, p_n$ which measures the maximum absolute distance between each estimated coefficient and their respective true coefficient. This value was calculated for each trial for a total of 1000 and then the mean of these maximums was taken for each combination of sample size, correlation, and number of variables. Figure 39 shows these curves for the various sample sizes and for different numbers of variables. From these curves, we see that the error decreases as the sample size increases while the error increases over increasing correlations in which the difference between the variables' estimated and true coefficients becomes larger

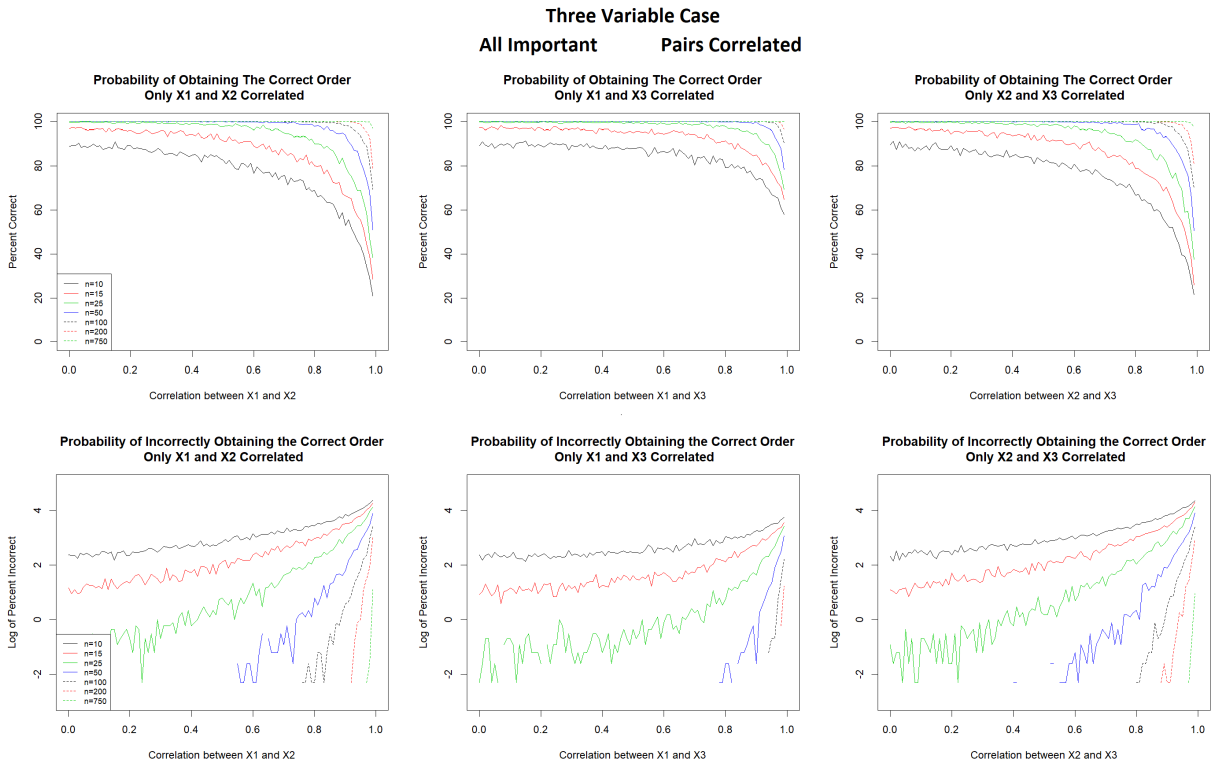

Figure 35: Probability of correctly (top three) and log of the probability of incorrectly (bottom three) obtaining the correct order of the important variables via the estimated coefficient for different sample sizes over various correlations for each pair of the three variables.

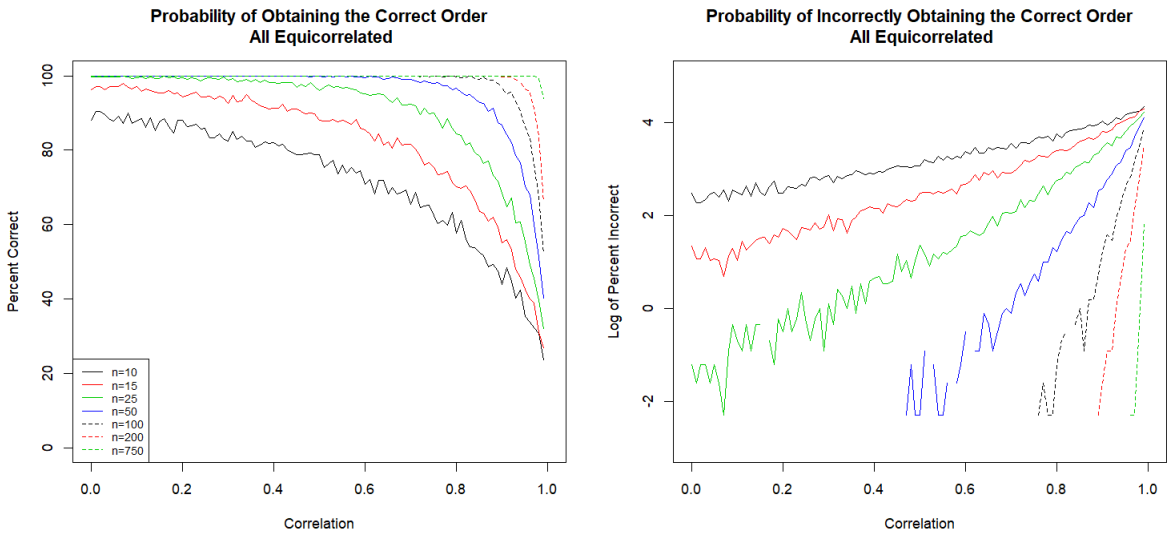

Figure 36: Probability of correctly (left) and the log of the probability of incorrectly (right) obtaining the correct order of the important variables via the estimated coefficient over various correlations between all variables (all equally correlated) and for different sample sizes.

making it increasingly difficult to extract the important variable. In order to explore the effect sample size has on the difference between these curves of different numbers of predictors, the ratios between these curves were calculated and plotted in Figure 40. From this we see that there are very similar ratios between the curves with the exception of the ratio between $p_n = 50$ and $p_n = 100$ which is different for the smallest sample size than for the larger sample sizes.

These mean maximum absolute errors were also plotted over $\frac{1}{1-\rho}$ in Figure 41 to show how the lines may become straighter and more linear in their trends. The same setting was repeated then, but for $\rho = 1 - \frac{100}{n}$ in Figure 42. This simulation also shows a more linear trend in the mean maximum absolute error versus the correlation ρ when it's related to the sample size n. This particular plot also gives rise to lower errors leading to a suggestion in which one may choose a sample size based on the correlation such that $n = \frac{100}{1-\rho}$. The reasoning is that no matter how large the distance is between the estimated and the true coefficient, a maximum absolute error half the true distance or less between the estimated

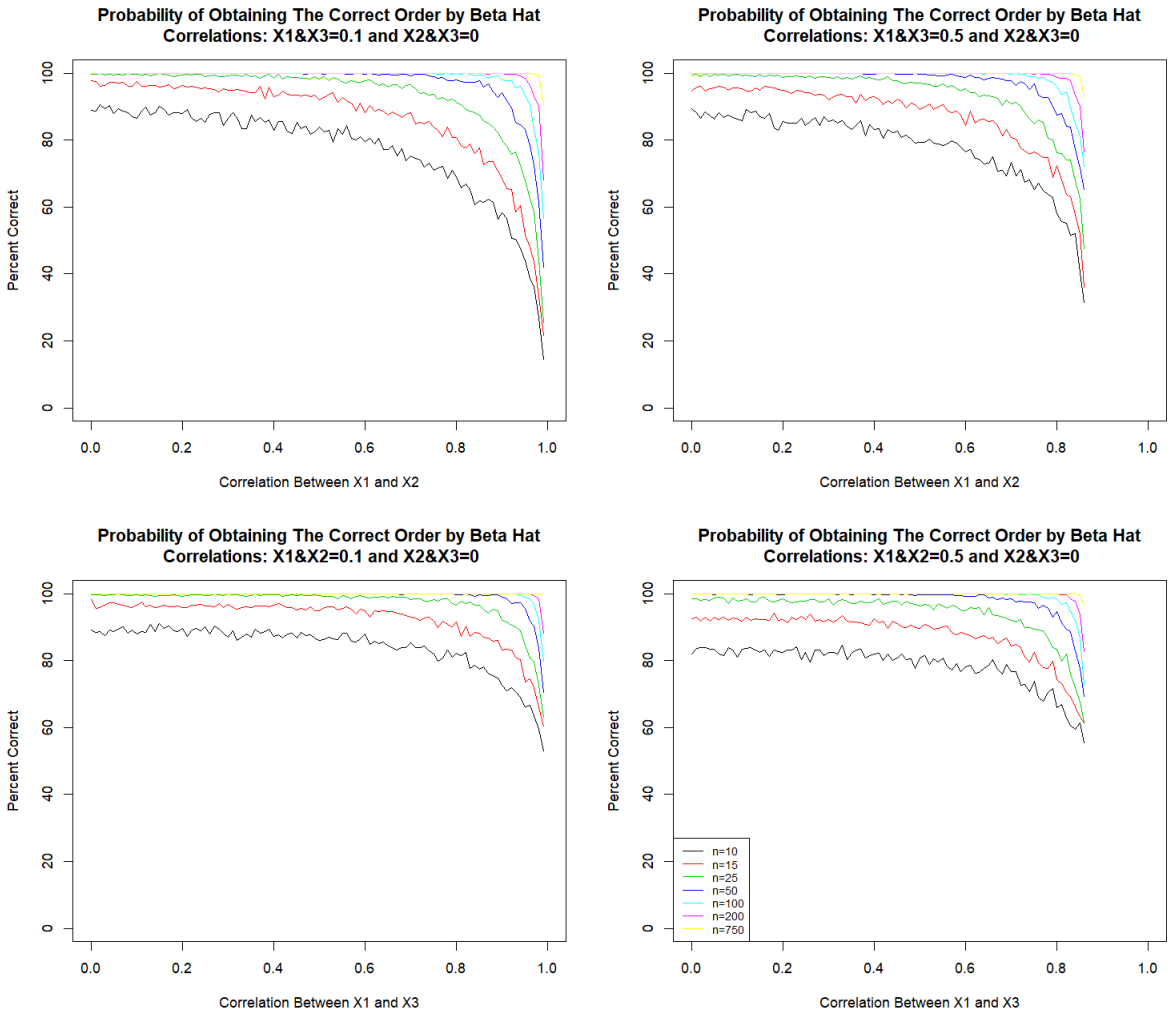

Figure 37: Probability of correctly obtaining the correct order of the important variables via the estimated coefficient for different sample sizes over various correlations for when the first and second variables and the first and third variables are correlated while the second and third are not correlated.

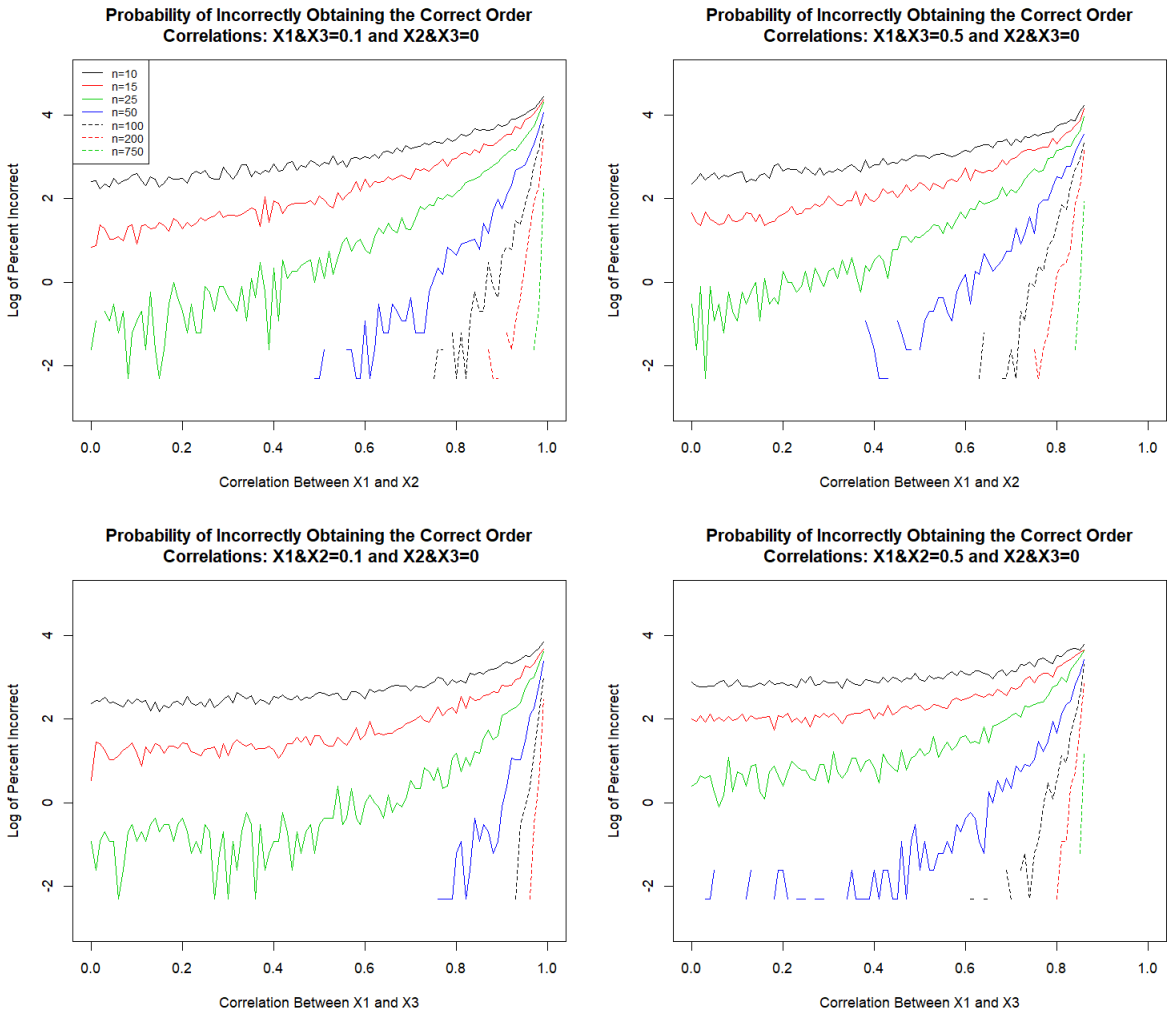

Figure 38: Log of the probability of incorrectly obtaining the correct order of the important variables via the estimated coefficient for different sample sizes over various correlations for the first and second variables and the first and third variables being correlated while the second and third are not correlated.

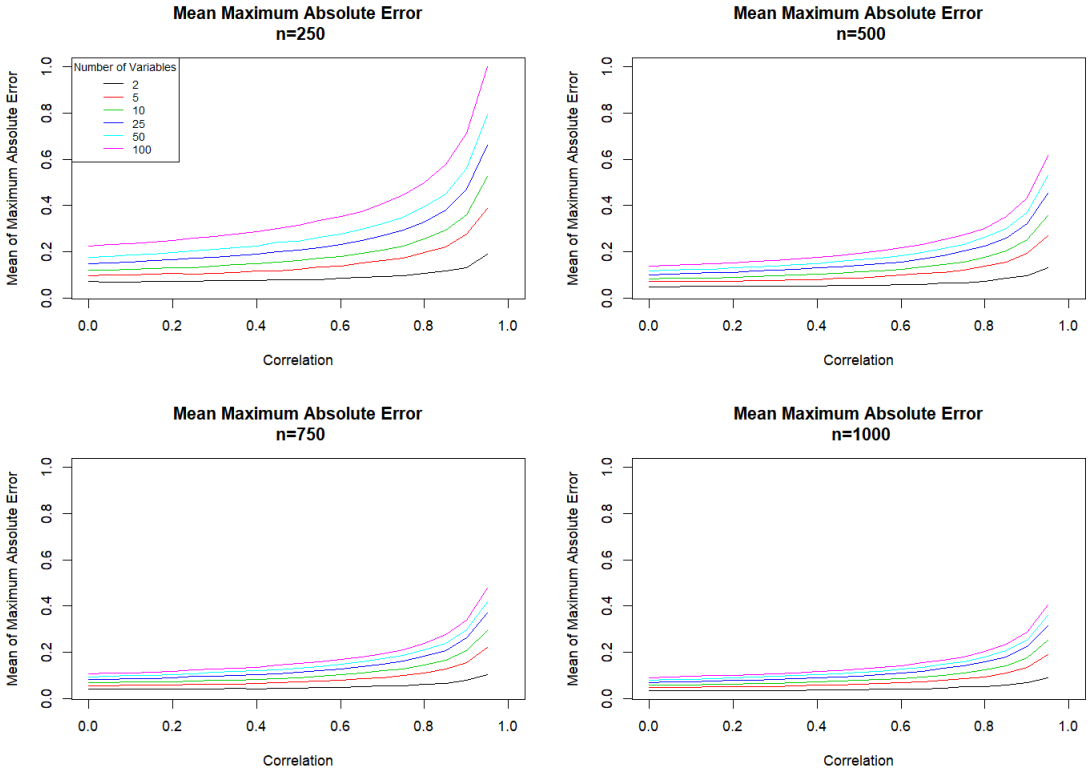

Figure 39: Mean maximum absolute error between the estimated coefficient and their true value over various correlations between the equicorrelated number of predictors and for different sample sizes.

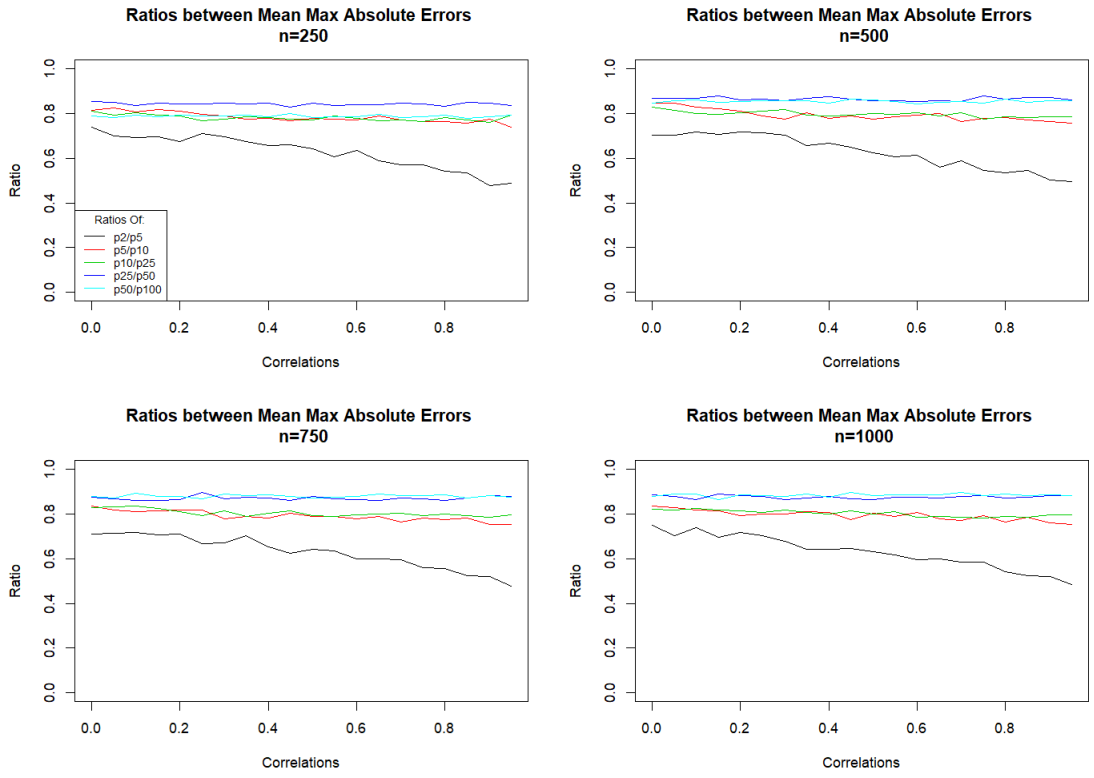

Figure 40: Ratios between mean maximum absolute errors in Figure 39 over various correlations between the equicorrelated number of predictors and for different sample sizes.

and true coefficients is needed for the probability to go to one as long as the estimates are uniformly estimated over the predictors.

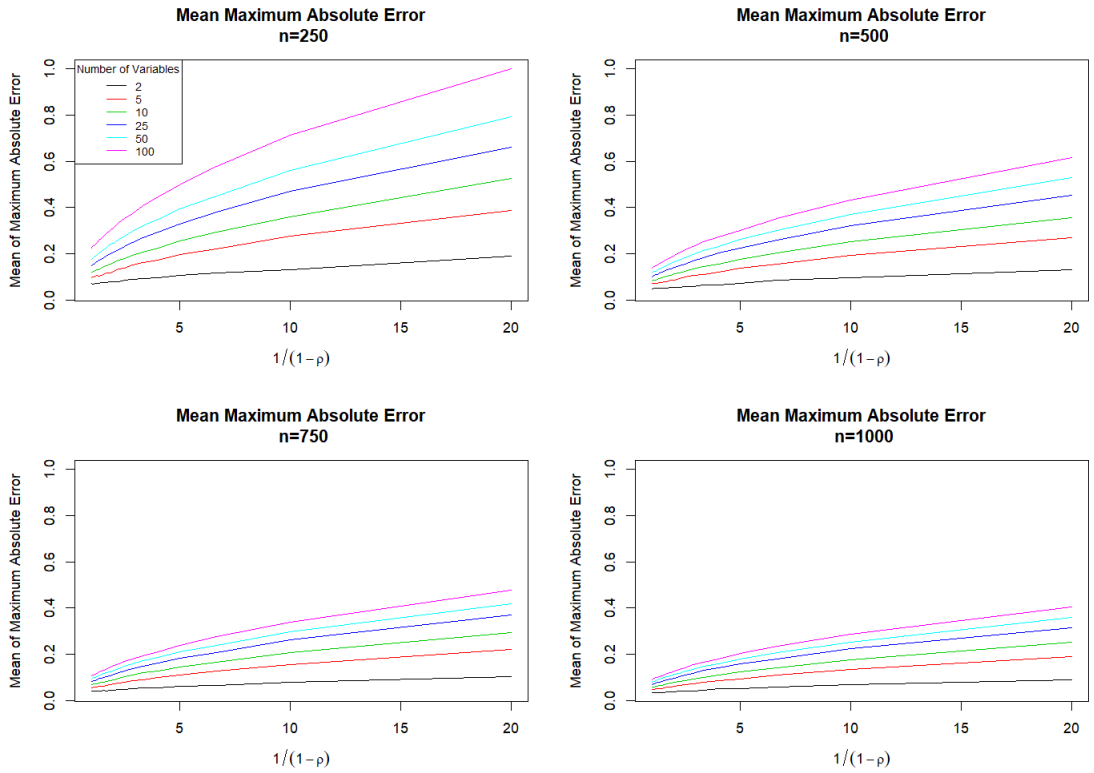

Figure 41: Mean maximum absolute error between the estimated coefficient and their true value over $1/(1-\rho)$ for various correlations, ρ, between the equicorrelated number of predictors and for different sample sizes.

The asymptotic distribution for the error between the estimated and true coefficients of p_n predictors can be used which nullifies the need for creating standardized data, calculating the response, and estimating coefficients which greatly improves the computation time of these simulations. In particular, we simulate the distribution as $\mathbf{e} \sim N(0, S^{-1})$ with S being the correlation structure. Thus, p_n errors could be simulated quickly and then the absolute value taken of each and the maximum absolute error calculated over all the errors. This may be repeated 1000 times with the mean maximum absolute error calculated per correlation, ρ, and number of variables, p_n, combinations as in Figure 43. The ratios between these mean maximum absolute error curves are also computed for comparison between this simulation

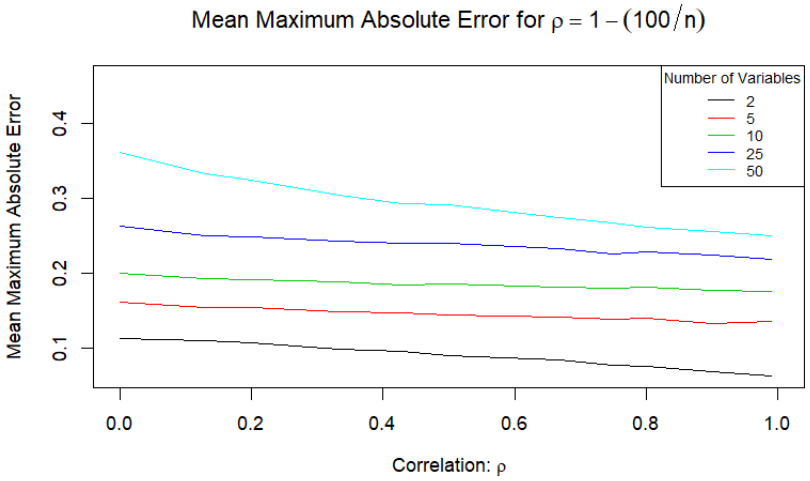

Figure 42: Mean maximum absolute error between the estimated coefficient and their true value over various correlations between the equicorrelated number of predictors and for different sample sizes selected as $\rho = 1 - (100/n)$ or $n = 100/(1 - \rho)$.

with the asymptotic distribution and the full simulations which started by creating data. These results are most similar to the full simulation when the sample size equalled 1000 in Figures 39 and 40. This indicates that the asymptotic distribution is verified by the assumption that the sample size is extremely large for it to hold.

From these simulations, we can conclude that as the correlation increases, the difficulty of obtaining the important variable or the correct order of the important variables also increases. As expected, the difficulty increases as the sample size decreases. Additionally, if the variables are ordered and the most important is correlated only with the least important (or with lower ordered predictors), then it is much easier to correctly order the variables by the estimated coefficients. However a question still begs, do these simulations hold for logistic regression?

The logistic regression case with two variables was explored as shown in Figure 44 and with three variables, two being unimportant but correlated, was explored in Figure 45. To simulate this data, the process is similar to that of linear regression, however the response

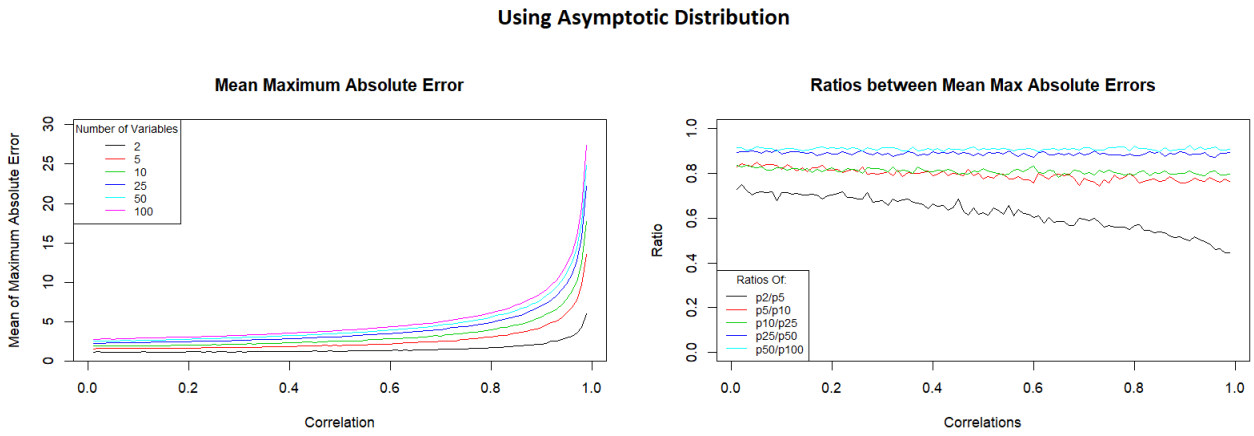

Figure 43: Using the asymptotic distribution of the coefficients, the mean maximum absolute error between the simulated coefficients and the true coefficients is plotted over various correlations between the coefficients and for different numbers of variables (left). The ratios between the mean maximum absolute error curves were also plotted (right).

variable takes on a Binomial distribution. Thus, step 3 from the linear regression simulations above is changed to:

- $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta}$
- $\mathbf{Y} \sim Binom(1, p)$ where $p = \frac{1}{1+\exp(-\mathbf{Z})}$

Figure 44 shows similar results as with linear regression in that as the correlation increases or as the sample size decreases, the percentage of correctly obtaining the important variable decreases meaning it becomes harder to correctly obtain the important variable. However, when the logistic regression case is compared to the linear regression case, these effects of sample size and correlation are higher. For example, with the sample size of $n = 10$, the percent of correctly identifying the important variable over the increasing correlation hovers above 60% until it drops below at high correlation values (around 0.8 and above) for the logistic case whereas for the linear case the percentages are above 90% until a correlation around 0.4 and above 80% until a correlation of 0.8. In terms of the three variable case with one important variable and the two unimportant variables being correlated, the previous findings are solidified as shown in Figure 45. Thus, it is harder to obtain the important

variable in the logistic case as compared to the linear case and with a small sample size or high correlations.

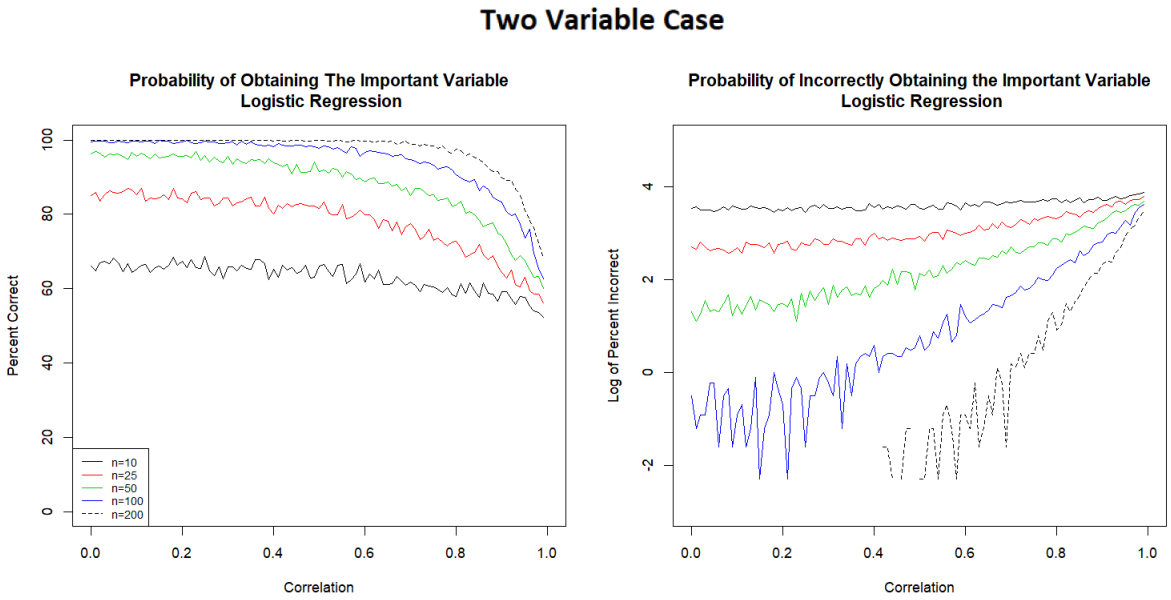

Figure 44: Probability of correctly (left) and the log of the probability of incorrectly (right) obtaining the important variable via the estimated coefficient over various correlations and for different sample sizes with only two variables in logistic regression. (Two variable case)

The counterexample from above where $n = c\log(p_n)$ with different c and p_n values with no correlation present was repeated for the logistic regression case as shown in Figure 46. Similar to the above simulations, it shows that for logistic versus linear regression, it is more difficult to obtain the important variable with lower probabilities for the same situations. However, the same conclusions can be made in that if the number of predictors is growing faster than the sample size, one can no longer correctly identify the important variable using the coefficients for logistic regression even with no correlation present.

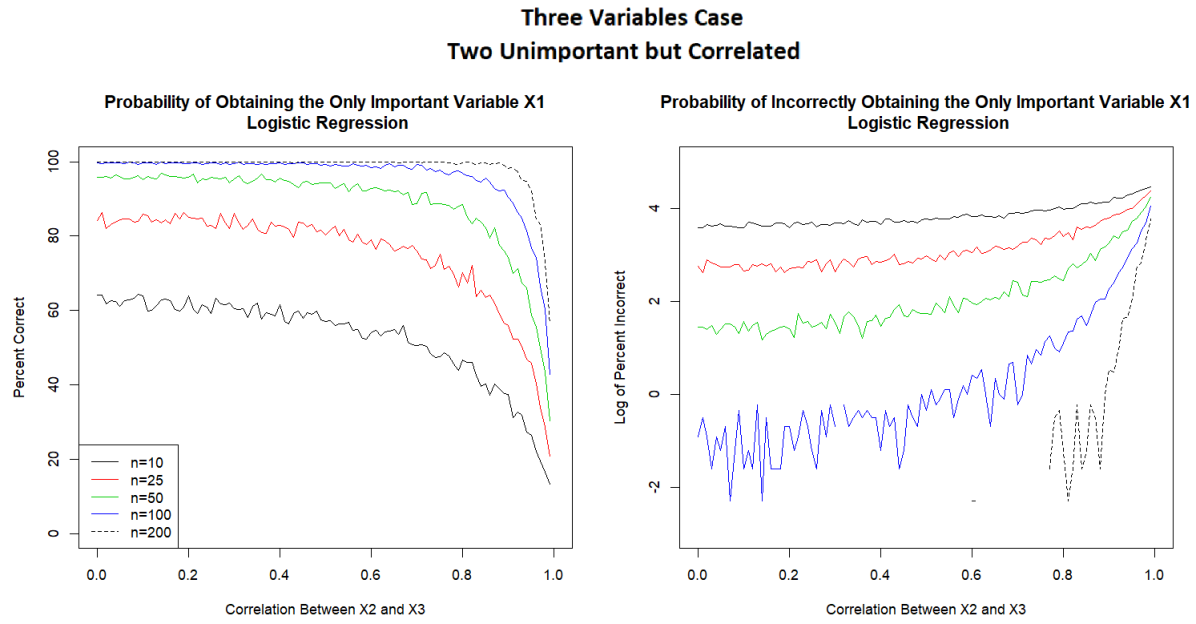

Figure 45: Probability of correctly (left) and the log of the probability of incorrectly (right) obtaining the only important variable via the estimated coefficients over various correlations between the two unimportant variables and for different sample sizes in logistic regression. (Three variable case)

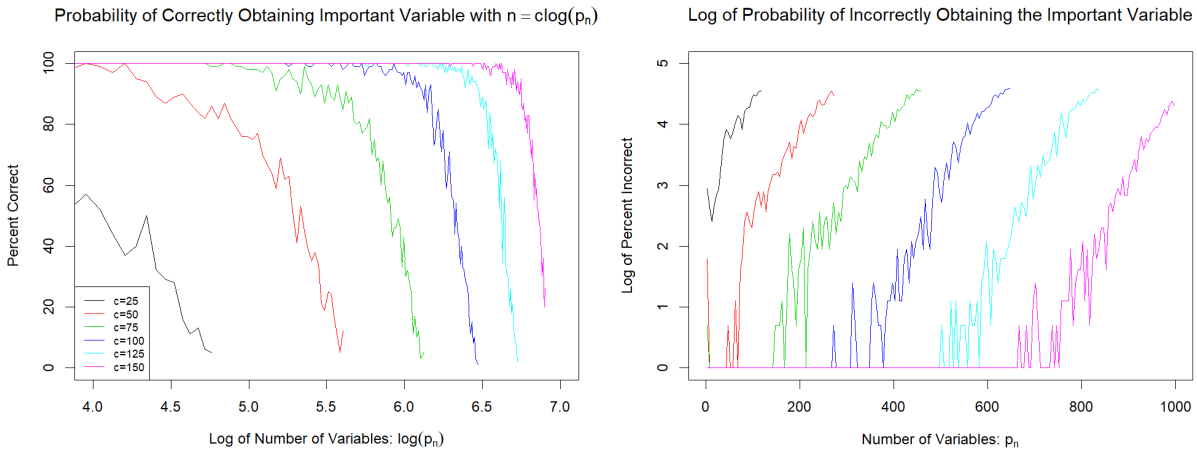

Figure 46: The probability of correctly obtaining the only important variable when $n = clog(p_n)$ where c is a constant, p_n is the number of variables for sample size n over the log of the number of variables when no correlation is involved (left). The probability of incorrectly obtaining the only important variable versus the number of variables for logistic regression (right).

2.7 Potential VIMP Measures and Methods

Due to the numerous and various VIMP measures which are model-specific, other metrics which deal less with the specific model and more with overall performance should be explored. Thus, VIMP could be assessed as a degradation in performance of the log-likelihood, sum of squares error, and R^2 after a random shuffle of the variable's observed values as long as the predictors are uncorrelated. With the case of correlated variables, the group of correlated variables may be assessed together through a random permutation of the observations within the set of the correlated predictors. This means that each variable within the correlated set will have its observed values shuffled. Then this set of correlated variables will be evaluated collectively therefore obtaining a VIMP measure for the whole group which will show the unified importance of the group. This method was proposed by Parr et al. and seems like a viable option in order to avoid the broken correlation structures which would happen if only a single predictor was permuted at a time [Parr et al., 2018].

In addition to the above suggested metrics, the following could also be assessed for classification: Area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, positive predictability, F1 score, and Score1. In each case of classification, after one has built a model and obtained the predicted classes, the misclassification matrix can be obtained as in Table 5. All of the above scores for classification are reliant on this matrix. First let's define sensitivity, specificity, and positive predictability (PP) where the values come from the misclassification matrix.

$$Sensitivity = \frac{TP}{TP + FN}$$
$$Specifiicity = \frac{TN}{TN + FP}$$
$$PP = \frac{TP}{TP + FP}$$

Table 5: Observed vs predicted counts.

	Observed Class 1	Observed Class 0
Predicted Class 1	True Positive (TP)	False Positive (FP)
Predicted Class 0	False Negative (FN)	True Negative (TN)

The ROC curve plots the sensitivity on the y-axis and one minus the specificity on the x-axis. The ROC curve shows the tradeoff between sensitivity and specificity meaning an increase in sensitivity will be followed by a decrease in specificity. If the ROC curve is a straight line at a 45 degree angle and thus the area under the curve (AUC) is 0.5, then the method is uninformative [Agresti, 2014]. When this occurs, sensitivity plus specificity equals one, hence the difference of sensitivity and specificity would be zero which means a positive difference is an improvement. Thus, the higher the AUC, the better the method and the more arched towards the top left the ROC curve will be.

A newer approach of evaluation leads from a 2012 PhysioNet Challenge of adults in intensive care units (ICU) with a low occurrence of interested events (mortality rate was 14.2%) [Silva et al., 2012]. This innovative tactic is referred to as Score1 and is calculated using sensitivity and the positive predictive (PP) value. The decision threshold is varied with the sensitivity and the PP being calculated for each value of the decision threshold. When the sensitivity and PP are plotted across the decision thresholds, the optimal decision cutoff is chosen as the value where the sensitivity and PP are closest. Score1 is then the minimum of sensitivity and PP at this decision threshold. This measure is said to be a "reasonable tradeoff between accuracy of discrimination and prognostic value" [Silva et al., 2012]. Generally, the PP value is preferred over specificity, leading to a reason Score1 may be preferred over the AUC of the ROC curve.

Next, let's define recall and precision in order to define the F1 score.

$$precision = \frac{TP}{TP + FP} = \frac{TP}{TotalPredictedPositive}$$
$$recall = \frac{TP}{TP + FN} = \frac{TP}{TotalActualPositive}$$

From these equations, we see that precision is actually PP and recall is the sensitivity. As for the F1 score, it is a function of these two measures:

$$F1 = 2\frac{precision \cdot recall}{precision + recall} = 2\frac{PP \cdot sensitivity}{PP + sensitivity}.$$

This measure is the harmonic mean of precision and recall from the equation above. Thus, this score will punish extreme values unlike the usual mean. In order to create a model where the precision and recall are balanced, the F1 score would need to be maximized [Koehrsen, 2018]. Collectively, these measures could be assessed depending on the outcome variable through simulations and on real life data.

3 Growth Modeling of Bangladeshi Children

Thus far for growth modeling on subsets from the PROVIDE study or on similar cohorts, functional principal component analysis (FPCA) has been implemented for height-for-age (HAZ) growth patterns over the first two years of life. This involved quantifying deviations of individual's growth from WHO standards by FPCA while linear regression was used to identify risk factors associated with growth faltering [Zhang et al., 2017]. Another analysis included several penalized linear regression methods used to select biomarkers of environmental enteropathy where the main outcome was still HAZ but at one year of age [Lu et al., 2017]. A current working manuscript on a similar data set of children in Bangladesh (the Preschool cohort) extends a comparable analysis using FPCA and penalized regression together in order to model the growth, specifically the height, from ages 3 to 18 when the heights are also observed irregularly. This working article focuses on a procedure which estimates the growth functions of the individuals which then may be used in further methods such as functional regression to explore the relationships with the covariates.

Due to the lack of current growth models which also consider the long list of various risk factors, the PROVIDE cohort is being further explored with the goal of creating a more comprehensive model of height and development with multiple covariates. In order to explore and view relationships, various plots were created for these variables and for each individual. Relationships between growth and the variables dealing with the environment, maternal health, income, and many bodily factors from the infant were also explored using traditional methods such as correlations and penalized linear regressions along with newer machine learning techniques including random forests. Even deep learning methods were considered. Using these methods leads to a more comprehensive model which includes several covariates that have been selected in order to best explain how these children are growing and developing over time, especially within the first two years of life. Thus, interventions may be applied in order to improve the overall growth, development, and health of children

in similar situations. Improving the lives of these children will carry through to improve their lives as adults which in turn would improve their offspring and the overall society as it has been shown that the first few years of life are of utmost importance for success later in life.

3.1 The PROVIDE Study Cohort and WHO Standards

The PROVIDE birth cohort consisted of 700 infants born in Mirpur which is an urban slum in Dhaka, Bangladesh from May 2011 to November 2014. Children were recruited at birth and followed over a two-year period with in-home visits twice a week and irregularly scheduled clinical visits where blood or stool samples were occasionally taken. A more detailed description of the study design, recruitment, and follow-up were described previously [Kikpatrick et al., 2015]. This study was approved by the Ethical Review Board of the ICDDR,B (FWA 00001468) and the Institutional Review Boards of the University of Virginia (FWA 00006183) and the University of Vermont (FWA 00000727). A large set of biomarkers for nutrition and systemic inflammation were calculated from the available stool and blood samples along with numerous survey results, developmental measures, and growth measures including over 900 potential predictor variables (see Table 6 for selected predictors). Even metabolomic and metagenomic data was collected from infant samples in order to gain knowledge about the gut bacteria. A few predictors were collected over time including but not limited to neopterin, CRP, calprotectin, activin A, amino acids, metabolomics, and the number of cumulative episodes of diarrhea. Thus, data is available from all sorts of sources and on numerous aspects of these children's lives ranging from their environment, their mother's health and background, and their own health in various formats. The primary outcome of interest is stunted growth by two years of age defined as a height-for-age Z score (HAZ) below -2 at two years. HAZ is a measure normalized for the child's age and gender from standards released from the WHO Multicentre Growth Reference Study Child Growth Standards. Stunting has been shown to be correlated with subsequent outcomes in

later life such as diminished survival, weakened learning capacity, and lower annual incomes [Dewey and Begum, 2011][Hoddinott et al., 2008]. HAZ is a commonly used measurement for malnutrition due to it's ability to capture the cumulative effects through childhood, however we would like to go back to the raw heights in order to explore this specific population. The heights, weights, and body mass index (BMI) for these children were measured at most 16 times over the two year period. These times are at enrollment and then weeks 6, 10, 12, 14, 17, 18, 24, 39, 40, 52, 53, 65, 78, 91, and 104 where 104 weeks indicates the two year mark. In addition to the initial time frame of two years, additional data was collected on a subset of individuals to include at least yearly outcomes for height, HAZ, weight, WAZ, WHZ, and BMI up to year 7. Developmental tests and scores were also collected including Bayley Scales of Infant and Toddler Development, Third Edition (Bayley's) at 1.5, 2, and 3 years of age, Mullen Scales of Early Learning (Mullen's) at 3 and 4 years, and Weschler Preschool and Primary Scale of Intelligence WPPSI (Weschler's) at 4 and 5 years. Bayley's was replaced by Mullen's at 3 years of age and Weschler's replaced Mullen's at 4 years making the main outcomes Bayley's at 2 years, Mullen's at 3 years, and Weschler's at years 4 and 5. For these neurocognitive outcomes, multiple articles from the PROVIDE cohort have been published including Donowitz et al., Moreau et al. and Jensen et al. [Donowitz et al., 2018] (coauthor) [Moreau et al., 2019] (coauthor) [Jensen et al., 2019]. Data from EEGs (electroencephalogram) was collected on a subset of subjects beyond the two year mark. The EEG net containing electrodes which sense electrical signals from firing neurons was placed on a child's head. Specifically, visual evoked potentials (VEP) were measured in response to a pattern-reversing checkerboard. This means the subject watched a monitor 65cm away with a Tobii X2-60 ete-tracking system attached. These signals were heavily processed as described in Jensen et al. [Jensen et al., 2019]. The latency and amplitude of a component relative to the previous component was measured for the VEP data as shown in Figure 47 after the VEPs were averaged per child. In addition, event related potentials (ERP) were collected using the idea that one person's face will be shown 70% of the time with the other

30% being a new and different face. The ERP data thus tries to measure the subject's ability to distinguish familiar vs unfamiliar faces. The hopes of using the VEP or ERP data is that these could be culturally independent measures of the children's development unlike the Bayley's or similar tests which cannot be directly compared between cultures. Due to the rolling enrollment, some children have not reached age six or seven at the time of this research and thus no time points beyond year 5 was further considered. In whole, this data set holds many complexities especially with the amount of covariates, the irregular times, and relationships between the predictors as shown in Figures 48, 49, and 50. The correlations from some selected variables and the incremental changes in height are shown in Figure 48 showing how variables can have an effect in the change in height at later time points. In particular, Alpha-1-Antitrypsin has a negative effect on the growth rates. Figure 49 shows the subset of metabolomics from the children's stool samples at week 40 which encompasses 85 variables alone (there are also sets of metabolomics from the plasma of the infants at week 40 and the breast milk at week 6). One can see the strong relationships between some sets of metabolomics but the lack of relationship with the incremental changes in height after they were measured. The sets of metabolomics were analyzed by Moreau et al. in which certain sets of these metabolites were associated with either growth or neurocognitive outcomes [Moreau et al., 2019] (coauthor). Figure 50 shows a selected child from the PROVIDE data set whose change in height from enrollment to two years was in the lower quartile (they had one of the lowest increases in height over the two years). The figure demonstrates the sparsity of the data where various data points are missing along with the general irregularity and the multiple types of measurements taken as also stated in 6.

The WHO growth standards come from the WHO Multicentre Growth Reference Study (MGRS) which was completed from 1997-2003. About 8500 children were involved in this study from Brazil, Ghana, India, Norway, Oman, and the USA with all types of backgrounds being considered allowing the impact of the environment to be lessened. Thus, these standards represent growth for all children up to age five where the breastfed infant is considered

Table 6: The Short List of Covariates.

Type/Source	**Measurements**
Maternal	Height and Weight Age at enrollment, first marriage, and first pregnancy Number of living children and marriages Plasma cytokines: IL-8 and TNF-α Breast milk metabolomics Breast milk cytokines: IL-7 and PDGFBB Breast milk lipids: LA, ALA, EIC9, DGLA, GLA, LLA, PLA, STE, EDA, DPA, ELA, MYR
Infant	Gender Number of cumulative episodes of diarrhea Birth order of enrolled infant Number of wheezing episodes over 2 years Cytokines from plasma: IL-1β, IL-4, IL-5, IL-6, IL-7, IL-10, TNF-α, MIP-1β, IFNG, GMCSF Mannitol and Lactulose concentrations Alpha lipopolysaccharide Ferritin Vitamin D Retinol binding protein Activin A CRP Amino acids from plasma such as arginine, tyrosine, and tryptophan Plasma and stool metabolomics such as malate or sphingomyelins and serotonin or spermidine respectively Filamentous hemagglutinin, a virulence factor Vaccine responses from measles, haemophilus influenza, and diptheria Virus Indicators for Adenovirus, astrovirus, B. fragilis, C. difficile, C. jejuni/coli, Campy pan, E. histolytica, rotavirus, salmonella, and more
Environmental/Household	Income and Expenditure (in Taka) Principal flooring, roofing, and/or wall material Number of rooms in home Number of household members Number of people usually sleeping in home Type of fuel used for cooking Food availability, is there usually a deficit? If the toilet facility shared with other households Water drinking source How often the newspaper is read

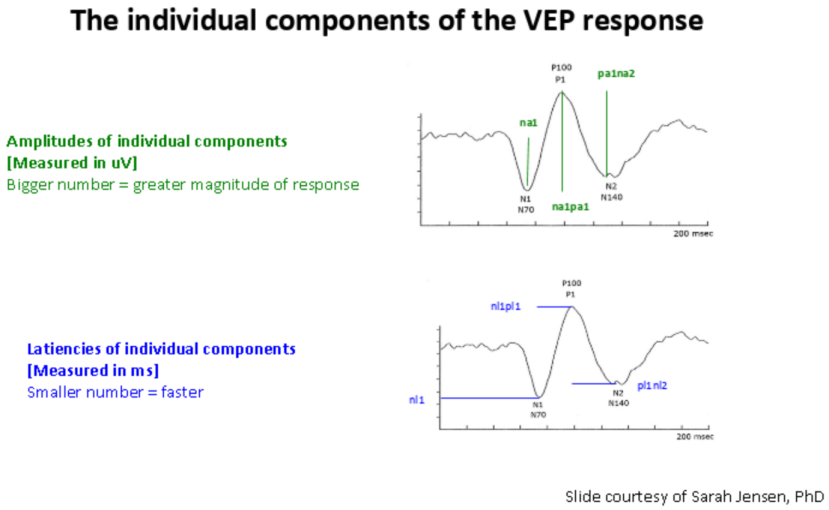

Figure 47: The individual components for the VEP outcome with measures of latency and amplitude.

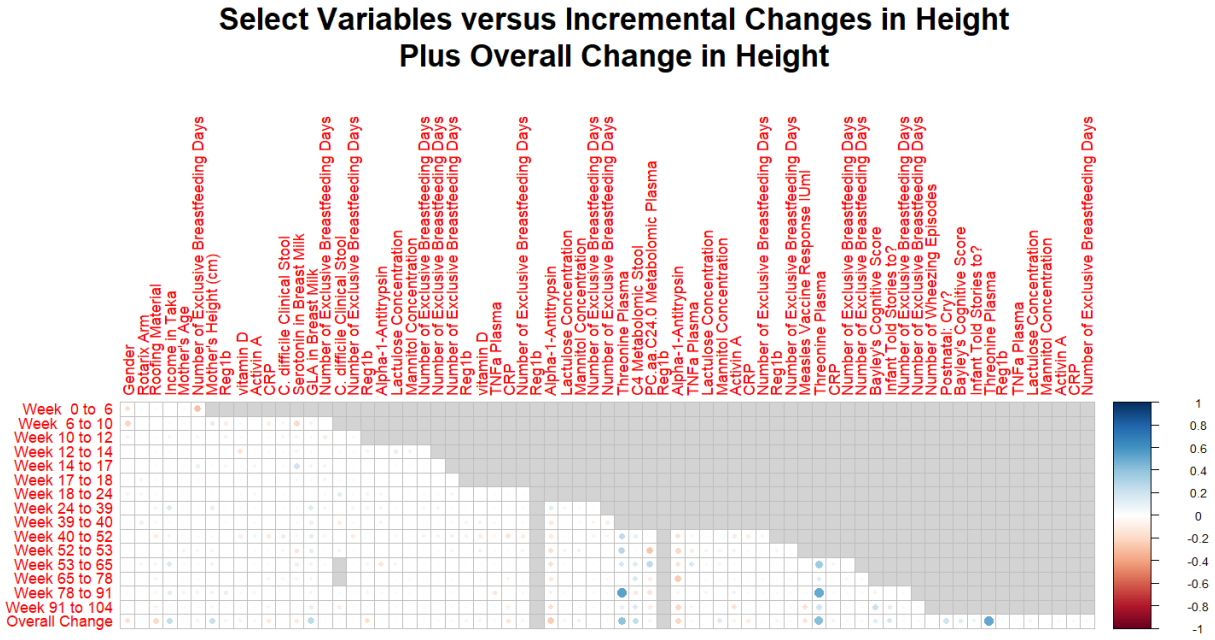

Figure 48: The Spearman correlation between selected variables at the time points and the incremental changes in height between each of the 16 time points.

the norm. The children involved in the study were healthy and living in conditions likely to allow them to reach their full potential. These standards are separated by sex since it is a known effect on growth. Multiple statistical techniques were used to create the height per-

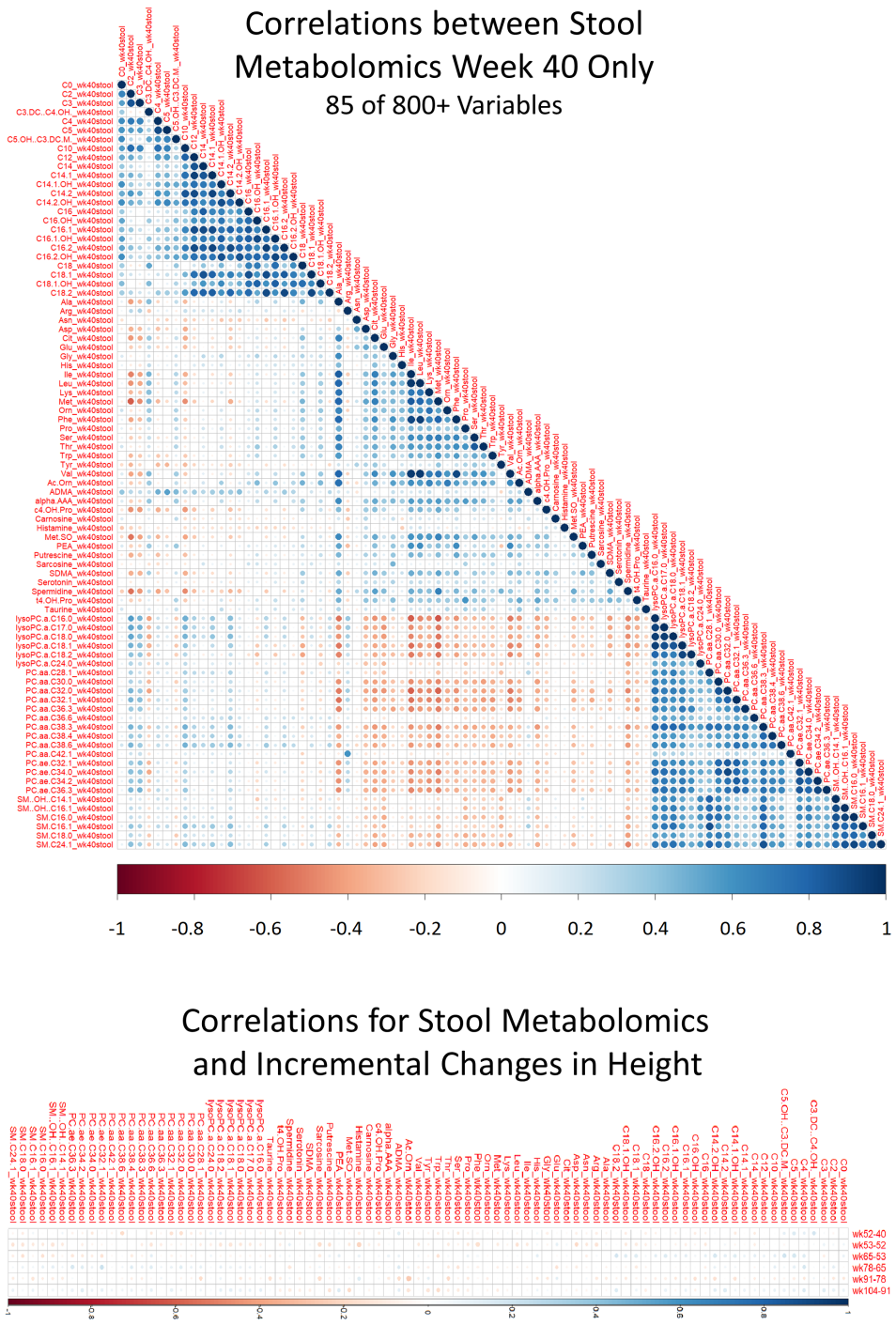

Figure 49: The Spearman correlation between the metabolomics from the stool of the infants at week 40 (left) and between these metabolomics and the incremental changes in height for each of the 16 time points (right).

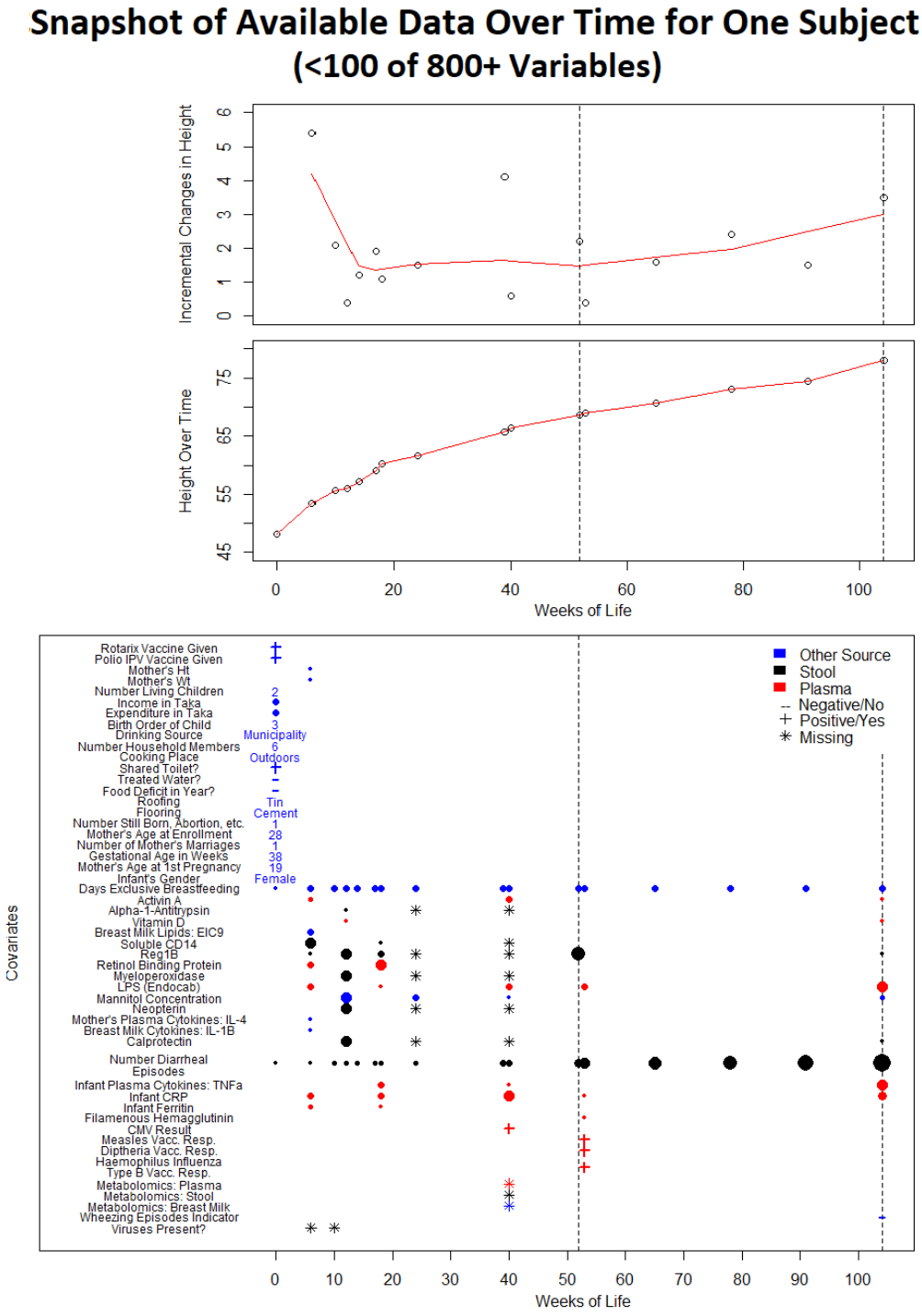

Figure 50: A selected female child within the lower quartile of overall change in height from the PROVIDE data. The color indicates the source of the measurements: black=stool, red=plasma, and blue=other. The $*$ indicates missing data for this subject and a $+$ or $-$ is the result from an indicator variable while the size of the point demonstrates the magnitude.

centile curves which are given as the standards and are seen on Figure 51 as the black lines whereas the colored lines (blue for males and pink for females) indicate the PROVIDE study individuals' heights over the first two years of life. One can see from this figure that at the two year mark, the majority of subjects in the PROVIDE study are already below the 50^{th} percentile mark by this WHO reference. This reference also gives individuals calculations to convert their data into the height-for-age (HAZ), weight-for-age (WAZ), weight-for-height (WHZ), and such which was completed within the PROVIDE study. As previously mentioned, a HAZ below -2 indicates that the child is stunted often indicating malnutrition [WHO MGRS Group, 2006]. These standards are the go to reference for healthy growth of children.

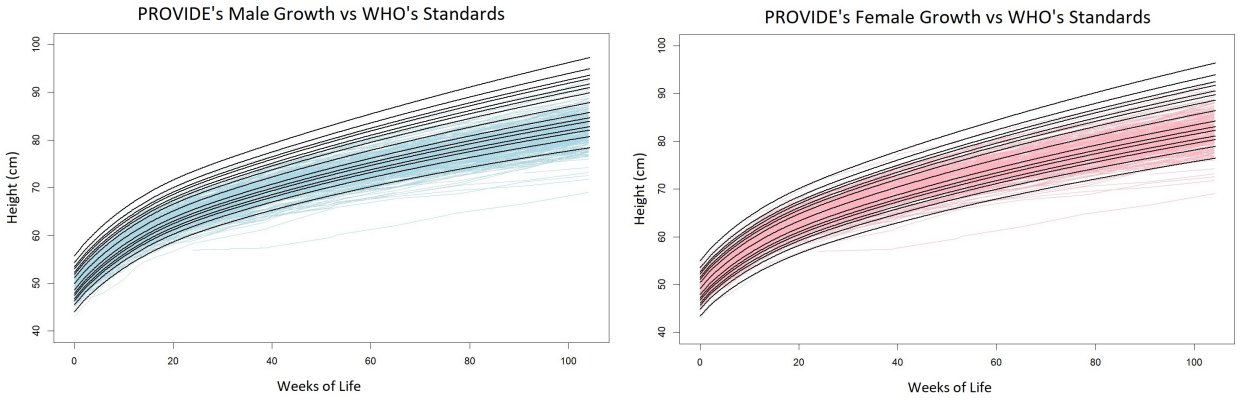

Figure 51: The light blue lines indicate the heights for the males within the study while the pink lines indicate the females' heights in the PROVIDE study. The black curves indicate the percentiles from the WHO standards for the respective sexes.

3.2 Current and Typical Growth Models

Many growth models have been created and explored. Among the different types polynomial models, Berkey-Reed, Jenss-Bayley, the Count model, the von Bertalanffy, Gompertz (in many different forms itself), logistic, and exponential are included. Several of these belong to the Unified Richards family of growth models including Unified-Gompertz and unified versions of the logistic and the von Bertalanffy [K. Tjørve and E. Tjørve, 2017]. Chirwa et al. explored the Berkey-Reed, Count model, Jenss-Bayley (and an adaptation thereof), and 2nd and 3rd order polynomial models for children in Africa aged 3 months to 10 years using Stata and SAS [Chirwa et al., 2014]. The findings suggest for modeling height that the adapted Jenss-Bayley and Berkey-Reed have similar and good performance. Another result showed the importance of sex within these models which is a widely known effect on height [Chirwa et al., 2014]. Polynomial models may be useful in certain situations such as the study by Troutman et al. where Excel was used [Troutman et al., 2018]. In that study, an age-specific growth model uses polynomial equations for weight and height split by sex and split further by gestational age. The preterm infants were shown to have a slower gain in weight and height than full term neonates, especially for those at the earliest gestational ages. Additionally, the catch up times for growth was explored where the females exhibited a faster catch up for lower gestational ages than males [Troutman et al., 2018]. An interesting approach to modeling growth involves a biased random walk. The resulting line resembles stair steps representing growth bursts where the growth increments come from a time varying distribution. With this model, each individual has their own curve in which height may be predicted [Suki and Frey, 2017]. Most of these methods can be used in conjunction with nonlinear mixed effects (NLME) which is described below and was proposed in this setting by Lindstrom and Bates [Lindstrom and Bates, 1990]. From these growth models mentioned, we will explore the Gompertz, logistic, and exponential in more detail.

The method put forth by Lindstrom and Bates has been widely used especially for growth modeling [Lindstrom and Bates, 1990]. The method is NLME models when repeated measures are the outcome. This method uses least squares estimators and maximum likelihood estimators (MLE) for the nonlinear fixed effects and the linear mixed effects respectively. In addition, a Newton-Raphson method is implemented for the estimation [Lindstrom and Bates, 1990]. The general NLME for the jth observation on the ith individual is given as:

$$y_{ij} = f(\boldsymbol{\phi}_i, \mathbf{x}_{ij}) + e_{ij}$$

where y_{ij} is the jth response on the ith individual, $\mathbf{x}_{ij}$ is the regressor vector for the jth response on the ith individual with no restrictions, f is a nonlinear function for the regressor vector and the vector of parameters $\boldsymbol{\phi}_i$ with length r and which is allowed to vary from individual to individual, and lastly e_{ij} is normally distributed random noise. The parameter vector may be included into the model as

$$\boldsymbol{\phi}_i = \mathbf{A}_i\boldsymbol{\beta} + \mathbf{B}_i\mathbf{b}_i$$

where $\mathbf{b}_i \sim N(\mathbf{0}, \sigma^2\mathbf{D})$,$\boldsymbol{\beta}$ is a fixed population parameters vector of length p, and $\mathbf{b}_i$ is the vector of length q for the random effects for individual i. Also, $\sigma^2\mathbf{D}$ is the covariance matrix. The $\mathbf{A}_i$ and $\mathbf{B}_i$ are design matrices respectively of sizes $r \times p$ and $r \times q$. These matrices can simplify the model. $\mathbf{A}_i$ may allow different groups to have varying fixed effects while $\mathbf{B}_i$ can give different groups separate random effects. The model for the ith individual may be written for the entire response vector as

$$\mathbf{y}_i = \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) + \mathbf{e}_i$$

where

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}, \mathbf{e}_i = \begin{bmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{bmatrix}, \text{ and } \boldsymbol{\eta}_i(\boldsymbol{\phi}_i) = \begin{bmatrix} f(\boldsymbol{\phi}_i, \mathbf{x}_{i1}) \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{i2}) \\ \vdots \\ f(\boldsymbol{\phi}_i, \mathbf{x}_{in_i}) \end{bmatrix}$$

and $\mathbf{e}_i \sim N(\mathbf{0}, \sigma\boldsymbol{\Lambda}_i)$ in which matrix $\boldsymbol{\Lambda}_i$ depends on i but only through its dimension. So, if one wants M individual models within one overall model, we must let

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix}, \boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \vdots \\ \boldsymbol{\phi}_M \end{bmatrix}, \text{ and } \boldsymbol{\eta}(\boldsymbol{\phi}) = \begin{bmatrix} \boldsymbol{\eta}_1(\boldsymbol{\phi}_1) \\ \boldsymbol{\eta}_2(\boldsymbol{\phi}_2) \\ \vdots \\ \boldsymbol{\eta}_M(\boldsymbol{\phi}_M) \end{bmatrix}.$$

In addition, we also will have $\tilde{\mathbf{D}} = diag(\mathbf{D}, \mathbf{D}, \ldots, \mathbf{D})$ and $\boldsymbol{\Lambda} = diag(\boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \ldots, \boldsymbol{\Lambda}_M)$. Hence the overall model is

$$\mathbf{y}|\mathbf{b} \sim N(\boldsymbol{\eta}(\boldsymbol{\phi}), \sigma^2\boldsymbol{\Lambda})$$

with $\boldsymbol{\phi} = \mathbf{A}\boldsymbol{\beta} + \mathbf{B}\mathbf{b}$, $\mathbf{b} \sim N(\mathbf{0}, \sigma^2\tilde{\mathbf{D}})$, and $\mathbf{B} = diag(\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M)$. We also have

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_M \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_M \end{bmatrix}.$$

Since the model is specified, the next step would be to estimate the parameters. For the estimation of $\boldsymbol{\beta}$ and $\mathbf{b}$, a least squares problem is created by augmenting the data with "pseudo-data" and then the MLE are found in both the linear and nonlinear cases. Then, a two-step algorithm is followed with the first step being the pseudo-data (PD) step where a nonlinear least squares estimation is completed and the second being the linear mixed

effects (LME) step where Newton-Raphson may be used. The method completes this two-step algorithm until convergence. To begin using the method and in particular for the first PD step, one must specify starting or initial values for the parameters. Usually these can be inferred from the data. A common example from Lindstrom and Bates is of orange tree growth in terms of circumference on five trees and is given using the logistic model here:

$$y_{ij} = \frac{\beta_1 + b_{i1}}{1 + \beta_2 e^{\beta_3 x_{ij}}} + e_{ij}$$

where x_{ij} are the time points the circumferences were measured, e_{ij} are iid $N(0, \sigma^2)$, and $b_i \sim N(0, \sigma^2 \mathbf{D})$ and $\mathbf{D}$ is a 1×1 matrix [Lindstrom and Bates, 1990]. Starting values are given to the method and after convergence, the parameter estimates are shown and the curves are plotted from the restricted MLE estimates. Note that since the circumferences were gathered over time, it's natural to think there may be serial correlation for each individual's measurements. This is not included within the noise term in the model, but is it assumed to be taken care of within the random effects structure [Lindstrom and Bates, 1990]. This method has been implemented in numerous cases, but has a great flexibility which leads individuals to apply NLME.

K. Tjørve and E. Tjørve (2017) and Chriwas et al. (2014) both advocate the use of NLME for modeling height and weight using covariates [K. Tjørve and E. Tjørve, 2017] [Chirwa et al., 2014]. The reasoning these authors among others promote this modeling technique is due to the incredible flexibility since it can deal with irregularly spaced times and missing data as well as model on the individual and population levels. Usually, the fixed effect within the model represents the mean structure or general population curve and the random effect allows for individual variations in growth. Another highlight of NLME is the availability of comparison between models making it easier to select a proper growth model along with the addition of covariates. Most of the previous studies involving low and

middle income countries have not used NLME for modeling height up to two years of age since longitudinal data is not readily available in these settings [Chirwa et al., 2014].

The Gompertz growth model is very common and may be used to model the growth for various biological beings such as plants, birds, fish, mammals, tumors, bacteria, and survival of cancer patients. Perhaps this model is so widely used due to its age. It was first suggested as a probability density function in 1825 where Makeham stated the model in the more common cumulative form. Thus, this model has a long history of multiple uses including from insurance for mortality. There are multiple versions of the Gompertz including a three-parameter or four-parameter version, the Zwietering modification (modified Gompertz), the Zweifel and Lasker re-parameterisation, the Gompertz-Laird, and the Unified-Gompertz. For most of these versions, the growth parameter values will not be directly comparable to growth coefficients from other methods and are often difficult to interpret. Thus, the following is one of the two models which are the Unified-Richards [K. Tjørve and E. Tjørve, 2017]. This particular model is the W_0 form where the W_0 is the initial value (height at birth):

$$W(t) = A_U \left(1 + \left(\left(\frac{W_0}{A_U}\right)^{1-d} - 1\right) \cdot \exp\left(\frac{-k \cdot t}{d^{d/(1-d)}}\right)\right)^{1/(1-d)}.$$

A_U denotes the upper asymptote (the highest possible height), d is the parameter that shifts the inflection value, t is time, and k represents the relative growth rate. Taking this Unified-Richards model, we can obtain a unified version of the logistic model by setting $d = 2$. The W_0 form of the Unified-Gompertz is then:

$$W(t) = A_U \left(\frac{A_U}{W_0}\right)^{-\exp(-e \cdot k \cdot t)}.$$

W_0, A_U, and t all have the same meaning, however k now becomes the maximum relative growth rate. So the absolute growth rate would then be $A_U \cdot k$. Note that in the original paper by K. Tjørve and E. Tjørve, this model showed a negative trend when plotted whereas when a negative sign was added as in the current representation in this research to the

first exponential the growth curve then became positive [K. Tjørve and E. Tjørve, 2017]. Switching the A_U and W_0 in the fraction also gives the same effect and values. This Unified-Gompertz has an inflection point which is set at 36.8% of the upper asymptote and is calculated by A_U/e whereas for the Unified-Richards, the inflection point is calculated by $A_U/d^{1/(1-d)}$ [K. Tjørve and E. Tjørve, 2017]. Thus, this model is a good alternative to other versions due to the interpretability of the parameters.

In addition to Gompertz, the logistic and exponential models will be explored for comparison since they are general and simpler to implement. First, the logistic growth model may go by several other names (and may be connected as above to the Richards family of growth curves), but has the simple S-shape which is commonly known. In this case, we will need a starting point, or lower asymptote which is nonzero. Thus, the following formula provides the logistic growth curve

$$h(t) = A_L + \frac{A_U - A_L}{1 + e^{(k-t)/\delta}}$$

where A_L is the value for the lower asymptote (here this would be the starting height), A_U is the upper asymptote, k is the growth rate, t is time, and δ is a shape parameter which determines the steepness of the curve. Similar to the formula for the logistic, the exponential growth curve may be written as

$$h(t) = A_U - (A_U - A_L)e^{-(kt)}$$

where again, A_U is the upper asymptote, A_L is the lower asymptote, k is the growth rate, and t is the time. To have similar notation, we may also write the Gompertz curve as

$$h(t) = A_L + (A_U - A_L)e^{-\exp(-k(t-I))}$$

where I is the inflection point which can be set as previously described ($I = A_U/e$) or by

the formula $I = (A_U - A_L)/e$ [Henderson and Seaby, 2006]. Using the previous formula for Gompertz and this formula with the two different inflection points gives Figure 52 for a child whose starting height would be 45 cm and reaches 85 cm over the two years. For our comparisons, it seems as though that none of the representations deem a great fit to the PROVIDE data.

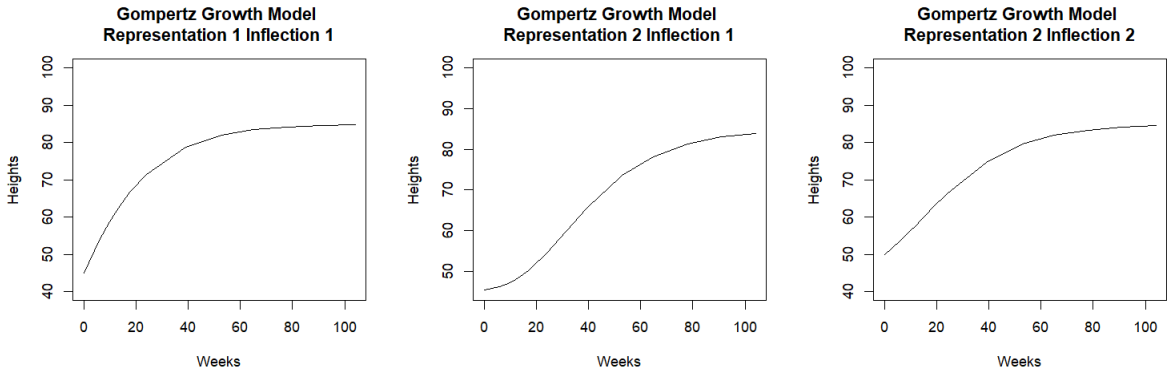

Figure 52: The first plot corresponds to the representation by K. Tjørve and E. Tjørve (2017) while the other two correspond to the representation featured in Henderson and Seaby's (2006) Growth II program with two separate inflection points.

In general, these nonlinear models in Figure 53 were created since they could be estimated with certain simplicities. However, we need a model which is a bit more specific and has a changing growth rate which may depend on several covariates. This need is displayed by Figure 53 since none of them fully capture the characteristics of the data.

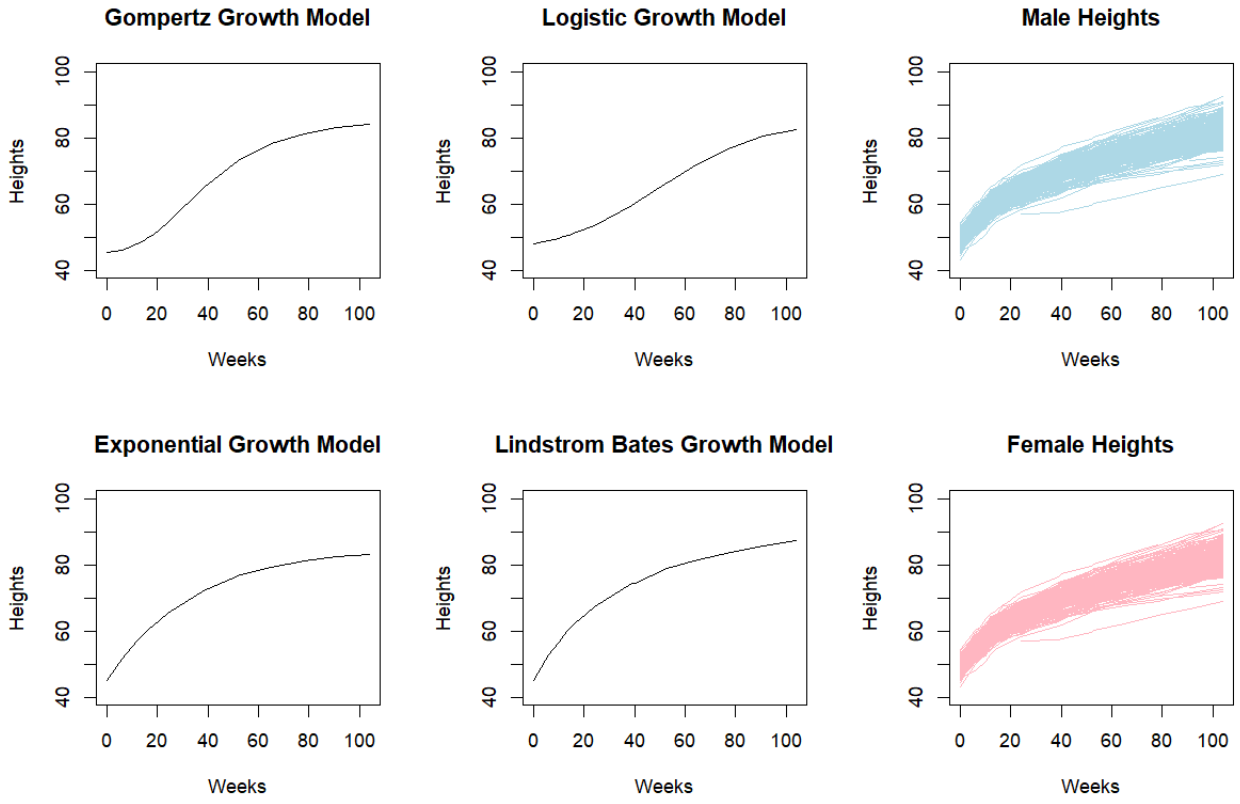

Figure 53: The first two columns of plots are currently used growth models on the same scales as the PROVIDE data but for a given subject who was born at 45 cm and will reach a max of 85 cm over the two years. The last column of plots are the actual heights for males (light blue) and females (pink) from the PROVIDE data.

3.3 Proposed Growth Model for PROVIDE Cohort

The model we are proposing to use for the PROVIDE cohort is unlike the others explored above. This model will be specific for the individuals in Bangladesh and has a goal to incorporate multiple covariates in order to predict height and apply interventions before stunting occurs. Figure 54 shows an example of the proposed model's shape for a particular individual with no missing data. For this example, no covariates have been added into the model.

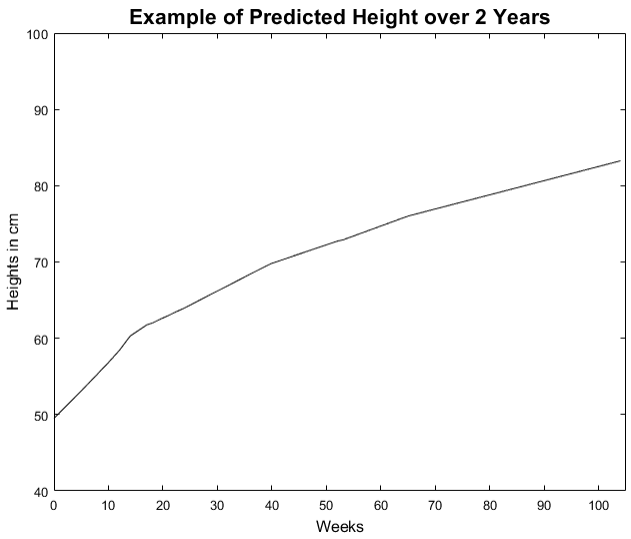

Figure 54: The graph displays an example of the proposed model for a particular subject with no missing data without covariates added.

The proposed model is derived as follows:

$$
\begin{aligned}
h_t &= h_0 \prod_{j=1}^{t} (1 + \delta_j \Delta_j) \\
&= h_0 \exp \sum_{j=1}^{t} \ln(1 + \delta_j \Delta_j) \qquad \text{Let } \ln(1 + \delta_j \Delta_j) \approx \delta_j \Delta_j \\
&= h_0 \exp \sum_{j=1}^{t} \delta_j \Delta_j \\
&\approx h_0 \exp \int_0^t \delta_s ds \\
&= h_0 \exp (g_t t)
\end{aligned}
$$

where Δ_j are the local time units since data was collected irregularly, δ_j are the growth rates per each time point, and h_0 is the height at birth.

The average growth rate up to time t is $g_t = \frac{1}{t} \int_0^t \delta_s ds$. Thus, as $t \to \infty$, $g_t t \to c$ where c is the person's adult height and so $g_t \to 0$. Random variation should also be added to the

model due to individual growth rates in which we would have:

$$h_t = h_0 \exp\left(g_t t + \sigma_t Y_t\right).$$

This is similar to financial models for stock prices where Y_t is Brownian motion in which we can take advantage of their estimation methods to help build the model. Additionally, the Ornstein-Uhlenbeck process is a continuous analog of a discrete AR(1) time series. Using the integrated version, we can have a basis on which to build. One may also add in the covariates to the $g_t t$ term as $g_t t + \mathbf{X}$ where $\mathbf{X}$ would be the covariates selected to help explain the growth of these children.

3.3.1 Constraints and Regularization

Due to the nature of the data, there will be some natural constraints that must occur. For example, the starting predicted value should be the same as the initial height. Thus, the first growth rate at time zero should also be zero. Since humans do not lose height specifically within the first two years of life, another natural constraint is that the growth rates are nonnegative. Another thought was the dependence between growth rates. From Figure 55, we see that there is very little correlation (≈ 0.2 or less) between the incremental changes in height which may show reason against our initial instinct that the growth rates would be dependent on each other. However, these incremental changes are not normalized which may lead to the lack of associations.

Possible regularizations could be applied to the model as well. Usually, regularizations are easily applicable and often penalize the coefficients with some methods assigning values of zero to unimportant variables. However, in our case, we not only have numerous covariates (900+), but they are measured at various times with some at multiple time points over the two years. Most of the individuals also do not have all the possible measurements per each variable meaning some individuals have very sparse data. In addition, we do not have just

one particular outcome, we have multiple due to the measures, again, being over time. This adds numerous complexities and is unlike most problems where regularization is applied. Hence, a new regularization technique would give the following conjecture.

Conjecture 3.1. *For time point t where $t = 1, \ldots, T$, a matrix of covariates $\mathbf{M}_t$ with size $p \times n$ have been measured across individuals $i = 1, \ldots, n$ along with a $1 \times n$ vector of incremental changes in height $\mathbf{D}_t$ for the time point t across the individuals. The first value of the incremental changes vector at time $t = 1$ will effectively be zero due to this time point being the starting time and thus growth has not yet occurred. In order to explain these incremental changes in height per each time point, a model must be specified.*

For each time t, the following will be optimized over all subjects:

$$min \sum_{i=1}^{n} (D_{t,i} - \hat{\boldsymbol{\beta}}_{t,i} \mathbf{M}_{t,i})^2 + \lambda \left(\|\hat{\boldsymbol{\beta}}_{t,i}\|^2 \right)$$

with $\lambda \geq 0$ and where $D_{t,i}$ is the incremental change at time t for individual i, $\hat{\boldsymbol{\beta}}_{t,i}$ is the $1 \times p$ vector of coefficients for the corresponding $\mathbf{M}_{t,i}$ which represents the $p \times 1$ vector of covariates at time t per individual i.

Thus, the second term, $\lambda \left(\|\hat{\boldsymbol{\beta}}_{t,i}\|^2 \right)$, will allow for regularization of these coefficients at time t therefore reducing the number of covariates needed at each time point.

This Conjecture may also be connected to the variable importance (VIMP) described in 3.4.3. One might choose to instead use VIMP per each time point t in order to select from the numerous covariates where a value above zero with random forest (RF) or conditional random forest (CRF) would indicate a predictor that has an effect on the outcome. Another way to use the VIMP would be to regularize on the VIMP values which is similar to wavelet thresholding.

These frameworks proposed here could lead to a fantastic model, but remain unimplemented in this current research as the selection of covariates was more thoroughly explored.

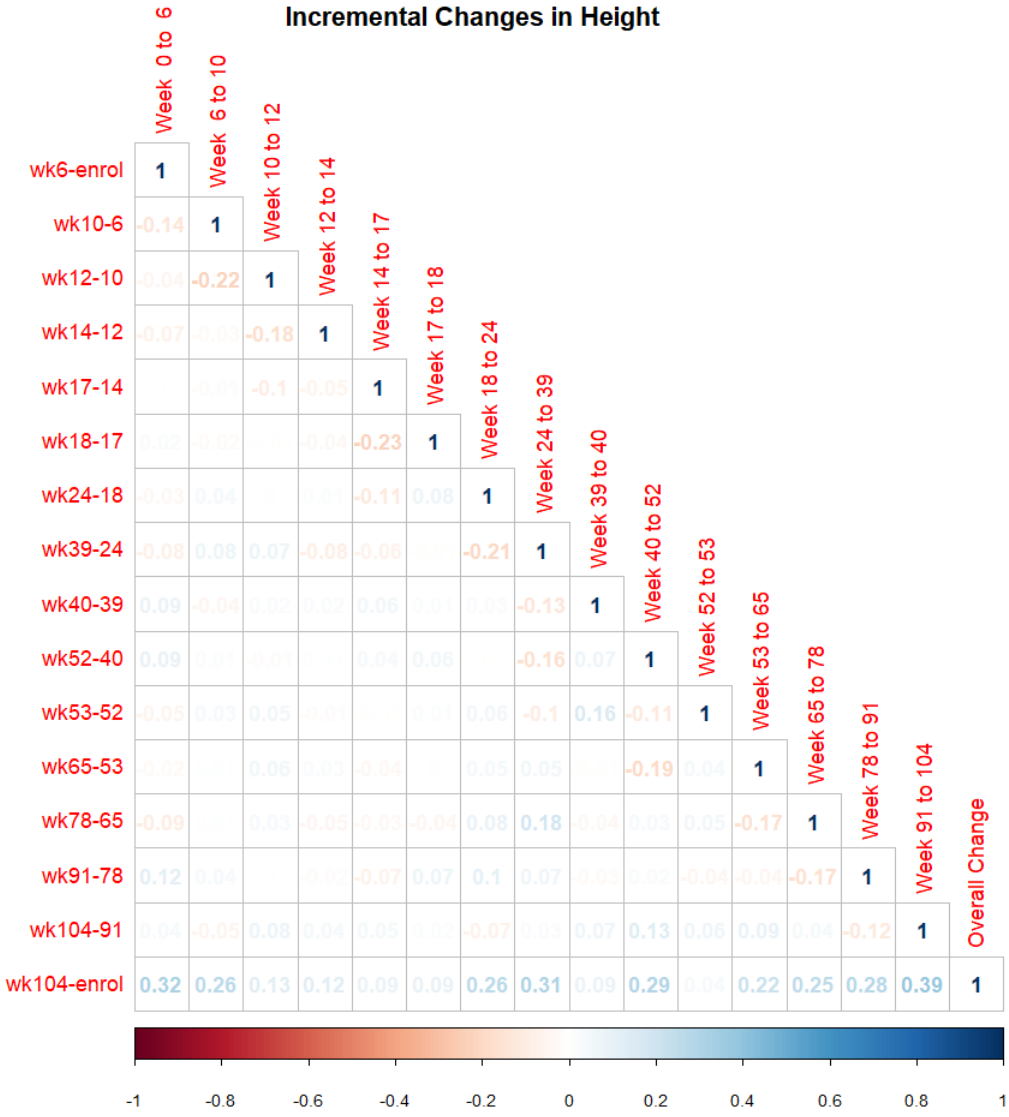

Figure 55: The Spearman correlations between the incremental changes in height over the time points.

3.4 Methods of Exploring and Selecting Covariates

3.4.1 Simple and Exploratory Methods

Due to the shear number of covariates available to predict height and other outcomes from the PROVIDE cohort, some preliminary exploratory analyses had to be completed. With a subset of this data and with only 47 predictors included, the conditional variable importance (VIMP) was previously used to explore which variables may predict the height-for-age z-scores (HAZ) [Donowitz et al., 2018] (coauthor). The results lead to mother's weight and height, initial HAZ value, income, and a few biomarkers being deemed important (ranked within the top 10 in terms of conditional VIMP) as seen in Figure 56. Specifically, mother's weight and the initial HAZ of the child were the two most important predictors of HAZ at two years of age [Donowitz et al., 2018] (coauthor). These results have since then led to additional funding where nutritional interventions in mothers are being studied.

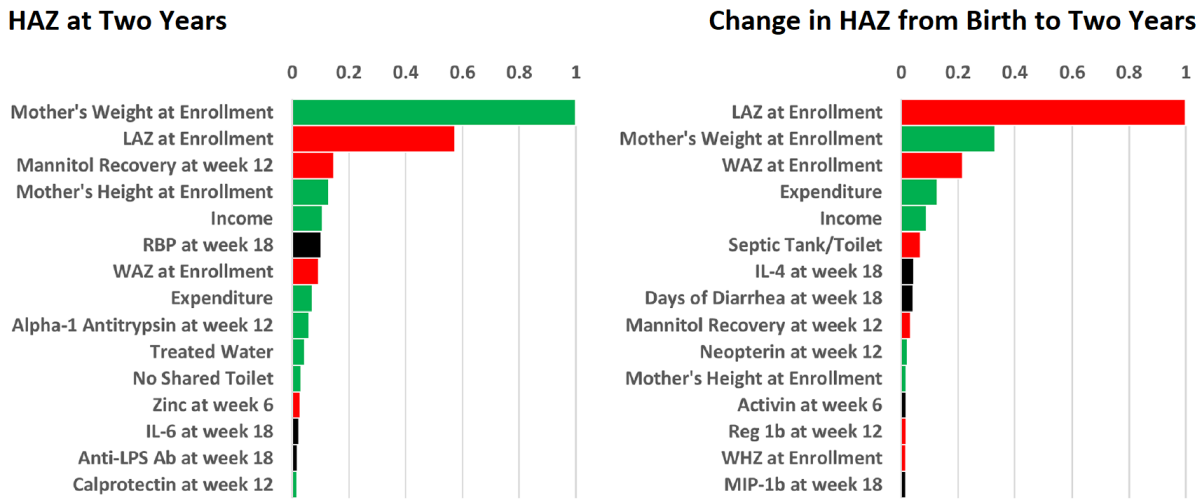

Figure 56: The variable importance measures are based on a conditional random forest model with the conditional variable importance calculated for the PROVIDE data. Only the top 15 variables are shown [Donowitz et al., 2018].

In addition to this previous analysis, a simple look through graphical relationships between all possible variables and height and change in height over two years. These simple

graphs give a first look and general idea to which predictors may show up in the coming methods. Another simple way the relationships were viewed included graphical measures such as Figure 50 which again shows the sparsity of the data. Additionally, Pearson and Spearman correlations were computed along with the results from a test of the correlation being significantly different from zero for the variables at each time point and the next time's incremental height, height, HAZ, and stunted indicator. For the most part, these two types of correlation measures overlapped fairly well with the exception of the incremental changes in height outcome where the two methods showed slightly different variables being correlated with this particular outcome. These results show that the incremental changes are more difficult to analyze especially since they have not been normalized. However, these simple calculations only give a first look at which variables may help predict growth and thus we still have the goal of explaining the growth patterns while exploring the effects of covariates at certain times in which these first looks are valuable to give intuition.

3.4.2 Penalized Linear Models with L1 and L2 Regularizations

A next step involves somewhat simple manners which explore the selection of covariates through L1 and L2 regularization, also known as lasso (least absolute shrinkage and selection operator) and ridge regression respectively. Both of these techniques are linear regression but with a penalty term added to the loss function of $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The penalty term penalizes the parameters in the model, here the coefficients of the variables. These methods may be used to avoid overfitting the data or for the selection of variables. For lasso regression, the penalty takes the form of the absolute value of the coefficients:

$$SSE + \lambda \sum_{j=1}^{P} |\beta_j|.$$

The penalty term for ridge regression is the sum of the squared coefficients:

$$SSE + \lambda \sum_{j=1}^{P} \beta_j^2.$$

In both cases, P is the number of predictors and β_j is the coefficient for predictor j. If $\lambda = 0$, then we have the ordinary least squares regression model. However, if λ is large, the model may underfit the data with too much penalization. As λ increases in lasso regression, coefficients will become zero, effectively selecting some variables (those with nonzero coefficients), however in ridge regression, the coefficients will be shrunk to zero but not equal to absolute zero. Thus ridge regression is not a feature selection method. These implications lead to ridge shrinking the coefficients of correlated covariates towards the same value while lasso will choose one of the correlated variables and ignore the rest. These two penalties may be combined to create the elastic net where the penalty with the loss becomes:

$$SSE + \lambda_1 \sum_{j=1}^{P} |\beta_j| + \lambda_2 \sum_{j=1}^{P} \beta_j^2.$$

This method will thus incorporate feature selection along with the shrinkage of the coefficients which has been suggested to deal better with groups of highly correlated covariates which is the case here. All three of these methods do require tuning of the λ parameters in order to achieve the best performance [Kuhn and Johnson, 2016].

For logistic regression, the ideas are similar however the loss function is the binomial likelihood function $L(p) = \binom{m}{k} p^k (1-p)^{m-k}$ where m is the number of trials, k is the number of successes, and p is the probability of success. The ridge regression like penalty, $\log L(p) - \lambda \sum_{j=1}^{P} \beta_j^2$, may stabilize the coefficients and help with correlated covariates. Similarly, a lasso like penalty may be used. The elastic net may also be implemented, however slightly differently:

$$\log L(p) - \lambda \left((1-\alpha) \frac{1}{2} \sum_{j=1}^{P} \beta_j^2 + \alpha \sum_{j=1}^{P} |\beta_j| \right).$$

In this case, α is the amount the two penalties are mixed together. If $\alpha = 1$ it is a lasso penalty whereas if $\alpha = 0$, then it is a ridge penalty. The λ still controls the amount of penalization [Kuhn and Johnson, 2016].

In the implementation of these models, the *glmnet* R package was used with the *cv.glmnet()* function which implements cross validation for the tuning of the λ parameter. Within this R function, the specific formula optimized for regression and for logistic regression respectively are

$$\frac{1/2 \cdot SSE}{n} + \lambda \cdot penalty$$
$$\frac{-\log L(p)}{n} + \lambda \cdot penalty$$

where $penalty = (1-\alpha)\frac{1}{2}\sum_{j=1}^{P}\beta_j^2 + \alpha\sum_{j=1}^{P}|\beta_j|$.

Since ridge regression does not effectively select variables, the analyses here focused on lasso and elastic net, although ridge regression was run in case the direction of a predictor's linear relationship with an outcome was needed. The outcomes assessed were the incremental heights, heights, HAZ, and the stunting indicator at the next time point. All predictors per each time point were used to predict the next outcome while in a separate round of analyses, the previous predictions were also used for the current time point as an additional predictor in order to improve the amount of information included within these models per time point. Then height, HAZ, and Bayley's at 2 years, Mullen's at 3 years, and Weschler's at 4 and 5 years were modeled with these methods as well.

3.4.3 Random Forests and Conditional Random Forests

Bagging (bootstrap aggregation) is an ensemble method that aggregates decision trees generated from bootstrapped samples as explained in Figure 57. Each of the trees is grown deep giving each tree low bias but high variability which is reduced when averaged. However there are no tuning parameters leading to overfitting issues. For a binary outcome, each tree can be thought of as casting a vote for which category that tree thinks the new observation should belong. The total number of votes for each category is then divided by the total number of

trees to produce the predicted probability for a new observation [Kuhn and Johnson, 2016]. These predicted probabilities can then be used to classify the new observation based on a decision threshold, which can be naively chosen as 0.5. One general downfall to aggregating trees is the loss of interpretability with a specific downfall for bagged trees being that the trees are not completely independent of one another. This is due to all of the predictors being considered at each split for every tree which leads to tree correlation. Tree correlation may prevent the method from optimally reducing the variance of the predictions since each tree can have comparable structures.

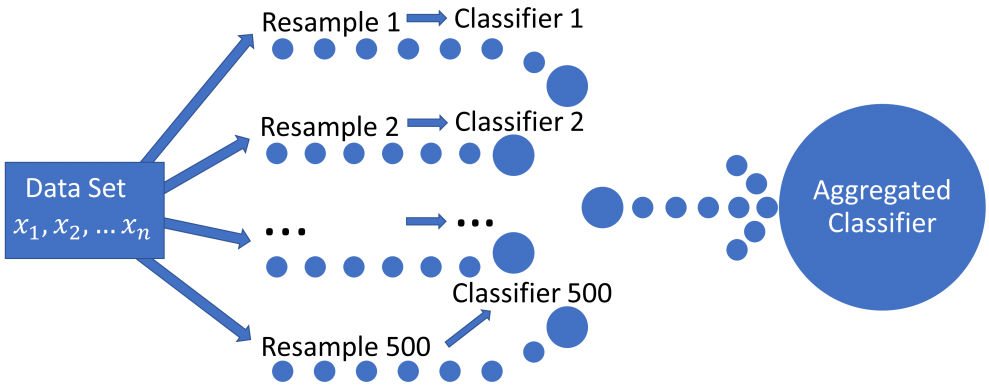

Figure 57: Flow chart of the general steps bagging, random forests, and conditional random forests follow.

Random forests (RF) is an improvement from bagged trees and is where the trees are built on bootstrapped samples (similar to bagged trees) but every time a split is considered, a random sample of predictors is chosen as split candidates which breaks this tree correlation. The split then only uses one of the sampled predictors at that decision node. Each tree is grown to the maximal depth and contributes equally to the final model, thus if the number of randomly sampled predictors equals the total number of predictors, RF becomes bagging. Due to the lessening of tree correlation, the number of trees created does not attribute much

to overfitting [Kuhn and Johnson, 2016]. When predictors are highly correlated, the variable importance may be overestimated making correlated predictors appear more important than uncorrelated ones [Strobl et al., 2008] [Boulestix et al., 2012] [Strobl et al., 2009]. The conditional random forests (CRF) and conditional variable importance takes into account these correlations to reflect the impact of a single variable in predicting the outcome. Thus, CRF uses an unbiased splitting criteria to avoid such issues. Both of these methods are nonparametric and do not make any assumptions about the data structure. Therefore, these models may include nonlinear relationships as well as interactions between predictors. While this is a great advantage, we are also losing the ability to peek inside of the model which would allow us to know the relationships within. A tuning parameter for both RF and CRF is the number of predictors to be randomly chosen at each decision node. It is important to tune this parameter since a small number of predictors chosen at each split can lead to choosing variables that are suboptimal and can lead to a loss of information [Boulestix et al., 2012] [Strobl et al., 2008]. In any situation, we would like to have informative predictors used to get the best predictions.

In terms of the variable importance (VIMP) for RF and CRF, the calculations are fairly simple. Since each tree in these methods is grown from a bootstrap sample, there are some subjects or cases that are left out, usually called out-of-bag (OOB). These samples may be used to assess the importance of a predictor in terms of how much a metric changes after the values for a certain predictor are shuffled. For example, in the regression setting, the mean squared error (MSE) is often used in some manner to assess a model. For RF, the MSE per tree is calculated using the OOB data, then each variable's values are shuffled one at a time and the MSE recalculated. The difference in the new MSE and the original MSE is found for each variable. Then, these differences per tree are averaged to get the whole forest's VIMP. In the classification case, the mean decrease in accuracy may be applied instead of the mean percent increase of MSE. These are often referred to as the permutation based VIMP metric within RF and it's the only type we will consider for the following analyses.

For the conditional random forest, one may consider the use of conditional VIMP in which a conditioning grid is create and the values of a certain variables are only permuted within the particular groups or values of the conditioning grid which depends on which variables the certain predictor is correlated.

Thus, two models (RF and CRF) with respectively the permutation based VIMP and the conditional VIMP are used for the incremental heights, heights, HAZ, and stunting indicator at each time point independently and with the predictions from the previous time point added to the current set of predictors. Additionally, the height, HAZ, Bayley's, Mullen's, and Weschler's were assessed with these methods. The VIMP measures were used to select variables with the cutoff being zero since a VIMP of zero or below indicates the variable is not helpful to the prediction of the outcome.

3.4.4 Deep Learning

In addition to more traditional methods such as penalized models and other but simpler machine learning techniques like RF, deep learning models were explored. Deep learning is a subfield of machine learning which emphasizes learning successive layers of information with increasingly meaningful representations of the data. The depth of the model refers to the number of layers used. Usually, these layers are neural networks which mimic the learning patterns of humans' natural biological neural networks. These networks consist of hidden units or neurons which are linear combinations of the original predictors which are not constrained. Thus there are weights per unit of the neural network, which in the deep learning setting, one could have many weights per each layer of the model. To optimize these weights, which are the parameters of the model, a backpropegation strategy is implemented where the weights are initially assigned to be random values. Thus, the loss score is very high but reduces with every case given to the model such that the weights are adjusted in the correct direction. Eventually, after many iterations over thousands of examples

(cases), the algorithm will output a model which optimizes the loss function by the weights [Chollet and Allaire, 2018].

During the explorations of these models, poor performance was found of the predictions for the testing set. This poor performance would lead to either the under and/or over estimates of children's growth and development which makes it difficult to say which interventions could be effectively applied. The other difficulty that arises with the deep learning aspect is the lack of samples in terms of the number of subjects for a regression or classification setting, at most 700 subjects with numerous missing values, and in terms of time points for the time series setting, at most 25 time points of heights with again several missing values. Having these low number of subjects or time points does not allow the method to effectively learn about the trends within this data in either type of setting. Therefore, the results from these types of models are not presented nor further examined.

3.5 Results for Models Predicting the Next Time Point's Outcomes

Within 3.5, the PROVIDE data was explored per time point taking each set of predictors available at a particular time point and predicting the next incremental change in height, height, HAZ, and whether or not the subject is stunted (HAZ < -2). For example, all the available covariates at week 6 were used to predict the change in height from week 6 to week 10, the height at week 10, HAZ at week 10, and stunting at week 10. Therefore, the largest amount of data is included at enrollment due to dropouts and missed clinic visits or measurements as time goes on in which the sample size per time point decreases. Due to this diminishing set of subjects, predictors were scanned before analysis to make sure an ample amount of subjects were included to select predictors and explore relationships while other subsets of data were created and separately analyzed such as the metabolomics from the stool samples at week 40. In addition to the data being explored independently per time point, the previous model's predictions were used as a new covariate for the next time point in the hopes that previous information would improve the next model's prediction. These results are summarized and then used to build the models in 3.6.

In general, all four methods' (lasso, elastic net, RF, and CRF) selection of variables overlapped fairly well, often with the RF or CRF selecting more predictors than the penalized linear regressions which often picked less than ten predictors. The correlations for the outcomes and the predictors per time point also corroborated with these models' results, although many more predictors had significant correlations (without adjusting for multiple comparisons since they were pairwise complete correlations) than what were selected by the different methods of models. However, these initial correlations gave way to the ideas of the types of predictors which may be selected from these other analyses. A non-exhaustive list of covariates over all the various time points over two years is included in Table 7. These

variables were then used in 3.6 for the creation of models predicting outcomes at two years up to five years.

Due to the lack of sample size for particular subsets of predictors, the metabolomics and TAC data from stool samples (which included information about which viruses were present) were analyzed in this first initial analysis, but were unable to be included for the models in 3.6. The TAC data did not choose many predictors throughout all four methods. However, we can compare the results from the metabolomics from blood samples at week 40 with Moreau et al. who used CRF for these plasma metabolomics plus a set at 36 months but arbitrarily chose the top 15 variables per outcome [Moreau et al., 2019] (coauthor). This article considered HAZ at 4 years, the change in HAZ from enrollment to 4 years, and Weschler's at 4 years while only height, HAZ, and stunting at the next time point and the incremental change in height were considered by the analyses in this research with this set of predictors. However, there is some overlap between these two analyses. In particular, the phosphatidylcholine (PC) species with either an ester (aa) or ether (ae) bond for the fatty acid chains that overlap between the HAZ at 4 year or change in HAZ from enrollment to 4 years outcomes (article) and next time point HAZ or next time point's incremental change in height (this research) are: PC aa C32:3, PC ae C34:2, PC aa C32:3, PC ae C34:1, and PC aa C36:2. The numbers in these metabolites indicate the total carbon chain length (x) followed by the number of double bonds (y) in the fatty acids (x:y). There are also a few sphingomyelin (SM) metabolites, specifically those that are hydroxy-sphingomyelins (SM-OH) which corroborate between the analyses such as SM-OH C14:1, SM-OH C22:1, and Total SM-OH which is the sum of these SM-OH metabolites. Lastly, acyclcarnitines chosen between the two include C6 (C4:1-DC), C12, and C14:1-OH. Therefore, there is a fair amount of overlap and one can say that these metabolites not only have a short term (12 weeks) prediction ability but a long term (3.25 years) ability.

In terms of performance, the R-squared values were calculated per time point while predicting the next time point's outcomes as in Figure 58. One must note that the R-

Table 7: A Short List of Selected Covariates from Lasso, Elastic Net, Random Forest, and Conditional Random Forest over Two Years of Life Predicting the Next Time's Outcome.

Type/Source	**Variables**
Maternal/Paternal	Mother's Height and Weight Mother's age at enrollment, at first marriage, and at first pregnancy Mother's number of living children, marriages, and number of stillborns, death, abortions, etc. Vitamin supplement types taken by mother during pregnancy Father's and mother's education levels and occupations MFGE8 and EGF from mother's breast milk
Week 6	Number of cumulative episodes of diarrhea Number of exclusive breastfeeding days LM ratio (lactulose to mannitol) Calprotectin Neopterin Alpha lipopolysaccharide (LPS) Retinol binding protein (RBP) Myeloperoxidase (MPO) Ferritin Vitamin D Activin A Reg 1B sCD14 CRP Cytokines from plasma: IL-1β, IL-4, IL-6, IL-10, TNF-α, etc. Filamentous hemagglutinin, a virulence factor Vaccine responses for measles, pertussis, diptheria, tetanus, etc.
Environmental/Household	Income and Expenditure (in Taka) Principal flooring, roofing, and/or wall material Number of rooms in home Number of household members Number of people usually sleeping in home Type of fuel used for cooking Food availability, is there usually a deficit? If the toilet facility shared with other households Water drinking source Do they own a clock?

squared values tend to be higher for predicting outcomes of time points that are closer to when the variables were collected especially since only the data at a certain time point is being used to predict the next time point's outcomes. For instance, predicting week 18 outcomes using week 17 data would be much easier than predicting week 39 outcomes from week 24 data. Additionally, the breast milk metabolomics subset, clinical stool TAC subset, and viral shedding stool TAC subset at week 6 were also explored, but the R-squared values were all essentially zero for each outcome of incremental change in height from week 6 to week 10, height at week 10, HAZ at week 10, and stunted height at week 10 when lasso and elastic net were used. Similarly for the TAC data from clinical stools and viral shedding stools taken at week 10, the R-squared values were basically zero for predicting the next time's outcomes at week 12. The subset of stool and blood metabolomics sampled at week 40 gave zero R-squared values except for the outcome of stunted height with lasso giving 3.17% and elastic net giving 2.43%. The subset for the Family Care Indicator (FCI) survey at week 78 also had zero R-squared values for the outcomes. However for RF and CRF methods, the R-squared values are as in Table 8. This shows some decent prediction power especially for the RF method. In terms of the additional exploration using the previous time's predictions as a variable, Figure 59 shows very similar R-squared values in which not much improvement is given, if any, when the predictions are included. However, for some of the subsets, there are great improvements when the predictions are included but since they cannot be included in 3.6, the specifics are not discussed. In general for both of these analyses, RF has the highest R-squared values, most likely due to this method selecting more predictors than the other methods. For the outcomes, incremental changes in height have the worst R-squared values over all the time points. This is most likely due to the fact that these values were not normalized, especially since the time points are not regularly spaced.

Table 8: R-squared percentage values for subsets of data at weeks 6, 10, 40, and 78 for the outcomes at the next time point for random forest (RF) and conditional random forest (CRF). Note: Inc. Ch. = Incremental Change in Height

Subset	**Method**	**Inc. Ch.**	**Height**	**HAZ**	**Stunted**
Week 6 Breast Milk Metabolomics	RF	79.31	76.62	77.00	84.53
	CRF	47.21	41.92	40.36	23.29
Week 6 Clinical Stool TAC	RF	18.99	28.62	54.49	74.58
	CRF	6.61	4.29	3.32	3.77
Week 6 Viral Shedding Stool TAC	RF	70.42	52.41	51.55	73.14
	CRF	8.08	10.22	18.54	6.79
Week 10 Clinical Stool TAC	RF	67.93	77.94	79.11	79.55
	CRF	7.23	10.43	20.66	3.53
Week 10 Viral Shedding Stool TAC	RF	56.78	64.13	58.47	64.65
	CRF	10.90	4.66	5.42	6.15
Week 40 Blood Metabolomics	RF	81.90	78.55	79.57	86.57
	CRF	48.80	45.01	49.36	45.89
Week 40 Stool Metabolomics	RF	78.31	77.98	78.41	86.57
	CRF	40.66	43.47	45.13	41.16
Week 78 FCI	RF	51.46	36.81	51.39	19.06
	CRF	23.95	26.50	25.79	20.45

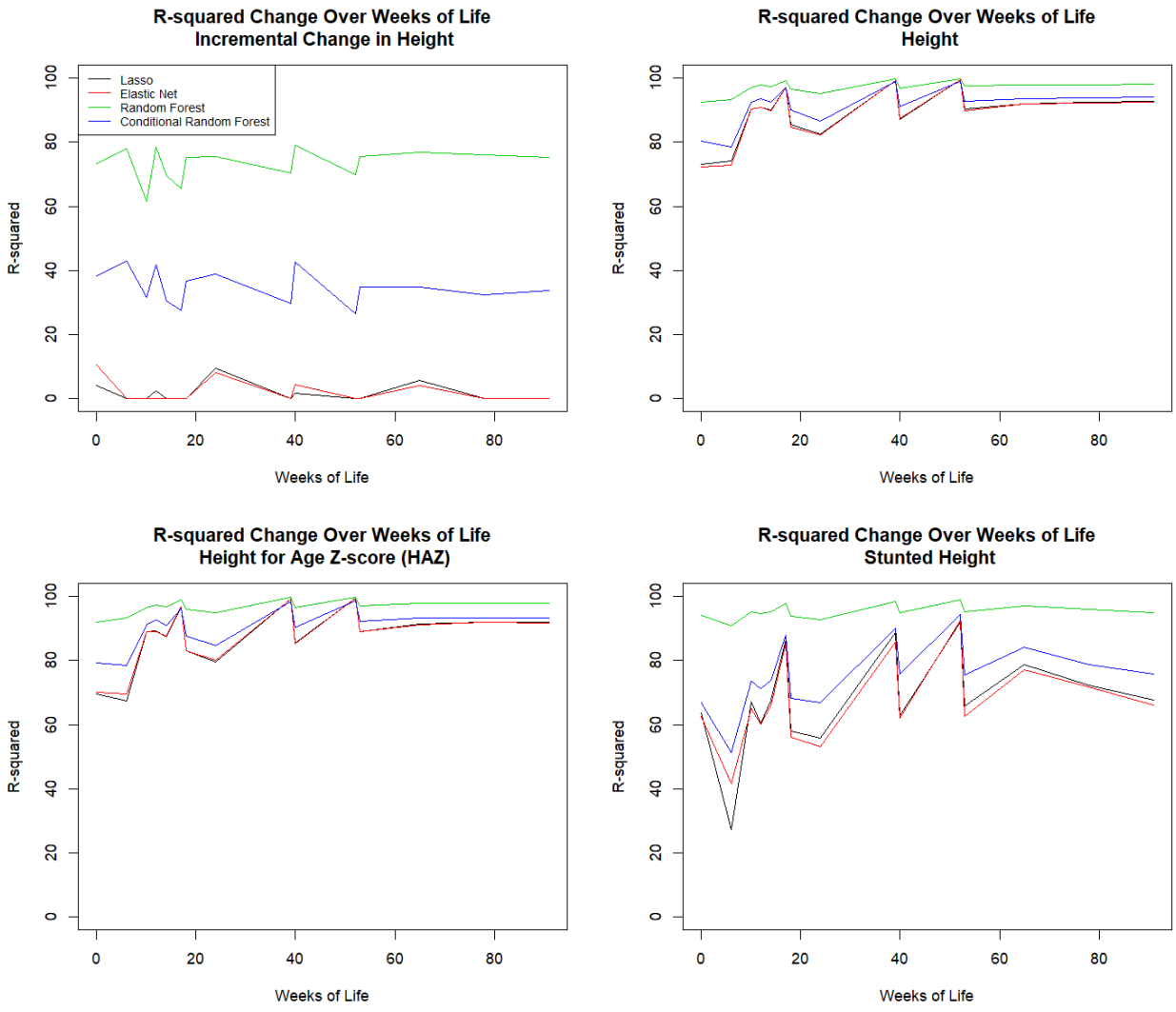

Figure 58: The R-squared values over time for predicting the next time point's outcome (incremental change in height, height, HAZ, and stunted height) for lasso, ridge, random forest, and conditional random forest.

3.6 Results for Models Predicting Outcomes at Two Plus Years

The results from selecting covariates in 3.5 were used to create models with a updating prediction for the outcomes which are the two year responses of height, HAZ, and Bayley's scores along with Mullen's at three years and Weschler's at four and five years. In these models, the predictors at enrollment are used to predict the two year and beyond outcomes using lasso, elastic net, RF variable importance (VIMP), and CRF with conditional VIMP.

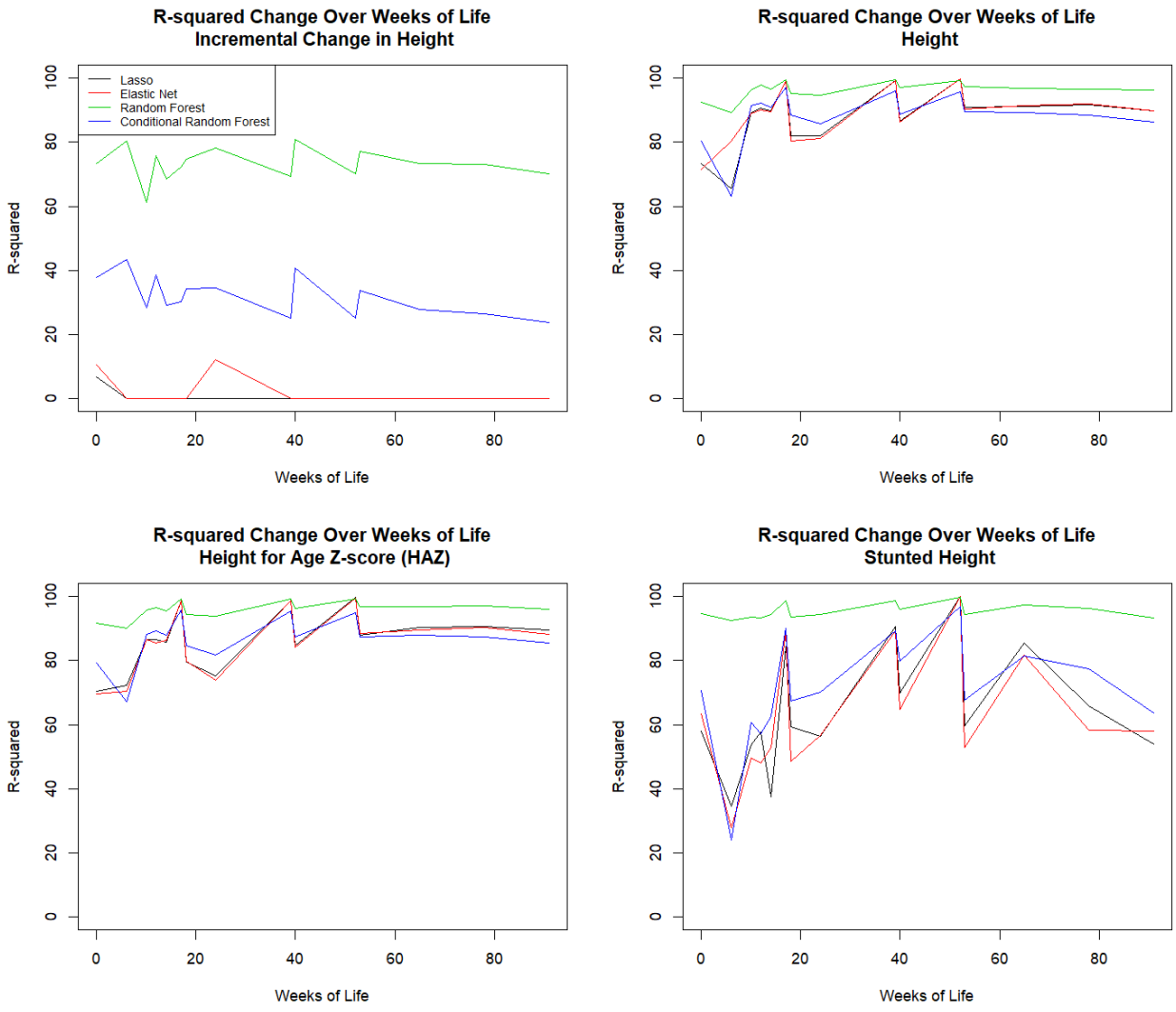

Figure 59: The R-squared values over time for predicting the next time point's outcome (incremental change in height, height, HAZ, and stunted height) with the previous time point's predictions as a new variable for lasso, ridge, random forest, and conditional random forest.

Predictors are selected from the enrollment time point and added to the predictors at week 6. This combined list is then used to again predict the outcomes at two years and beyond. Then variables are selected to be added to week 10 variables. The process keeps adding selected predictors and updating the model in this manner up to week 91. Thus, one is able to track how much the additional predictors add to the prediction of the outcomes by the percent variation explained, the R-squared. Specifically, when variable selection is used, one

can see which variables are selected when and if some tend to lose their informativeness as time goes on. When VIMP is used, one can rank the covariates in terms of importance which also tells about a variable's informativeness towards the outcomes over time.

For the change in R-squared over time, we can reference each method for each outcome of Bayley's at year two, Mullen's at year three, Weschler's at years four and five, and height and HAZ at year two in Figures 60, 61, 62, and 63 respectively. Collectively, lasso and elastic net perform poorly even as time goes on with the best R-squared values being for the height at two years outcome. The RF and CRF do much better than the penalized regressions at all time points and for all outcomes. One reason the RF method is uniformly better in terms of the R-squared values is due to the number of predictors included and that are deemed important in each model. Thus, the R-squared is over inflated due to the number of predictors included at each time and thus for each model. The CRF does select more predictors than the penalized regressions, however this method at least considers the correlation between predictors and is thus the method most trustworthy. In general, it seems as though height at two years is predicted the best followed by HAZ at two years. For the developmental outcomes, the R-squared values are similar across the time points and do not show much, if any, improvement over time unlike that of height at two years. Practically, a value for R-squared around 40% is fairly decent while ones near 80% are great as is the case for height at two years and the RF R-squareds.

As for the rank of VIMP of predictors over time for Bayley's scores, we can view Figures 64, 65, 66, and 67. All of the plots show different types of variables such as maternal, household/environmental, and biomarkers over the two year time frame with their rank by the conditional VIMP being plotted. For each of the Bayley's scores except for motor, with the maternal variables, MFGE 8 or EGF from the breast milk has a VIMP rank over time that stays fairly high or improves leading to the interpretation that this particular predictor is not only important for predicting these outcomes, but stays important as time increases even though additional variables are added each time. Some indicators for if a certain

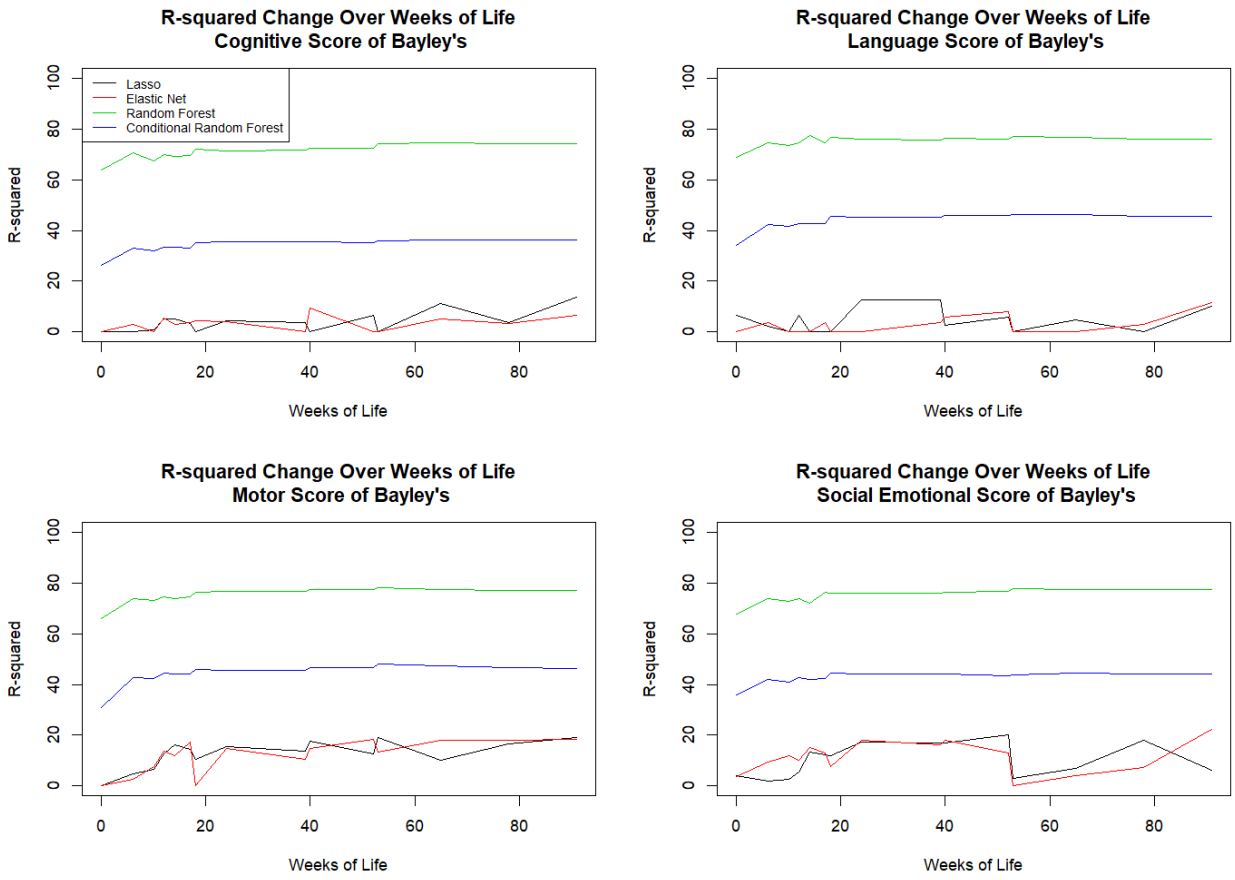

Figure 60: The change in R-squared values over time for predicting each score of Bayley's at two years of age for lasso, ridge, random forest, and conditional random forest.

vitamin supplement was taken before birth by the mother were important throughout the time frame, especially for the folic acid supplement in the top left panel of Figure 66. The indicator for household food deficit stays important under the household variables panel (top right) for the cognitive, language, and motor scores of Bayley's. It is very interesting to see some similarities between the types of variables in the Bayley's outcomes especially how important the maternal aspects can be and stay important over time. Figure 68 shows the final time point at week 91 and the conditional VIMP for the selected variables (those with a positive conditional VIMP) for each of the Bayley's scores. We can see that in addition to some of the variables plotted in Figures 64, 65, 66, and 67, predictors dealing with the number of episodes of diarrhea, the number of exclusive breastfeeding days, and height at

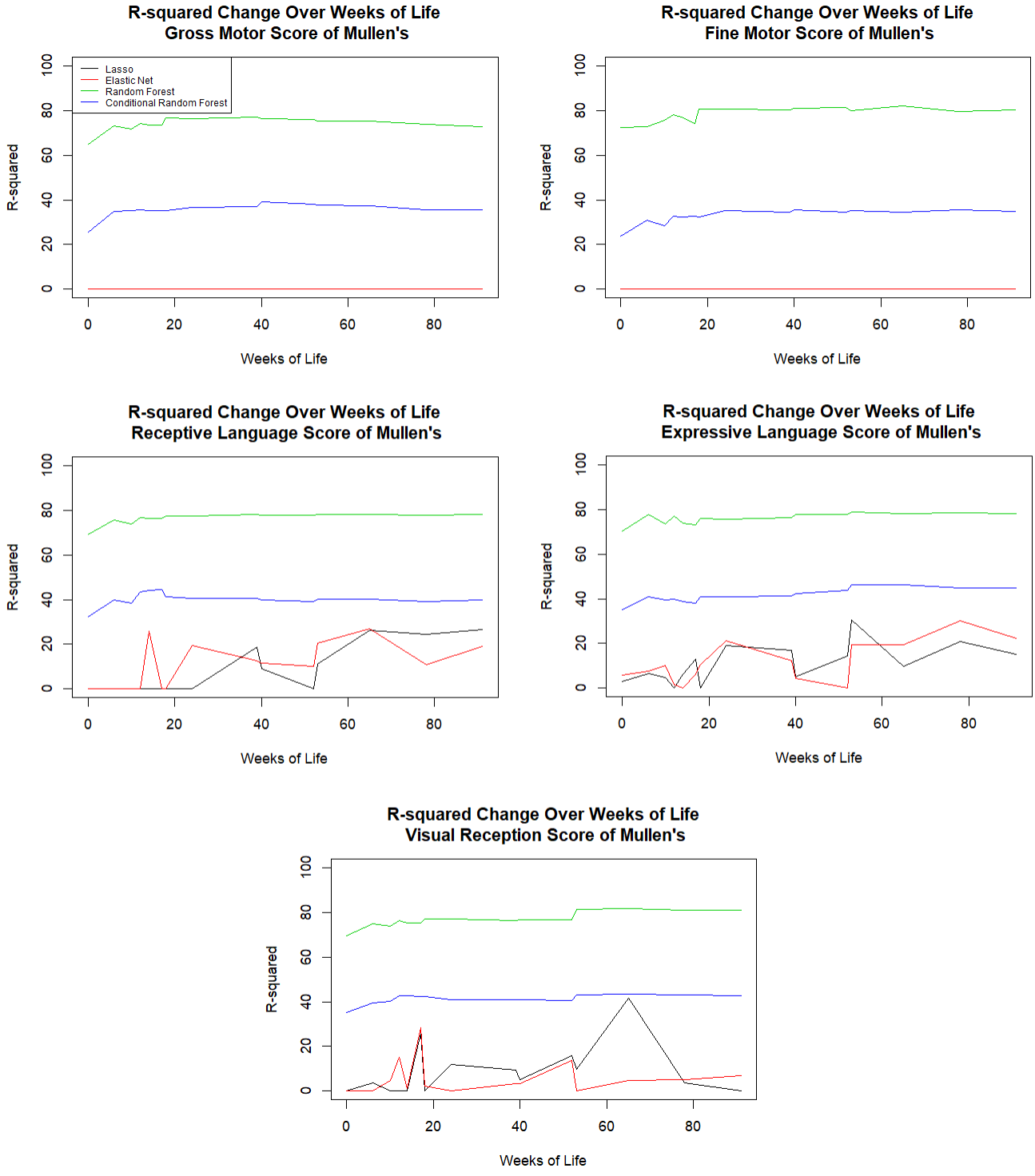

Figure 61: The change in R-squared values over time for predicting each score of Mullen's at three years of age for lasso, ridge, random forest, and conditional random forest.

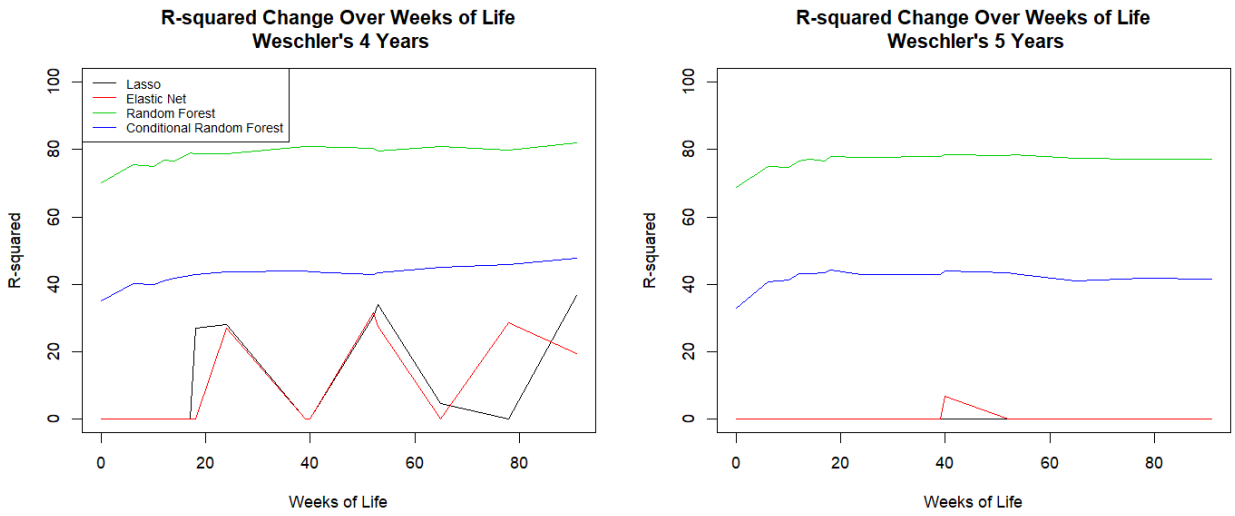

Figure 62: The change in R-squared values over time for predicting Weschler's score at 4 years (left) and 5 years (right) for lasso, ridge, random forest, and conditional random forest.

Figure 63: The change in R-squared values over time for predicting height (left) and HAZ (right) at two years of age for lasso, ridge, random forest, and conditional random forest.

given times are also selected and kept until the end. Cognitive scores seem to be affected more by the number of episodes of diarrhea than the other scores while cognitive and social emotional scores deal more with the number of exclusive breastfeeding days than the other

two. For the diarrhea episodes and exclusive breastfeeding predictors, the language score is not effected by those types of variables.

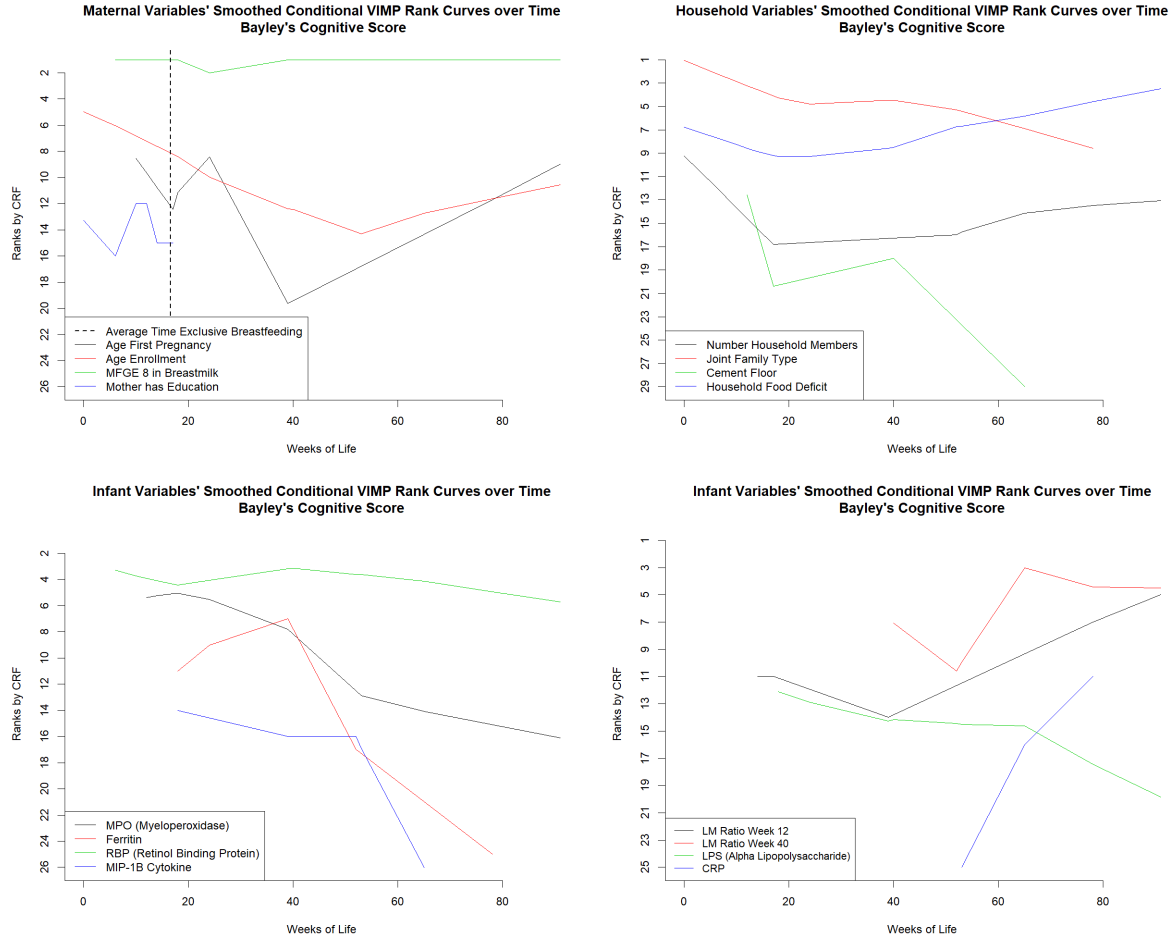

Figure 64: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Bayley's cognitive score.

Similarly, some predictors of height at two years over time is in Figure 69. For the height outcome, we can see that variables' VIMP rank such as for maternal weight decreases over time. This may be due to the added information about the child's own height as time increases thus overriding the mother's influence in combination with the new information. One can also see from this figure that mother's education is no longer selected around week 52 or 53 and thus loses it's importance. This is possibly due to the addition of not only the children's heights, but the vaccine responses as seen in the lower righthand panel of

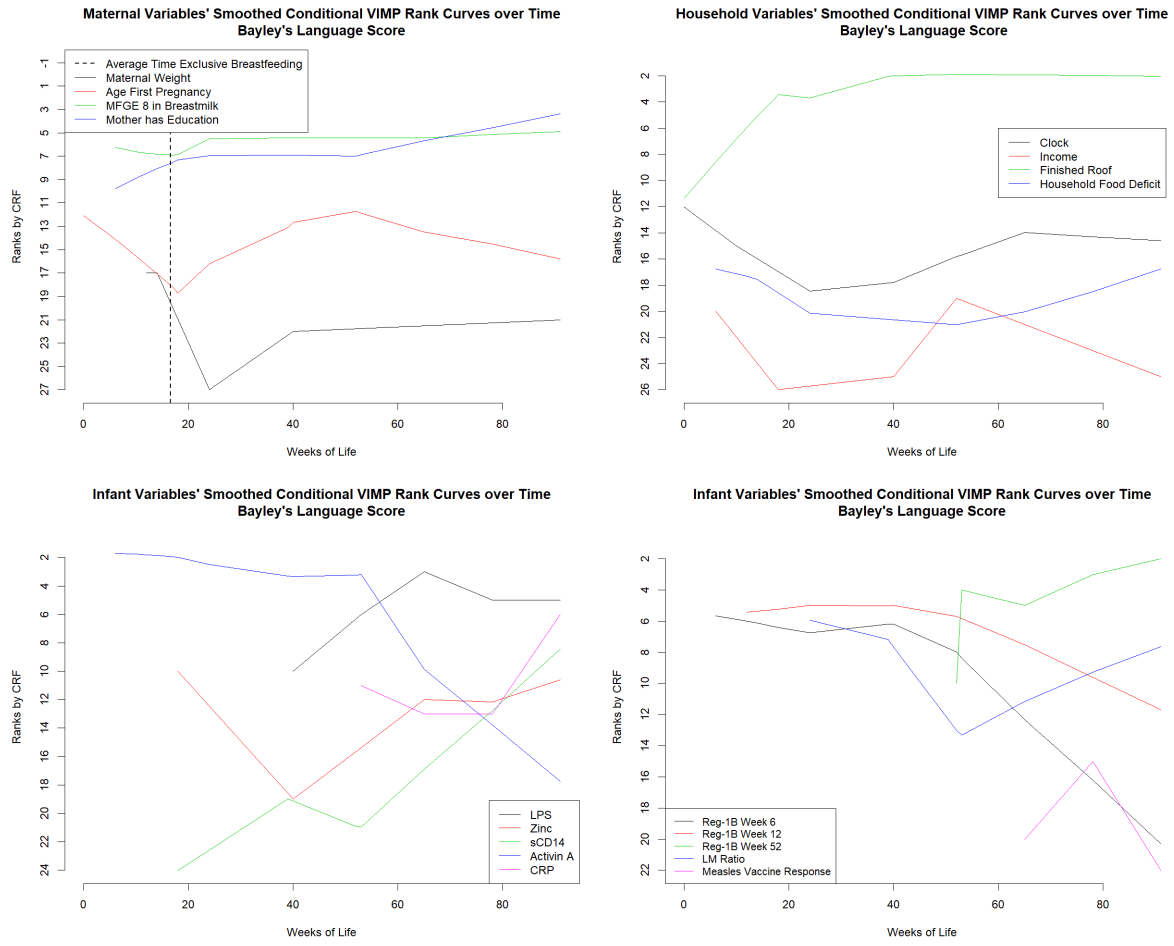

Figure 65: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Bayley's language score.

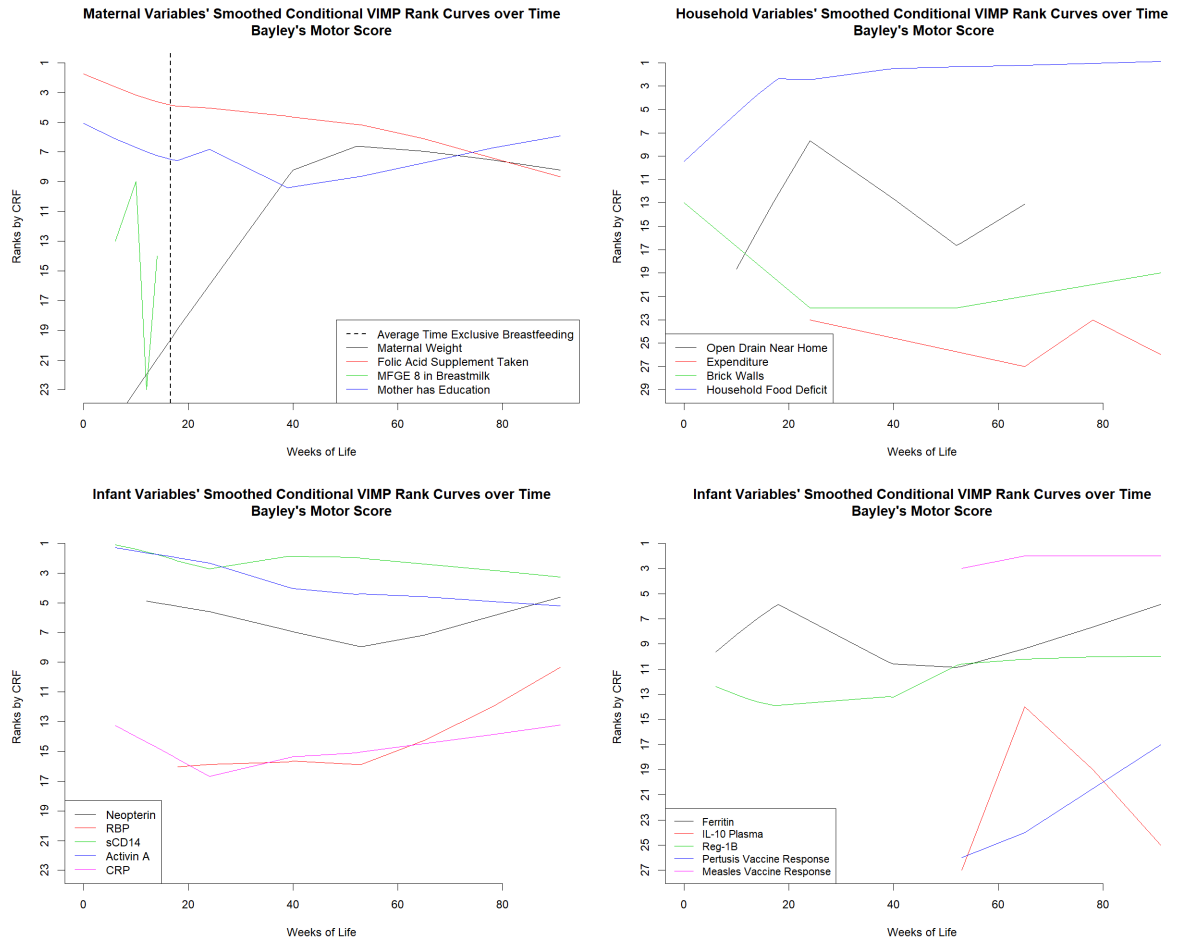

Figure 66: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Bayley's motor score.

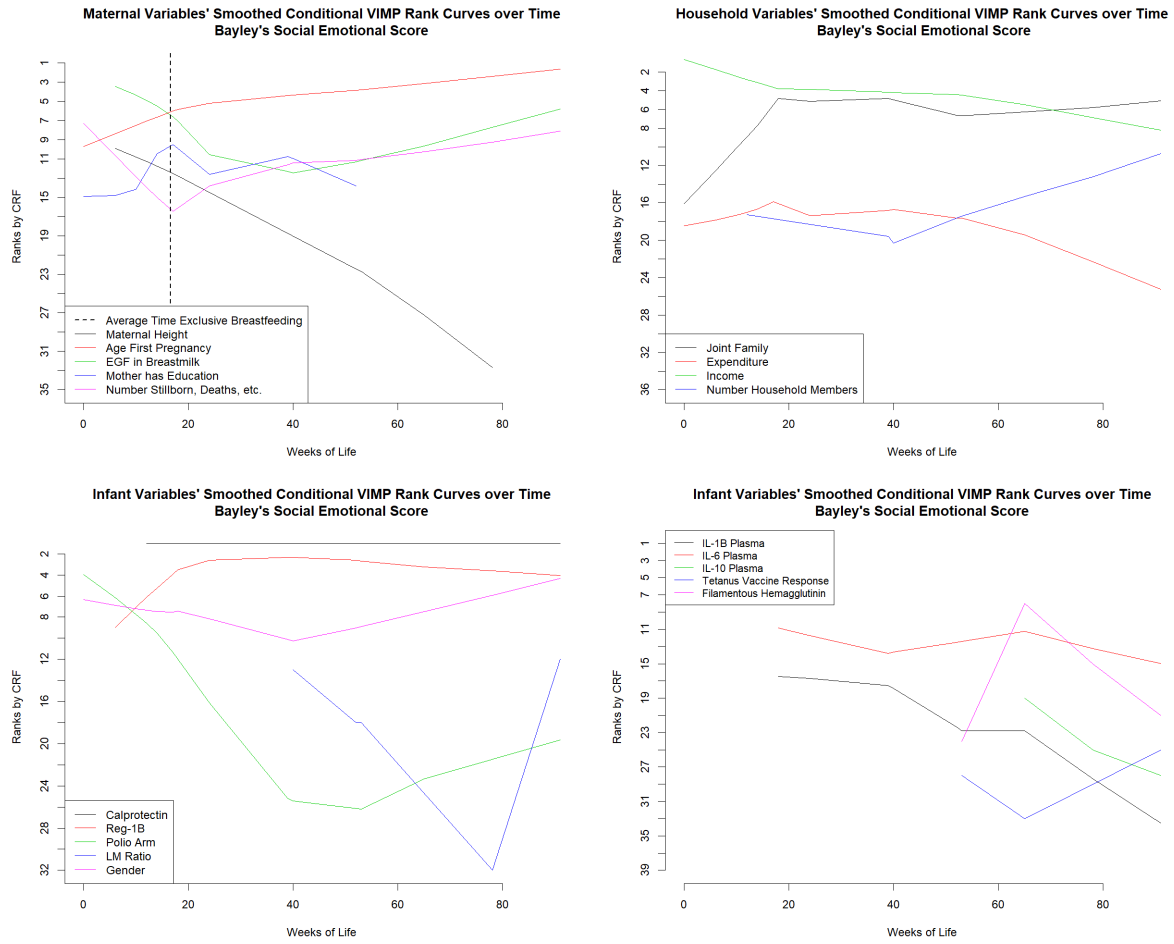

Figure 67: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Bayley's social emotional score.

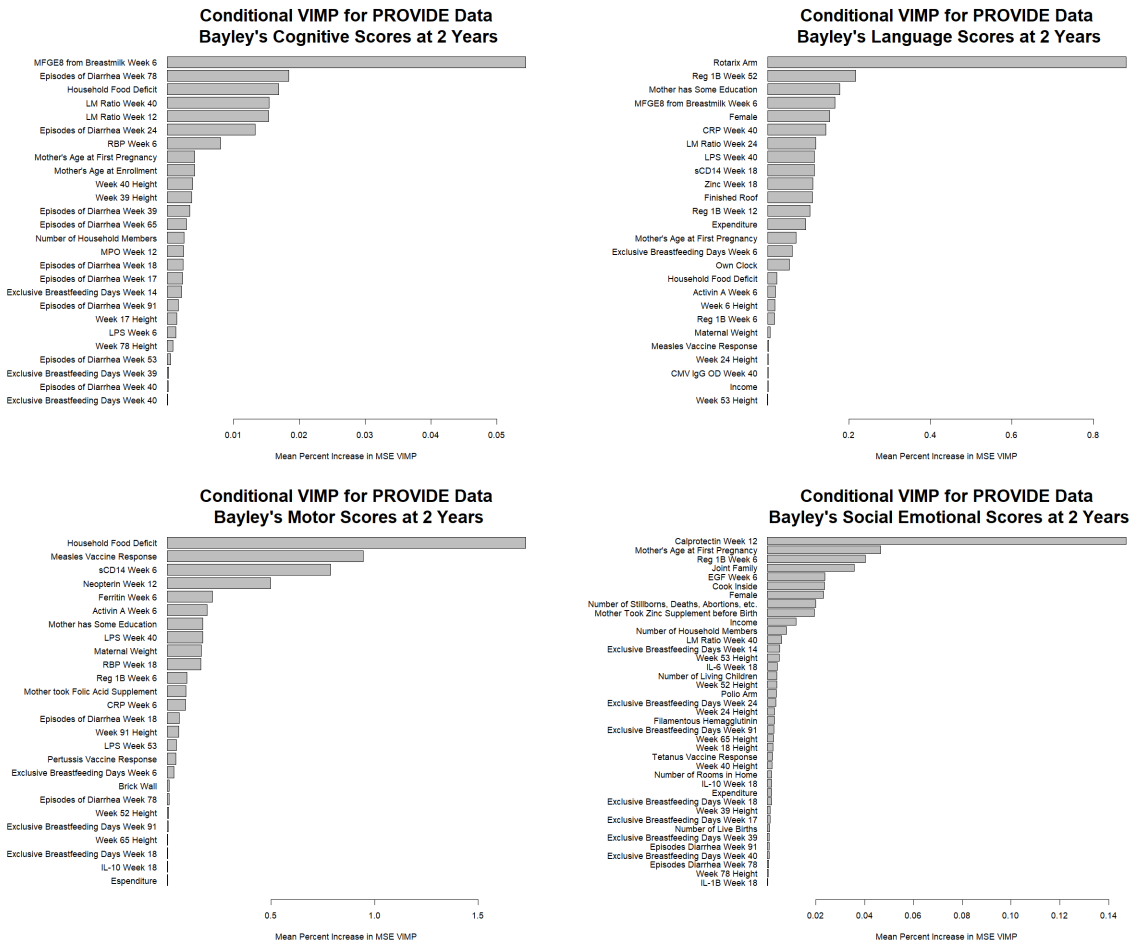

Figure 68: Bayley's positive conditional VIMP per predicting each score at two years: Cognitive (top left), Language (top right), Motor (bottom left), Social Emotional (bottom right).

Figure 69. Similar to the Bayley's scores, the indicator for household food deficit is also deemed important over time. From Figure 70 the top five predictors are of course the latest heights with the height at week 91 being clearly the most important variable. However, other variables are still selected in terms of having a positive conditional VIMP as seen more clearly in the right panel of Figure 70, but for the predictors with conditional VIMP values near zero, further analyses would need to be completed to state whether these variables are truly important.

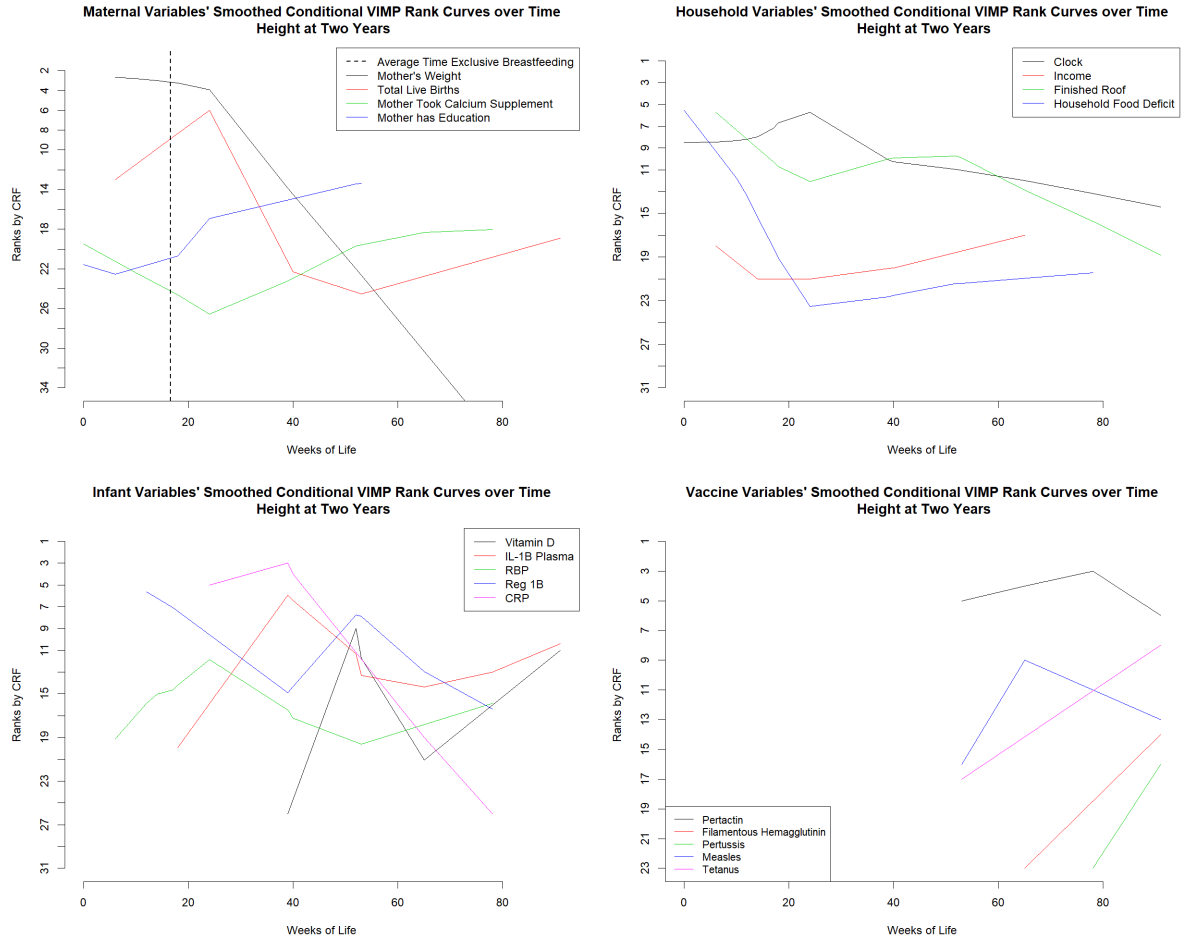

Figure 69: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting height at two years of age.

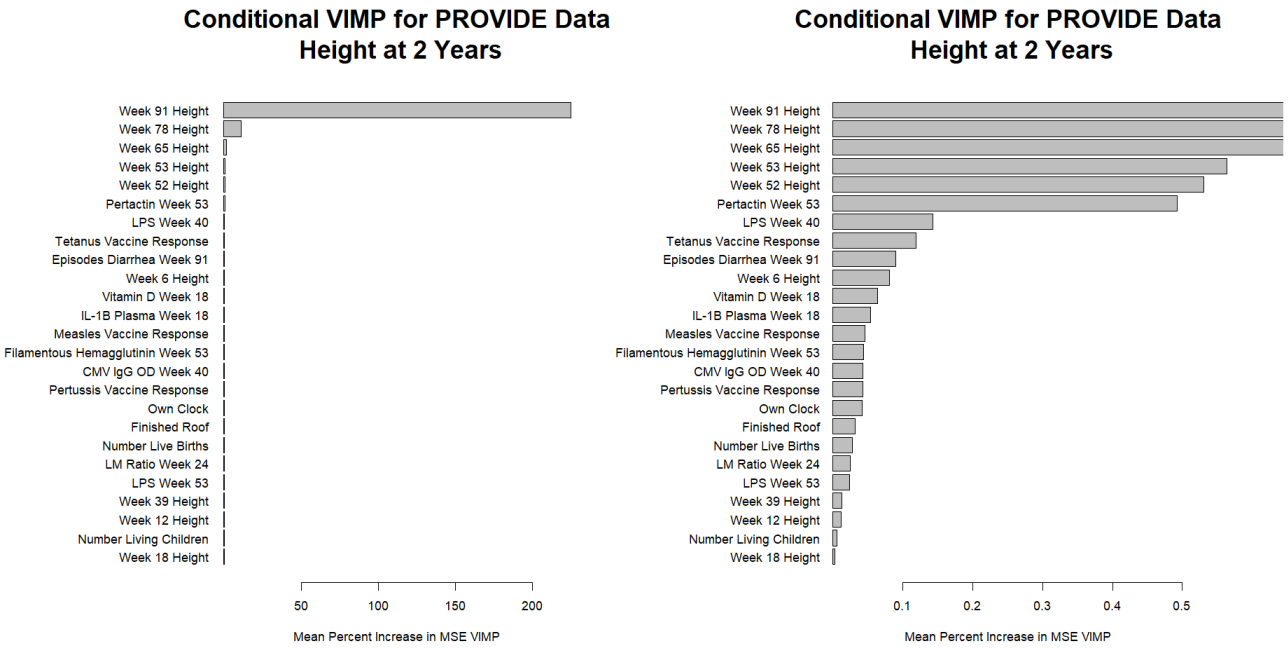

Figure 70: Predicting height at two years, the conditional VIMP at week 91 for predictors with positive conditional VIMP full scale (left) and the zoomed in version of the same graph (right).

Overall for height and these developmental outcomes, it's safe to say from the results of these analyses that the mother's effect is not only there, but adamant and thus confirms Donowitz et al. [Donowitz et al., 2018] (coauthor). As for other types of predictors, some vaccine responses seem to be predictive of development and growth including the measles vaccine response as shown for Bayley's language and motor scores in the bottom right panels of Figures 65 and 66 and for height at two years from the bottom right panel of Figure 69. Cytokines and other biomarkers tend to remain important for predicting Bayley's and height indicating that they are important in both types of child growth and development. Multiple other predictors measured from the infants are deemed important over time, however for height, more of the selected predictors (ones with positive conditional VIMP) are height variables from previous weeks as would be expected. Even though this outcome includes multiple previous height variables, we can see from the plots in Figure 69 that other variables are still considered important within the model for predicting the height at two years. Since

the conditional VIMP is used, there is no need to have hesitation about the correlation between any of these predictors.

For other outcomes such as Mullen's, Weschler's, and HAZ similar figures are included in the Appendix. The results for HAZ at two years are shown in Figures A and B. The maternal variables in this case seem to drop out before the last time point except for the number of stillborns, deaths, abortions, MC and DNRs the mother has had. Only the number of rooms in the home was selected and stayed through to week 91 from the household/environmental type predictors while several biomarkers and other such measures from the infants were included in the week 91 model. As for vaccine responses, only the pertussis vaccine response and pertactin were kept. Several of the previous HAZ values were also included while only one time point for each of the number of exclusive breastfeeding days and the number of episodes of diarrhea were retained until week 91.

Figures C, D, E, F, and G in the Appendix show some of the variables selected and their rank by the conditional VIMP over time for the Mullen's scores. The predictors with positive conditional VIMP at week 91 are shown in Figure H. From all these figures, one can see which variables were selected throughout and at the last time point of week 91. For Mullen's gross motor response, no maternal predictors were selected and kept throughout the entire time frame with few being selected for all other outcomes of Mullen's: 1 for expressive language, 6 for receptive language, 3 for fine motor, and 4 for visual reception. Similarly, very few household/environmental predictors were selected with the highest number being only two for receptive language. The main types of variables chosen are previous heights, episodes of diarrhea, or biomarkers and other measures from the infants. Thus, the types of variables selected here are quite different from those selected within the Bayley's outcomes which is not quite to be expected, although they are different tests and scales which cannot be directly compared.

Finally, Weschler's positive conditional VIMP at week 91 are shown in Figure K while some variables over time are shown in Figures I and J. With both the 4 and 5 year outcomes,

there is a good mix of variables selected and kept to the end of week 91. Unfortunately, there is not much overlap in the exact predictors that were selected between the two time points most likely due to the fact that more subjects were included with the five year outcome of Weschler's due to the switch from Mullen's to Weschler's at four years. However, those that do overlap include: the female indicator, CRP week 18, number of rooms in the home, the father's occupation, LPS week 18, income, expenditure, and a few previous heights and episodes of diarrhea. However, in terms of comparison to Bayley's, the results from Weschler's are more similar than Mullen's, especially due to the types of variables showing up including maternal and household predictors.

4 Conclusions and Future Research

4.1 Conclusions

In general, the ideas behind variable importance (VIMP) encompass many situations and numerous different calculations which can be model specific. Throughout this research, the cases of classification for NICU data and regression for the PROVIDE data were explored for creating VIMP confidence intervals and a cutoff value which can be used to select variables. Logistic regression versus random forest (RF) were applied for the NICU data in which the absolute value of the test statistic from logistic regression as the VIMP was much more variable, especially in the rankings of the predictors, than for either VIMP measure in RF. However, the Gini index in RF is worse at correctly ranking the variables than the mean decrease in accuracy, both of which may only be calculated in the classification setting for RF. For finding a cutoff for the NICU or PROVIDE data, the maximum of the means and the maximum of the medians were similar for all three VIMP measures (mean decrease in accuracy, Gini index, and absolute value of the test statistic), however there is not one specific cutoff method which led to a consensus between VIMP measures. In the regression setting with the PROVIDE data, only RF was assessed with the mean percent increase in MSE VIMP where the rankings were very difficult to get correct due to the data's complex structure. To find the cutoff values for selecting predictors using VIMP, conditional random forest (CRF) with two VIMP measures and regular RF with the mean percent increase in MSE were assessed in which the maximum of the means led to the best cutoff for the PROVIDE data. In both classification and regression, the permutation based VIMP measures for RF, their standard deviations, and ranks stabilize fairly quickly although if one wants to ensure correct rankings, the number of trees needs to be increased until the VIMP measures stabilize, which depending on the data may be fairly large. In terms of computation time, creating a RF with a large number of trees takes about as long as implementing one of the

three procedures from Ishwaran and Lu [Ishwaran and Lu, 2018] or our per tree bootstrapping method, but the timing aspect also depends on the size of the data set for the number of observations and the number of predictors. Thus, if one is mainly interested in predictions, training a RF model and using the subsampling, delete-d jackknife, double bootstrapping, or per tree VIMP bootstrapping methods may be of interest to give confidence intervals for that particularly sized forest. If stable rankings are of interest for the predictors, then creating a large forest and plotting the mean and standard deviation as in 2.4 would be preferable.

For the theoretical aspects of the linear regression's estimated coefficients as the VIMP, higher correlations between predictors adds to the difficulty of obtaining the correct important variable(s). As is the case in most statistical topics, increasing the sample size improves the probability of obtaining the correct important variable or the correct order. However, there are cases where if the correlation is so close to one that the VIMP measures cannot distinguish between the predictors. Also, there is a balance between the sample size and the number of predictors in that if the number of predictors is growing faster than the sample size, then the probability of obtaining the correct order or important variable does not go to one. All of this indicates that a sample size larger than the number of predictors, and even in consideration of large correlations, should be used to obtain correct rankings and good estimates of VIMP measures, especially within the linear regression setting.

With the PROVIDE study, many covariates were selected in various stages using four different methods including penalized regressions (lasso and elastic net) and RF and CRF with respective VIMP calculations. These models were computed per time point predicting the next time point's incremental change in height, height, height-for-age z-score (HAZ), and stunted growth. This analysis gave an ideal about which predictors could be useful in the growth process over time. After these methods were compared, the list of covariates selected was then used in an updating model where predictions were made for height, HAZ, and Bayley's at two years, Mullen's at three years, and Weschler's at four and five years while covariates were selected at each time and added to the next time's predictor list. The

R-squared values from these updating models showed that height followed by HAZ were easiest to explain. The developmental tests all had similar values and trends across the time points and models. Lasso and elastic net performed uniformly poorly for all outcomes except height at two years.

In terms of predictors chosen, maternal factors are still considered important over the two years for predicting various developmental and growth outcomes. Additionally, household and environmental aspects such as having a household food deficit or having an open drain near the home are kept from enrollment over the entire time frame and deemed important for predicting a particular outcome. In some cases, the variables are not only considered important, but their importance may grow over the time frame. For the height at two years, the main types of predictors chosen naturally have to do with the previous heights, which are expected to be of high importance, however there are multiple other variable types chosen such as cytokines from the infants' plasma samples. In general, there is a consensus between all outcomes on the fact that all types of variables measured are important, however the specific predictors chosen may vary especially when comparing Mullen's to Bayley's or Weschler's outcomes. Thus, there is a discrepancy between Mullen's and the other developmental tests possibly showing how these three different tests measure separate aspects of the children's development.

4.2 Future Research

There is a great deal of information that has risen from this research, yet there are still numerous directions and questions left unanswered. Specifically, finding new ways to calculate confidence intervals for measures of VIMP or VIMP measures which are not model specific should be assessed including the change in metrics after a random shuffle of a predictor's values.

Within the PROVIDE growth modeling setting, a framework for a proposed model has been outlined but not explored or applied further. Thus, much work in this direction can be completed. Instead of using penalized regressions and random forests to select variables, screening methods could be applied to the large set of variables for PROVIDE along with implementing various imputation techniques in order to increase the sample size for certain predictors which have numerous missing values. There are various additional methods for selecting and modeling these types of outcomes in which multiple different paths may be taken to go forward in this research. In addition to these future paths, the EEG data can be analyzed to pick up on patterns and signals for the individuals these were measured on and in relation to the neurocognitive tests such as Bayley's, Mullen's, and Weschler's. The main goal for relating the EEGs to developmental tests would be that these EEGs are culturally independent while the tests depend heavily on the culture of the individuals taking it and thus cannot be compared across countries. Metagenomic and metabolomic data from other infants' samples are also available which may allow for information about how the gut bacteria affects the outcomes of interest. All in all, there is still a multitude of information that can be gathered from this PROVIDE study cohort, especially since it is still ongoing.

References

[Archer and Kimes, 2008] Archer, Kellie J.; Kimes, Ryan V. "Empirical characterization of random forest variable importance measures." Computational Statistics & Data Analysis, vol. 52, 2008, pp. 2249-2260. doi:10.1016/j.csda.2007.08.015.

[Agresti, 2014] Agresti, Alan. Categorical Data Analysis. 3rd ed. Hoboken, NJ: John Wiley & Sons, 2014.

[Auret and Aldrich, 2011] Auret, Lidia; Aldrich, Chris. "Empirical comparison of tree ensemble variable importance measures." Chemometrics and Intelligent Laboratory Systems, vol. 105, 2011, pp. 157-170. doi:10.1016/j.chemolab.2010.12.004.

[Boulestix et al., 2012] Boulestix, Anne-Laure; Janitza, Silke; Kruppa, Jochen; Konig, Inke R. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics." WIREs Data Mining Knowledge Discovery, vol. 2, no. 6, 18 Oct. 2012, pp. 493-507. doi:10.1002/widm.1072.

[Breiman, 2001] Breiman, L. "Random Forests." Machine Learning, vol. 45, no. 1, Oct. 2001, pp. 5-32. doi:10.1023/A:1010933404324.

[Chirwa et al., 2014] Chirwa, Esnat D.; Griffiths, Paula L.; Maleta, Ken; Norris, Shane A.; Cameron, Noel. "Multi-level modelling of longitudinal child growth data from the Birth-to-Twenty Cohort: a comparison of growth models." Annals of Human Biology, vol. 41, no. 2, 2014, pp. 168-179. doi:10.3109/03014460.2013.839742.

[Chollet and Allaire, 2018] Chollet, François; Allaire, J.J. Deep Learning with R. Manning, 2018.

[Dewey and Begum, 2011] Dewey, K. G.; Begum, K. "Long-Term Consequences of Stunting in Early Life." Maternal & child nutrition, vol. 7 Suppl 3, Oct. 2011, pp. 5-18. doi:10.1111/j.1740-8709.2011.00349.x.

[Donowitz et al., 2018] Donowitz, Jeffrey R.; Cook, Heather; Alam, Masud; Tofail, Fahmida; Kabir, Mamun; Colgate, E. Ross; Carmolli, Marya P.; Kirkpatrick, Beth D.; Nelson, Charles A.; Ma, Jennie Z.; Haque, Rashidul; Petri, William A., Jr. "Role of maternal health and infant inflammation in nutritional and neurodevelopmental outcomes of two-year-old Bangladeshi children." PLOS Neglected Tropical Diseases, vol. 12, no. 5, 29 May 2018, doi:10.1371/journal.pntd.0006363.

[Gasull et al., 2015] Gasull, Armengol; Jolis, Maria; Utzet, Frederic. "On the norming constants for normal maxima." Journal of Mathematical Analysis and Applications, vol. 422, no. 5, Feb. 2015, pp. 376-396. doi:10.1016/j.jmaa.2014.08.025

[Gregorutti et al., 2017] Gregorutti, Baptiste; Michel, Bertrand; Saint-Pierre, Philippe. "Correlation and variable importance in random forests." Statisitcal Computing, vol. 27, 2017, pp. 659-678. doi:10.1007/s11222-016-9646-1.

[Grömping, 2009] Grömping, Ulrike. "Variable Importance Assessment in Regression: Linear Regression versus Random Forest." The American Statistician, vol. 63, no. 4, 2009, pp. 308-319. doi:10.1198/tast.2009.08199.

[Grömping, 2015] Grömping, Ulrike. "Variable importance in regression models." WIREs Computational Statistics, vol. 7, Mar./Apr. 2015, pp. 137-152. doi:10.1002/wics.1346.

[Henderson and Seaby, 2006] Henderson, P. A.; Seaby, R. M. Growth II. Pisces Conservation Ltd., Lymington, England.

[Hoddinott et al., 2008] Hoddinott, J.; Maluccio, J. A.; Behrman, J. R.; Flores, R.; Martorell, R. "Effect of a nutrition intervention during early childhood on economic productivity in Guatemalan adults." Lancet, vol. 371, no. 9610, 2 Feb. 2008, pp. 411-416. doi:10.1016/S0140-6736(08)60205-6.

[Hua et al., 2018] Hua, Yuxiu; Zhao, Zhifeng; Li, Rongpeng; Chen, Xianfu; Liu, Zhiming; Zhang, Honggang. "Deep Learning with Long Short-Term Memory for Time Series Prediction." ArXiv.org, Cornell University, 24 Oct. 2018, arxiv.org/abs/1810.10161.

[Ishwaran, 2007] Ishwaran, Hemant. "Variable importance in binary regression trees and forests." Electronic Journal of Statistics, vol. 1, 2007, pp. 519-537. doi:10.1214/07-EJS039.

[Ishwaran and Lu, 2018] Ishwaran, Hemant; Lu, Min. "Standard Errors and Confidence Intervals for Variable Importance in Random Forest Regression, Classification, and Survival." Statistics in Medicine, vol. 38, June 2018, pp. 558–582. doi:10.1002/sim.7803.

[Janitza et al., 2013] Janitza, Silke; Strobl, Carolin; Boulestix, Anne-Laure. "An AUC-based permutation variable importance measure for random forests." BMC Bioinformatics, vol. 14, no. 119, 5 Apr. 2013, doi:10.1186/1471-2105-14-119.

[Jensen et al., 2019] Jensen, S.K.G; Kumar, S.; Xie, W.; Tofail, F; Haque, R. Petri, W.A.; Nelson, C.A. "Neural correlates of early adversity among Bangladeshi infants." Scientific Reports, vol. 9, no. 3507, 5 Mar. 2019, doi:10.1038/s41598-019-39242-x.

[Kaneshiro, 2014] Kaneshiro, Neil K. "Apgar Score: MedlinePlus Medical Encyclopedia." Edited by David Zieve, Isla Ogilvie, and A.D.A.M. Editorial Team, MedlinePlus NIH, U.S. National Library of Medicine, 20 Nov. 2014. medlineplus.gov/ency/article/003402.htm.

[Kikpatrick et al., 2015] Kirkpatrick, B. D.; Colgate, E. R.; Mychaleckyi, J. C.; Haque, R.; Dickson, D. M.; Carmolli, M. P.; Nayak, U.; Taniuchi, M.; Naylor, C.; Qadri, F.; Ma, J. Z.; Alam, M.; Walsh, M. C.; Diehl, S. A.; PROVIDE Study Teams; Petri, W. A., Jr. "The 'Performance of Rotavirus and Oral Polio Vaccines in Developing Countries' (PROVIDE) study: description of methods of an interventional study design to explore

complex biologic problems." The American Journal of Tropical Medicine and Hygiene, vol. 92, no. 4, April 2015, pp. 744-751. doi:10.4269/ajtmh.14-0518.

[Koehrsen, 2018] Koehrsen, Will. "Beyond Accuracy: Precision and Recall." Towards Data Science, Medium, 3 Mar. 2018, towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c.

[Kuhn, 2018] Kuhn, Max. "The Caret Package." Github Sites, 26 May 2018, topepo.github.io/caret/variable-importance.html.

[Kuhn and Johnson, 2016] Kuhn, Max; Johnson, Kjell. Applied Predictive Modeling. Springer Science+Business Media LLC New York, 2016. doi:10.1007/978-1-4614-6849-3.

[Kulaylat et al., 2018] Kulaylat, Audrey S.; Buonomo, Erica L.; Scully, Kenneth W.; Hollenbeak, Christopher S.; Cook, Heather; Petri, William A., Jr.; Stewart, David B., Sr. "Development and Validation of a Prediction Model for Mortality and Adverse Outcomes Among Patients with Peripheral Eosinopenia on Admission for Clostridium difficile Infection." JAMA Surgery. vol. 153, no. 12, Dec. 2018, pp. 1127-1133. doi:10.1001/jamasurg.2018.3174.

[Lindstrom and Bates, 1990] Lindstrom, Mary J.; Bates, Douglas M. "Nonlinear Mixed Effects Models for Repeated Measures Data." Biometrics, vol. 46, no. 3, Sept. 1990, pp. 673-687. https://www.jstor.org/stable/2532087.

[Lu et al., 2017] Lu, Miao; Zhou, Jianhui; Naylor, Caitlin; Kirkpatrick, Beth D.; Haque, Rashidul; Petri, William A., Jr.; Ma, Jennie Z. "Application of Penalized Linear Regression Methods to the Selection of Environmental Enteropathy Biomarkers." Biomarker Research, vol. 5, no. 9, 9 Mar. 2017, doi:10.1186/s40364-017-0089-4.

[Moreau et al., 2019] Moreau, G.B; Ramakrishnan, G.; Cook, H.L.; Fox, T.E.; Nayak, U.; Ma, J.Z.; Colgate, E.R.; Kirkpatrick, B.D.; Haque, R.; Petri, W.A. Jr. "Child-

hood growth and neurocognition are associated with distinct sets of metabolites." EBioMedicine, 24 May 2019, doi:10.1016/j.ebiom.2019.05.043.

[Murray and Conner, 2009] Murray, Kim; Conner, Mary M. "Methods to quantify variable importance: implications for the analysis of noisy ecological data." Ecology, vol. 90, no. 2, 1 Feb. 2009, pp. 348-355. doi:10.1890/07-1929.1.

[Olden et al., 2004] Olden, Julian D.; Joy, Michael K.; Death, Russell G. "An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data." Ecological Modeling, vol. 178, 2004, pp. 389-397. doi:10.1016/j.ecolmodel.2004.03.013.

[Oña and Garrido, 2014] Oña, Juan de; Garrido, Concepción. "Extracting the contribution of independent variables in neural network models: a new approach to handle instability." Neural Computing and Applications, vol. 25, no. 3-4, Sept. 2014, pp. 859-869. doi:10.1007/s00521-014-1573-5.

[Parr et al., 2018] Parr, Terence; Turgutlu, Kerem; Csiszar, Christopher; Howard, Jeremy. "Beware Default Random Forest Importances." Explained.ai, 26 Mar. 2018, explained.ai/rf-importance/index.html.

[Saldana and Feng, 2018] Saldana, Diego; Feng, Yang. "SIS: An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models." Journal of Statistical Software, vol. 83, no. 2, Feb. 2018. doi:10.18637/jss.v083.i02.

[Sandri and Zuccolotto, 2008] Sandri, Marco; Zuccolotto, Paola; "A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees." Journal of Computational and Graphical Statistics, vol. 17, no. 3, 2008, pp. 611-628. doi:10.1198/106186008X344522.

[Silva et al., 2012] Silva, Ikaro; Moody, George; Scott, Daniel J.; Celi, Leo A.; and Mark, Roger G. "Predicting In-Hospital Mortality of ICU Patients: The

PhysioNet/Computing in Cardiology Challenge 2012." Computing in Cardiology, vol. 39, 2012, pp. 245-48. Laboratory for Computational Physiology. MIT. https://lcp.mit.edu/pdf/SilvaCinC2012.pdf.

[Steiner et al., 2018] Steiner, Kevin L.; Ahmed, Shahnawaz; Gilchrist, Carol A.; Burkey, Cecelia; Cook, Heather; Ma, Jennie Z.; Korpe, Poonum S.; Ahmed, Emtiaz; Alam, Masud; Kabir, Mamum; Tofail, Fahmida; Ahmed, Tahmeed; Haque, Rashidul; Petri, William A., Jr.; Faruque, Abu S. G. "Species of Cryptosporidia Causing Subclinical Infection Associated with Growth Faltering in Rural and Urban Bangladesh: A Birth Cohort Study." Clinical Infectious Diseases, vol. 67, no. 9, 2018, pp. 1347-1355. doi:10.1093/cid/ciy310.

[Steiner et al., 2019] Steiner, Kevin L.; Kabir, Mamun; Priest, Jeffrey W.; Hossain, Biplob; Gilchrist, Carol A.; Cook, Heather; Ma, Jennie Z.; Korpe, Poonum S.; Ahmed, Tahmeed; Faruque, A.S.G.; Haque, Rashidul; Petri, William A. Jr. "Fecal IgA against a sporozoite antigen at 12 months is associated with delayed time to subsequent cryptosporidiosis in urban Bangladesh: a prospective cohort study." Clinical Infectious Diseases, 25 May 2019, doi:10.1093/cid/ciz430.

[Strobl et al., 2008] Strobl, Carolin; Boulesteix, Anne-Laure; Kneib, Thomas; Augustin, Thomas; Zeileis, Achim. "Conditional variable importance for random forests." BMC Bioinformatics, vol. 9, no. 307, 11 July 2008, doi:10.1186/1471-2105-9-307.

[Strobl et al., 2007] Strobl, Carolin; Boulesteix, Anne-Laure; Zeileis, Achim; Hothorn, Torsten. "Bias in random forest variable importance measures: Illustrations, sources and a solution." vol. 8, no. 25, 25 Jan. 2007, doi:10.1186/1471-2105-8-25.

[Strobl et al., 2009] Strobl, Carolin; Hothorn, Torsten; Zeileis, Achim. "Party on! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package". No. 050. University of Munich: Department of Statistics, 2009.

[Suki and Frey, 2017] Suki, Béla; Frey, Urs. "A time-varying biased random walk approach to human growth." Scientific Reports, vol. 7, no. 7805, 10 Aug. 2017, doi:10.1038/s41598-017-07725-4.

[Sullivan et al., 2018] Sullivan, B. A.; Wallman-Stokes, A.; Isler, J.; Sahni, R.; Moorman, J. R.; Fairchild, K. D.; Lake, D. E. "Early Pulse Oximetry Data Improves Prediction of Death and Adverse Outcomes in a Two-Center Cohort of Very Low Birth Weight Infants." American Journal of Perinatology, vol. 35, 28 May 2018, pp. 1331-1338. doi:10.1055/s-0038-1654712.

[Sullivan et al., 2016] Sullivan, Brynne A.; McClure, Christina; Hicks, Jamie; Lake, Douglas E.; Moorman, J. Randall; Fairchild, Karen D. "Early Heart Rate Characteristics Predict Death and Morbidities in Preterm Infants." The Journal of Pediactrics, vol. 174, 22 April 2016, pp. 57-62. doi:10.1016/j.jpeds.2016.03.042.

[K. Tjørve and E. Tjørve, 2017] Tjørve, Kathleen M. C.; Tjørve, Even. "The use of Gompertz models in growth analyses, and new Gompertz-model approach: An addition to the Unified-Richards family." PLOS ONE, vol. 12, no. 6, 5 June 2017, doi:10.1371/journal.pone.0178691.

[Troutman et al., 2018] Troutman, John A.; Sullivan, Mary C.; Carr, Gregory J.; Fisher, Jeffrey. "Development of growth equations from longitudinal studies of body weight and height in the full term and preterm neonate: From birth to four years postnatal age." Birth Defects Research, vol. 110, no. 11, 14 Mar. 2018, doi:10.1002/bdr2.1214.

[WHO MGRS Group, 2006] WHO Multicentre Growth Reference Study Group. "WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development." Geneva: World Health Organization, 2006. https://www.who.int/childgrowth/standards/Technical_report.pdf?ua=1

[Zhang et al., 2017] Zhang, Yin; Zhou, Jianhui; Niu, Feiyang; Donowitz, Jeffrey R.; Haque, Rashidul; Petri, William A., Jr.; Ma, Jennie Z. "Characterizing Early Child Growth Patterns of Height-for-Age in an Urban Slum Cohort of Bangladesh with Functional Principal Component Analysis." BMC Pediatrics, vol. 17, no. 84, 21 Mar. 2017, doi:10.1186/s12887-017-0831-y.

Appendix

Supplemental Figures: HAZ at Two Year Response

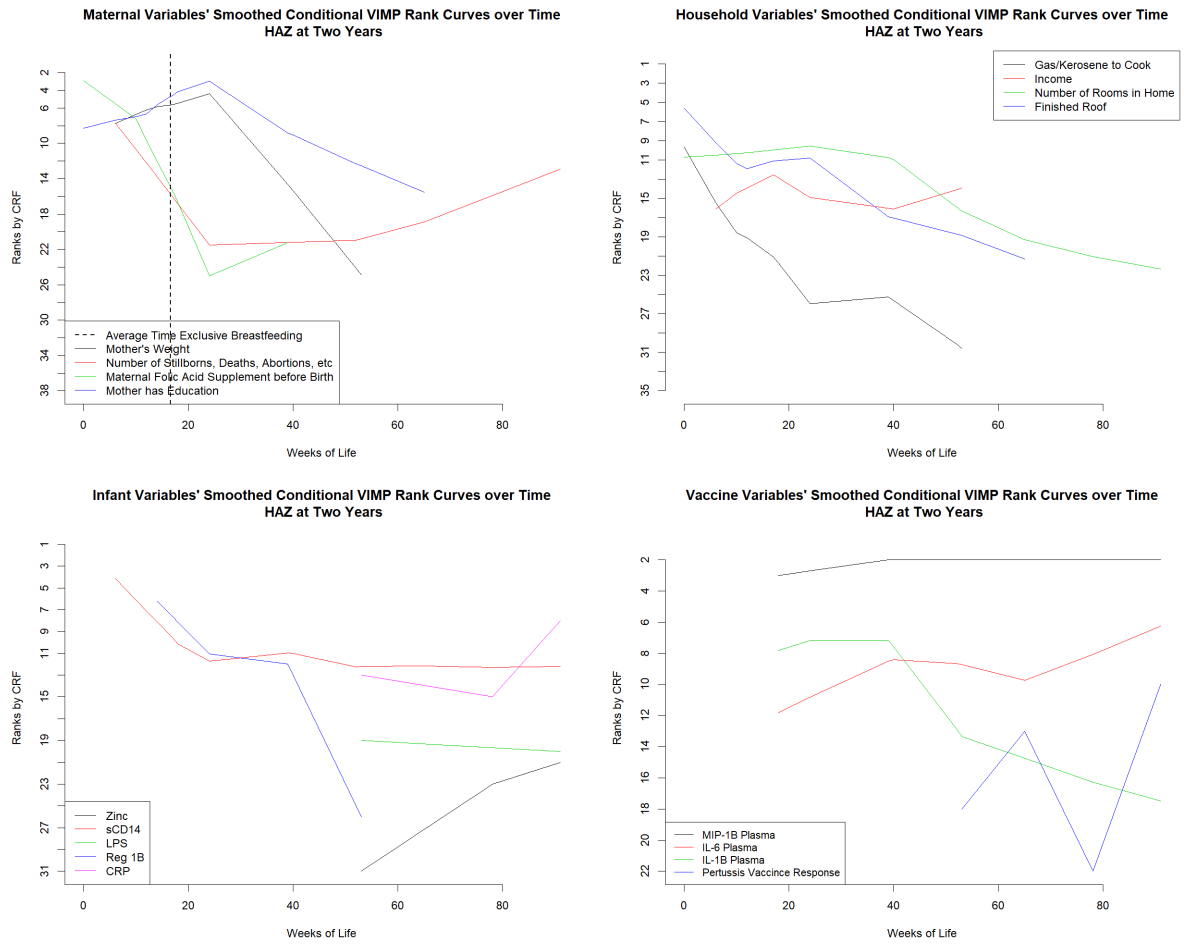

Figure A: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting HAZ at two years of age.

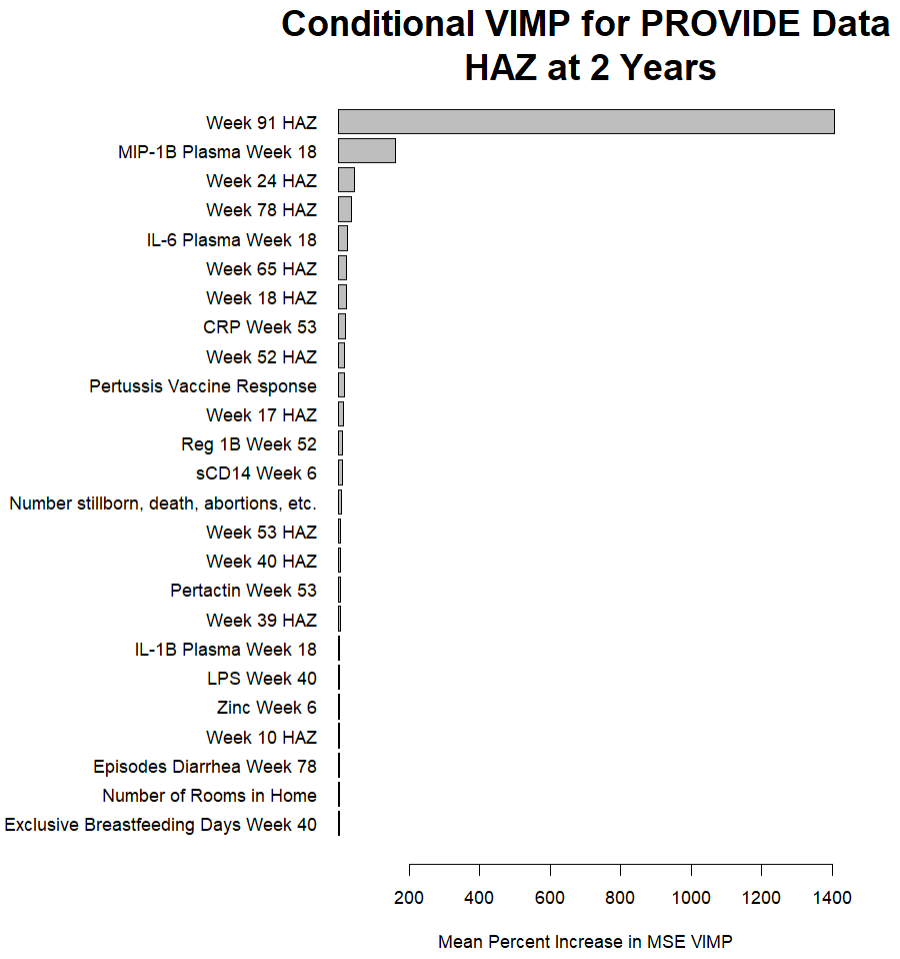

Figure B: Predicting HAZ at two years, the conditional VIMP at week 91 for predictors with positive conditional VIMP.

Supplemental Figures: Mullen's Responses

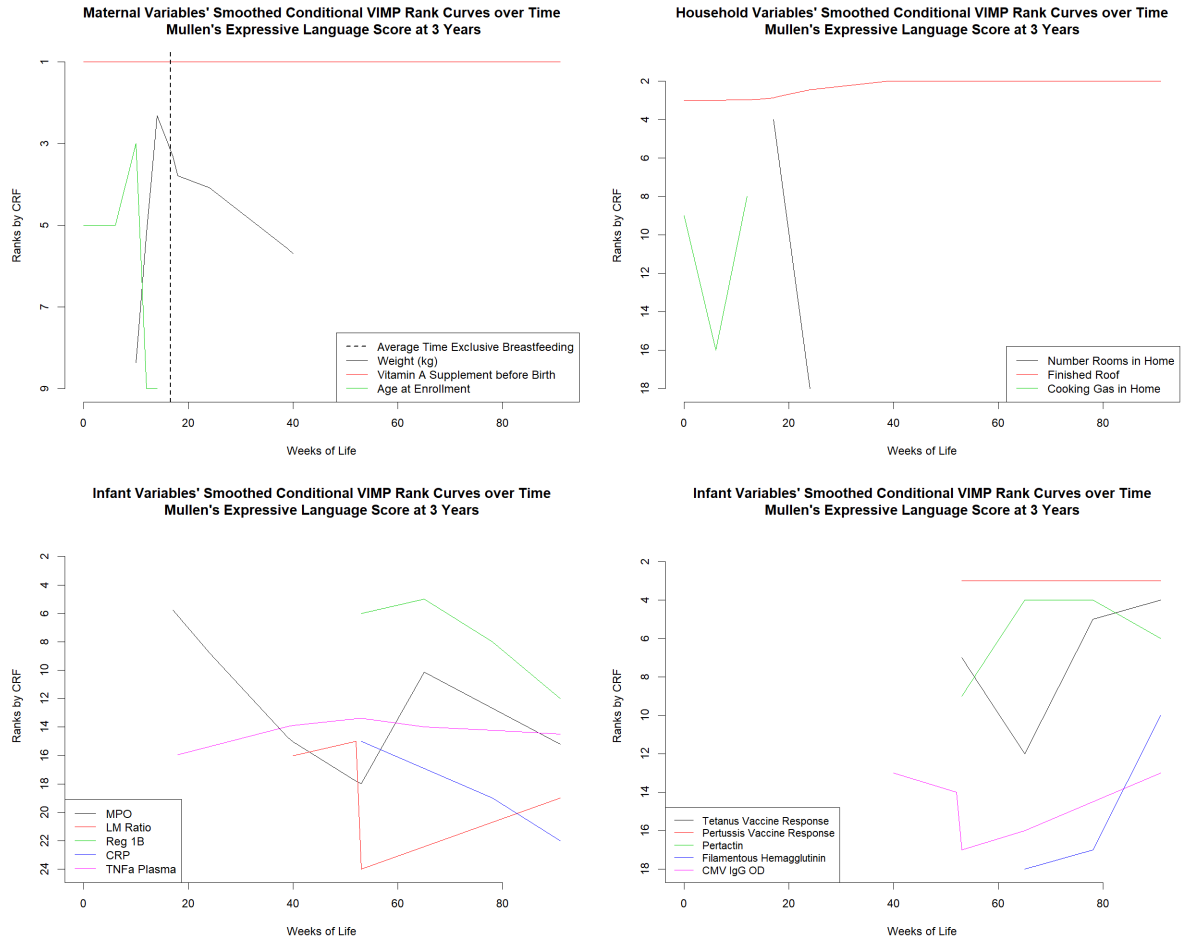

Figure C: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Mullen's Expressive Language at three years of age.

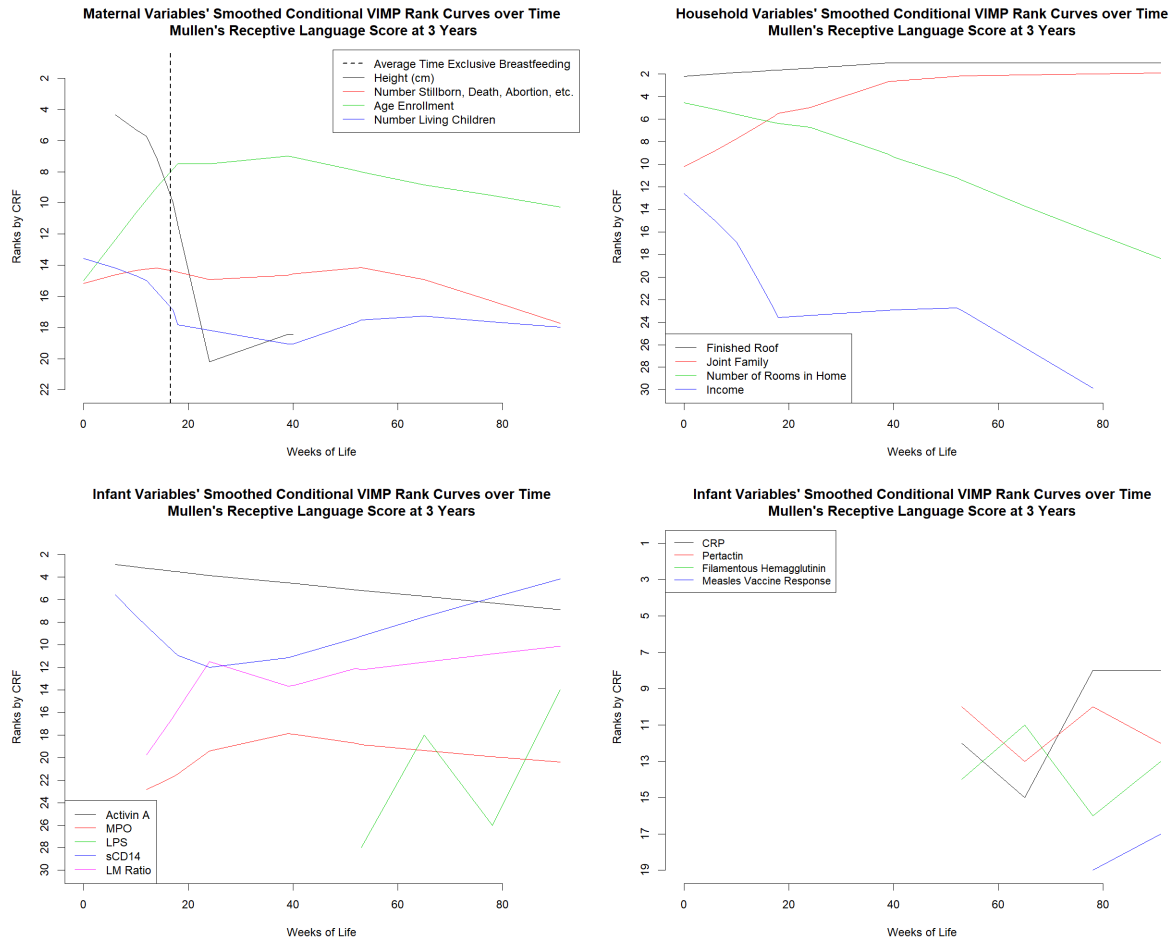

Figure D: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Mullen's Receptive Language at three years of age.

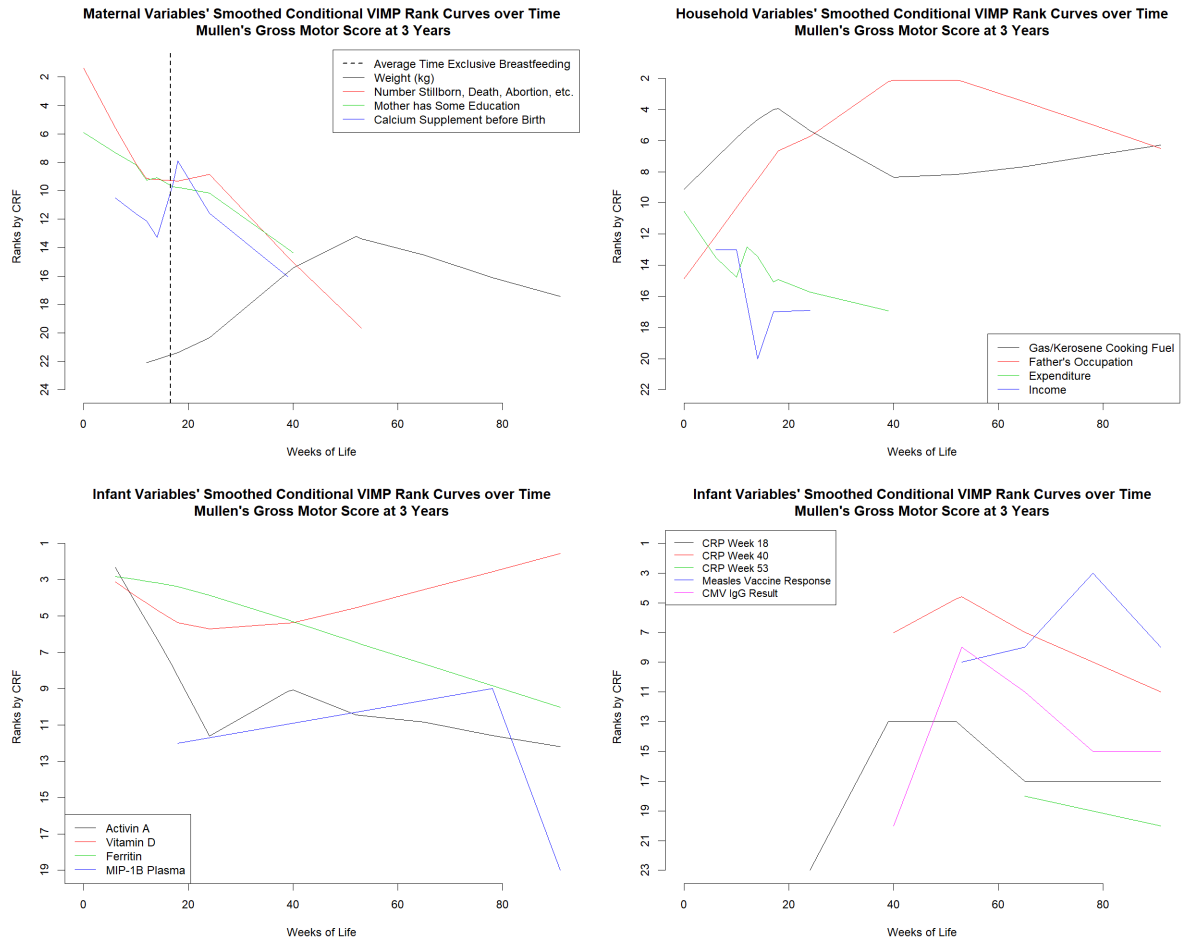

Figure E: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Mullen's Gross Motor at three years of age.

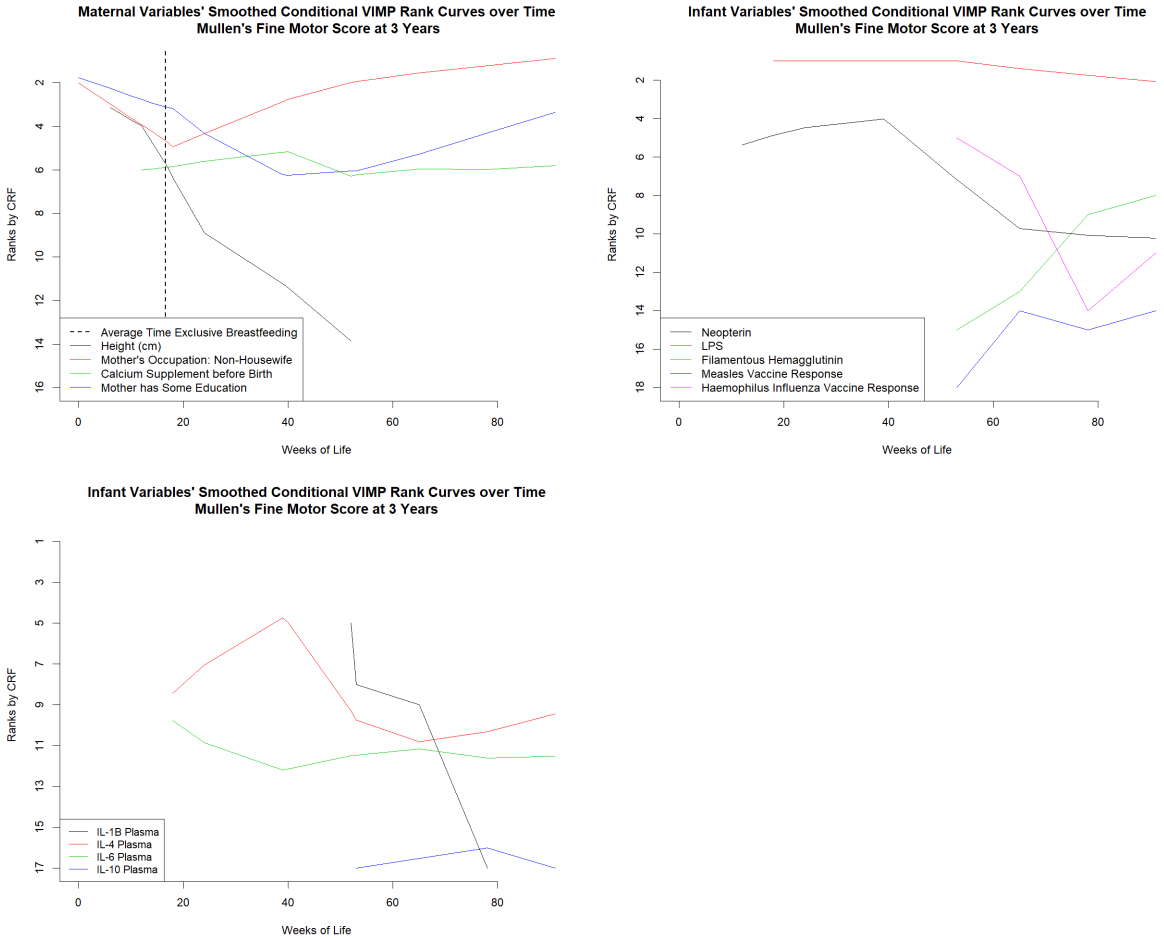

Figure F: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Mullen's Fine Motor at three years of age.

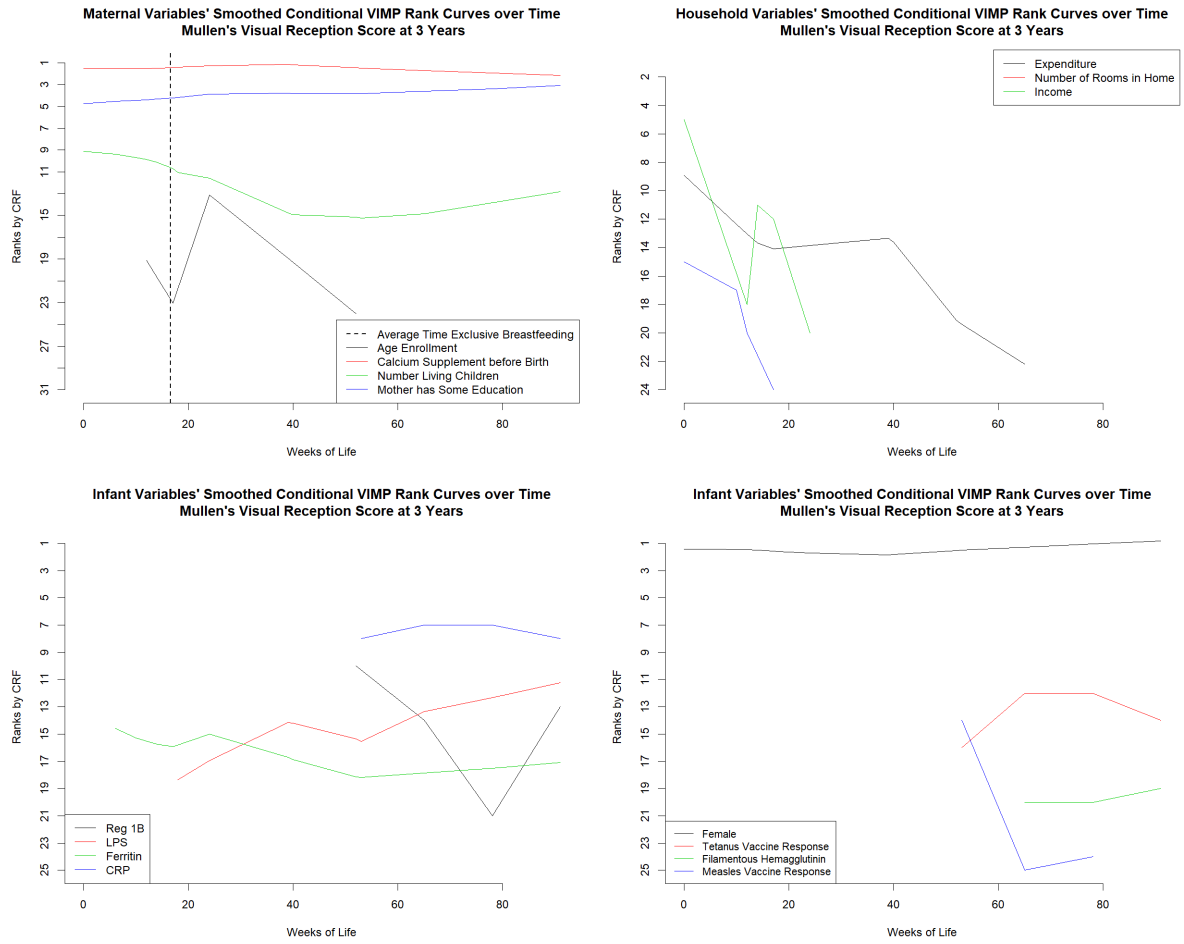

Figure G: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Mullen's Visual Reception at three years of age.

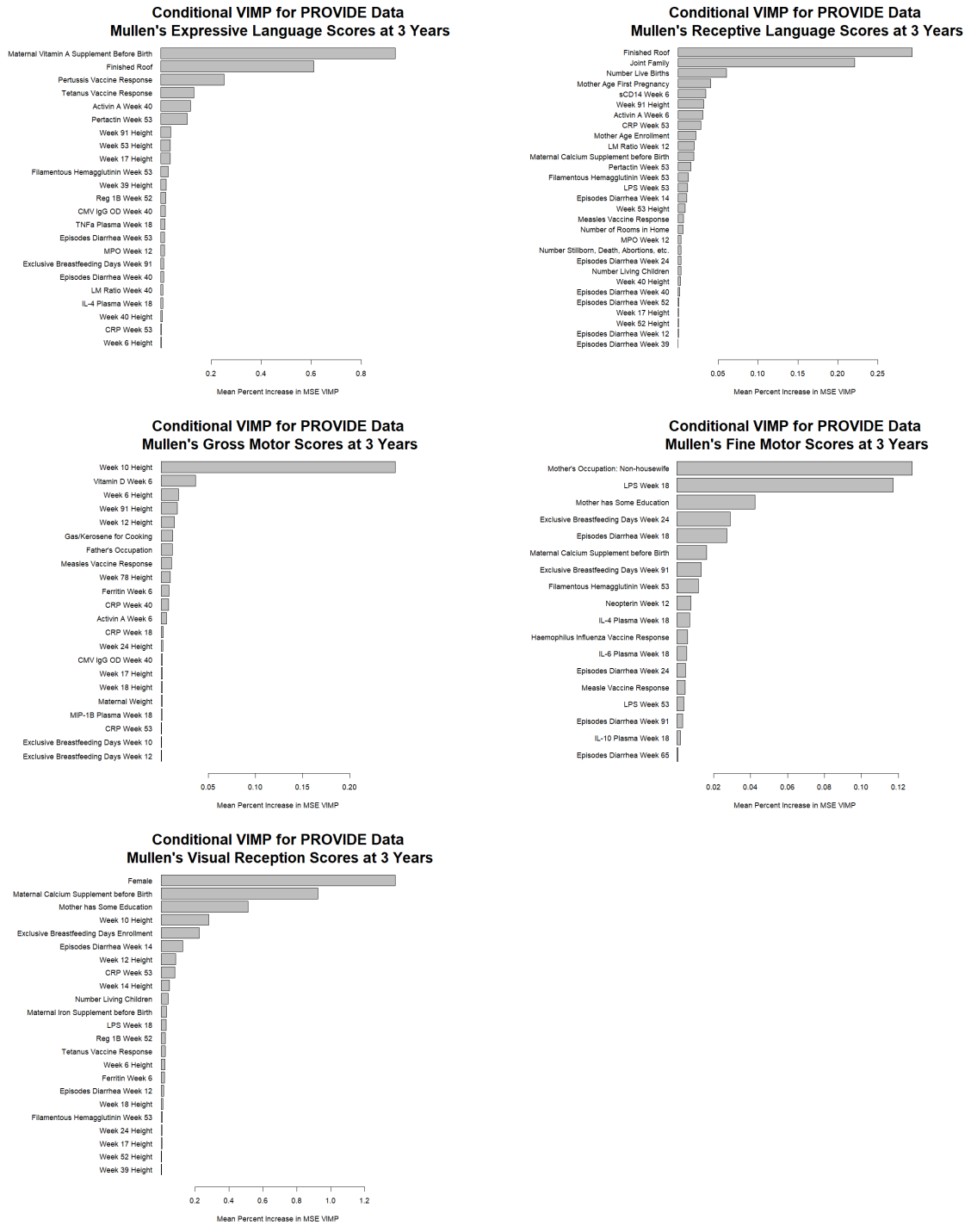

Figure H: Predicting Mullen's at three years, the conditional VIMP at week 91 for predictors with positive conditional VIMP.

Supplemental Figures: Weschler's Responses

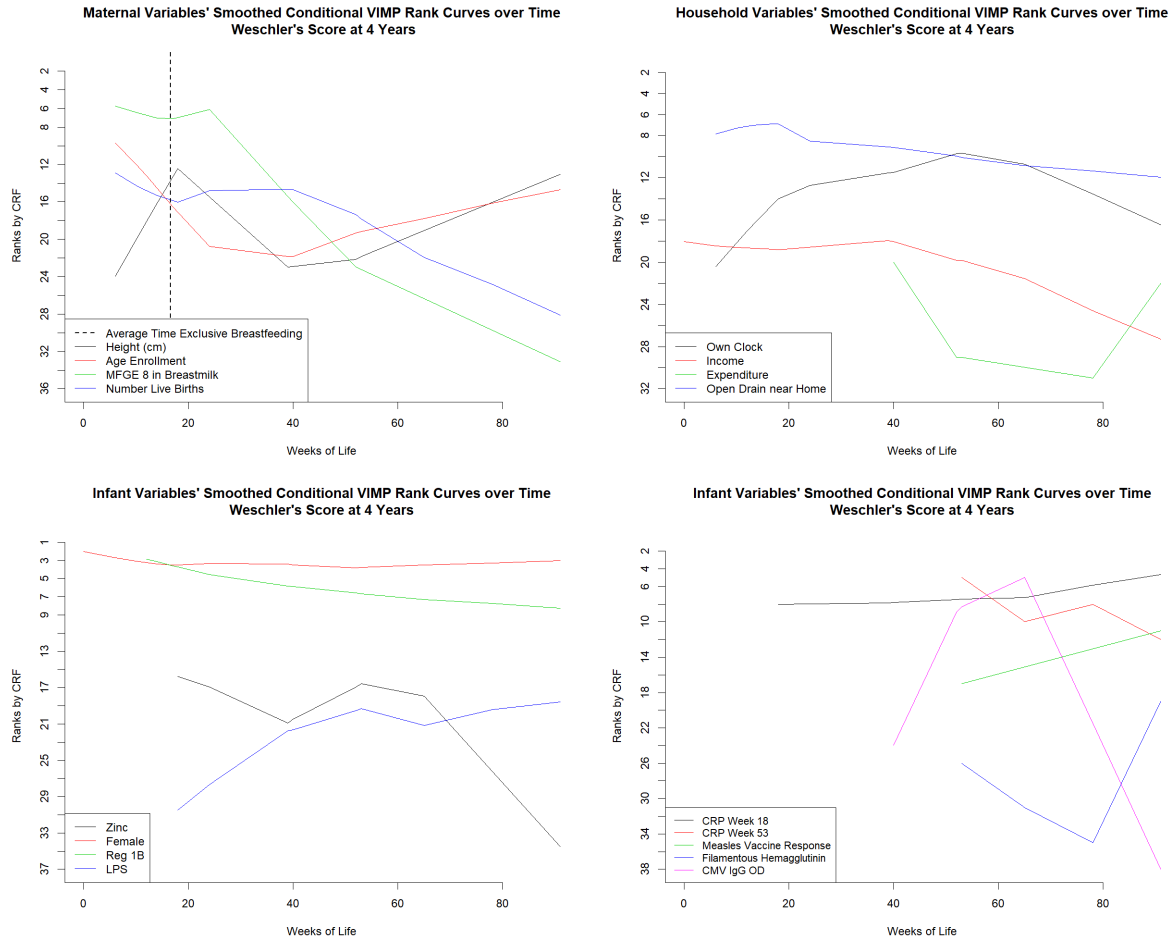

Figure I: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Weschler's at four years of age.

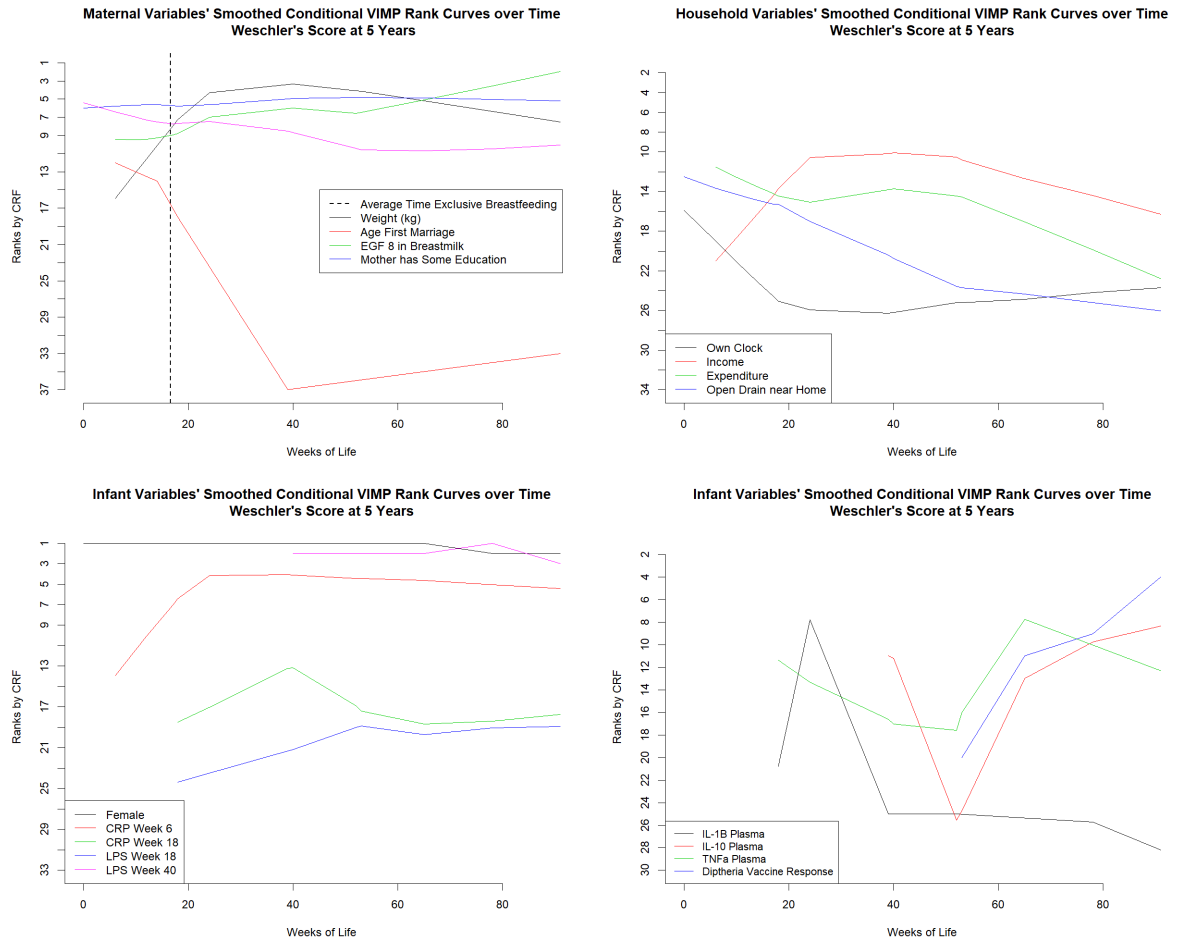

Figure J: Maternal factors' (top left), household variables' (top right), and other infant's measures' (bottom two) conditional VIMP rank over time predicting Weschler's at five years of age.

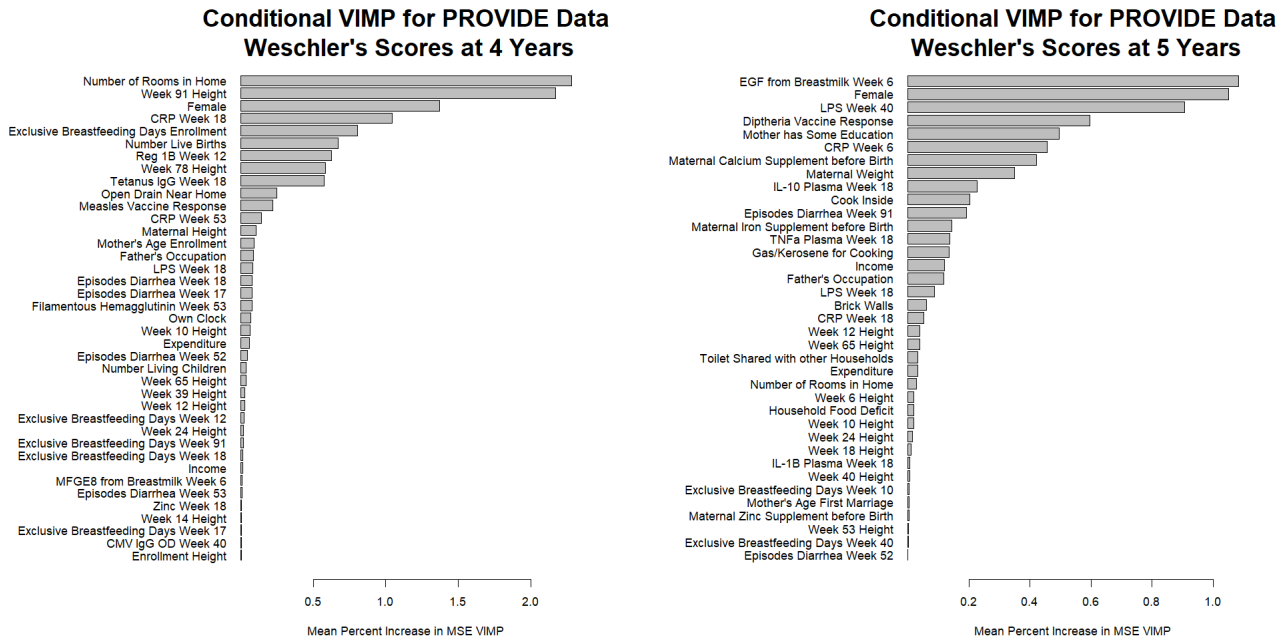

Figure K: Predicting Weschler's at four and five years, the conditional VIMP at week 91 for predictors with positive conditional VIMP.