Summer 2023

Approaches to Detecting and Modeling Over-and Underdispersion in Alternative Count Data Distributions and an Application of Logistic Regression and Random Forest Modeling to Improve Screening Tools for Tic Disorders in Children

Rebecca C. Wardrop

Approaches to Detecting and Modeling Over- and Underdispersion in Alternative Count Data Distributions and an Application of Logistic Regression and Random Forest Modeling to Improve Screening Tools for Tic Disorders in Children

by

Rebecca C. Wardrop

Bachelor of Arts
College of Wooster 2014

Master of Science
University of South Carolina 2016

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Biostatistics

Arnold School of Public Health

University of South Carolina

2023

Accepted by:

James W. Hardin, Major Professor

James R. Hussey, Committee Member

Robert R. Moran, Committee Member

James F. Thrasher, Committee Member

Ann Vail, Dean of the Graduate School

Acknowledgments

As I move out of this phase of my life, I am grateful for all of the support I've received and lessons I've learned academically, profesionally, and personally from the many people who have supported me along the way.

This endeavor would not have been possible without Dr. James W. Hardin, my committee chair and advisor. He provided invaluable advice and guidance. He was patient with me and gave me the confidence to advocate for myself and complete my studies. I'd also like to extend my sincerest thanks to Dr. James R. Hussey for his support throughout my graduate studies as well as his service on my committee. His class was one of the first I took at this university and his skill as a teacher assured that I had made the correct decision in coming here. Many thanks as well to my other committee members, Dr. Robert R. Moran and Dr. James F. Thrasher. Their questions throughout the process helped shape my dissertation.

I would also like to thank Dr. Adam B. Lewin, Dr. Heather R. Adams, Dr. Jennifer A. Vermilion, Dr. Steven P. Cuffe, Melissa L. Danielson, Dr. Rebecca H. Bitsko, and Dr. Bo Cai for their participation in my final dissertation project. Additional thanks to Deborah Salzberg for her support on that project.

Finally, I would like to acknowledge the role of my family. First, thank you to my husband, Jake, for being willing to forgo advancement in his own career to support my desire to continue my studies. I am deeply appreciative of his partnership and look forward to continuing to support each others' dreams in the future. As for the rest of my family, I will never be able to repay them for all of their support through my many years in school; they were understanding when I needed to take a step back

and encouraging when I needed the extra push to continue. A special thanks to my dad for his frequent reminders that a PhD is primarily about perseverance. He spent hours helping with homework, editing papers, and answering questions. Without his willingness to always lend a listening ear and a helping hand, I would not have made it anywhere near as far in my academic career.

Abstract

This dissertation focuses on theory and application of discrete data methods, particularly approaches to over- and underdispersion relative to the Poisson distribution and an application of random forest and logistic regression modeling. The first chapter derives a score test for over- and underdispersion in the heaped generalized Poisson distribution. Equi-, over-, and underdispersed heaped generalized Poisson and heaped negative binomial data are simulated to evaluate the performance of the score test by comparing the power it achieves to that of Wald and likelihood ratio tests. We find that the score test we derive performs comparably to both the Wald and likelihood ratio tests. The second chapter explores the application and limitations of a model for the dispersion parameter in the double Poisson distribution utilizing a logistic-like link function. Data are simulated under various dispersion structures and a set of models assuming different maximum dispersion values are estimated for each. Through the simulation and a case study, we assess the application of the proposed model and identify potential improvements to aid in its effective utilization. Finally, the third chapter evaluates the performance of, and identifies important items in, a screening and a diagnostic tool for tic disorders in children. We also compare their results in terms of their ability to correctly predict tics in children and determine that the random forest models are more effective at reducing Type II errors.

Table of Contents

List of Tables

List of Figures

Chapter 1

Derivation of a Score Test for Over- and Underdispersion Based on the Heaped Generalized Poisson Distribution

Two of the most common problems encountered when modeling count data are heaping and overdispersion. Heaping occurs when subjects fail to report exact counts and arises due to subjects' digit preferences or their tendency to round estimations. As heaping is a measurement error, it may lead to biased estimation due to increased variance if left unaddressed.

The term overdispersion indicates that the variance exceeds the mean, violating the assumption of equidispersion inherent in the Poisson distribution, that is $\mathrm{E}(Y)=\mathrm{Var}(Y)$. Failure to properly address excess variance can lead to underestimation of standard errors and, thus, incorrect inference about regression parameters. Although the Poisson model is the standard approach to analyze such count data, there are no inherent mechanisms in the distribution to deal with either phenomena.

One proposed technique to handle heaped count data is to utilize a mixture of re-scaled distributions [13]. A mixture of re-scaled Poisson distributions can be used in the case of equidispersion, however, in the presence of overdispersion, an alternative distribution must be used. Although there are many possibilities [19, 40, 11, 36], the distribution that will be the focus of this paper is the generalized Poisson [10].

To justify the use of a mixture of rescaled generalized Poisson distributions rather than a mixture of rescaled Poisson distributions, analysts must present evidence

against the assumption of equidispersion. Likelihood ratio and Wald tests provide two possible justifications, however Rao's score test has the advantage of only requiring estimation under the null hypothesis [32, 33]. Therefore, we present a derivation of the score test for overdispersion based on a heaped generalized Poisson model and evaluate the performance of the test's power through simulation.

1.1 Heaped Generalized Poisson

The heaped generalized Poisson is a mixture distribution comprising of $m+1$ mixture components and their associated weights. The mixture components are the generalized Poisson distributions P_C with mean μ and dispersion parameter α modeling reporting behaviors $B = 1, ..., m+1$. Their corresponding mixture weights assume a multinomial probability function given by P_M, which models the probability of each behavior.

The probability mass function (PMF) for the generalized Poisson is given by,

$$P_C(Y_i = y_i) = \frac{\theta_i(\theta_i + \alpha y_i)^{y_i - 1} \exp(-\theta_i - \alpha y_i)}{y_i!}, y_i = 0, 1, 2, ... \tag{1.1}$$

where $\theta_i > 0$ and $\max(-1, -\theta_i/4) < \alpha < 1$. The mean and variance are then,

$$\mathrm{E}(Y_i) = \mu_i = \frac{\theta_i}{1 - \alpha}$$

$$\mathrm{Var}(Y_i) = \frac{\theta_i}{(1 - \alpha)^3} = \frac{1}{(1 - \alpha)^2} E(Y_i)$$

From the noted relationship between the mean and variance, it is clear that when $\alpha > 0$, there is overdispersion relative to the mean and when $\alpha < 0$ there is underdispersion. We can also see that when $\alpha = 0$, the distribution reduces to the Poisson with parameter θ_i.

Behaviors $B = 2, ..., m+1$ correspond to reporting on the heaping values $k = k_2, ..., k_{m+1}$. Thus, a respondent under behavior m would provide a response that is heaped on a multiple of k_m. Note that, given the notation, we assume m different known heaping values. Under behavior 1, respondents provide exact counts. This can also be described as heaping on $k = 1$.

Suppose we have a vector of count responses $\mathbf{Y} = (Y_1, ..., Y_n)$ where Y_i and Y_j are independent and identically distributed for all $i \neq j$. To estimate the mixture components, we associate covariates $\mathbf{X}$ and parameters $\boldsymbol{\beta}$ with the mean μ_b via the log-link function, $\log(k_b\mu_b)$=$\mathbf{X}$ $\boldsymbol{\beta}$. Note that this is equivalent to $\log(\mu_b)$=$\mathbf{X}$ $\boldsymbol{\beta} - \log(k_b)$, thus the link functions for all behaviors are the same save for the offset term $\log(k_b)$. The mixture weights are estimated via a multinomial model using covariates $\mathbf{Z}$ and parameters $\boldsymbol{\gamma}_b$ [20]. Taking all this, the probability model is given by,

$$P(Y = y) = \sum_{b=1}^{m+1} P_{M_b}(B = b|\mathbf{Z}, \boldsymbol{\gamma}_b) P_{C_b}(Y = \frac{y}{k_b}|\mu_b = \exp(\mathbf{X}\boldsymbol{\beta} - \log(k)), \alpha) I_{k_b}, \quad (1.2)$$

where

$$I_{k_b} = I(y \bmod k_b = 0),$$

$$P_{M_1}(B = 1|\mathbf{Z}, \boldsymbol{\gamma}_b) = \frac{1}{1 + \exp(\mathbf{Z}\boldsymbol{\gamma}_2) + ... + \exp(\mathbf{Z}\boldsymbol{\gamma}_{m+1})}, \text{and}$$

$$P_{M_b}(B = b|\mathbf{Z}, \boldsymbol{\gamma}_b) = \frac{\exp(\mathbf{Z}\boldsymbol{\gamma}_b)}{1 + \exp(\mathbf{Z}\boldsymbol{\gamma}_2) + ... + \exp(\mathbf{Z}\boldsymbol{\gamma}_{m+1})}, b = 2, ..., m+1.$$

1.2 Derivation of the Score Test

Suppose $Y = Y_1, Y_2, ..., Y_n$, is an independent, identically distributed sample from the heaped GP density function heaped on multiples of $k_1, k_2 ... k_{m+1}$ where $k_1 = 1$. The log-likelihood is then,

$$\begin{aligned}\mathcal{L} &= \sum_{i=1}^{n} \log \left\{ \sum_{b=1}^{m+1} P_{M_b}(B = b|\mathbf{Z}, \boldsymbol{\gamma}_b) P_{C_b}(Y = \frac{y}{k_b}|\mu_b = \exp(\mathbf{X}\boldsymbol{\beta} - \ln(k), \alpha) I_{k_b} \right\} \\ &= \sum_{i=1}^{n} \log \left\{ \sum_{b=1}^{m+1} P_{M_b} P_{C_b} I_{k_b} \right\} \\ &= \sum_{i=1}^{n} \log L_i. \end{aligned} \tag{1.3}$$

Taking derivatives in terms of α gives,

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} \frac{\partial P_{C_b}}{\partial \alpha} I_{k_b} \tag{1.4}$$

$$= \sum_{i=1}^{n} \left\{ \frac{\sum_{b=1}^{m+1} P_{M_b} I_{k_b} P_{C_b} A_b}{\sum_{b=1}^{m+1} P_{M_b} I_{k_b} P_{C_b}} \right\} \tag{1.5}$$

and

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} = \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} \frac{\partial^2 P_{C_b}}{\partial \alpha^2} I_{k_b} - \sum_{i=1}^{n} \Big[\frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} \frac{\partial P_{C_b}}{\partial \alpha} I_{k_b} \Big]^2 \tag{1.6}$$

$$= \sum_{i=1}^{n} \left\{ \frac{\sum_{b=1}^{m+1} P_{M_b} I_{k_b} P_{C_b} (A_b^2 - B_b)}{\sum_{b=1}^{m+1} P_{M_b} I_{k_b} P_{C_b}} \right\} - \sum_{i=1}^{n} \left\{ \frac{\sum_{b=1}^{m+1} P_{M_b} I_{k_b} P_{C_b} A_b}{\sum_{b=1}^{m+1} P_{M_b} I_{k_b} P_{C_b}} \right\}^2, \tag{1.7}$$

where

$$A_b = \frac{-1}{1-\alpha} + (\mu_b - y_b) + \frac{(y_b - 1)(y_b - \mu_b)}{\mu_b(1-\alpha) + \alpha y_b} \tag{1.8}$$

$$B_b = \frac{1}{(1-\alpha)^2} + \frac{(y_b - 1)(y_b - \mu_b)^2}{[\mu_b(1-\alpha) + \alpha y_b]^2} \tag{1.9}$$

The score test statistic for testing $\mathrm{H}_0 : \alpha = 0$ against the alternative $\mathrm{H}_1 : \alpha \neq 0$ is then,

$$S(\alpha) = \frac{\Big[\frac{\partial \mathcal{L}}{\partial \alpha}\Big]^2}{E\Big[-\frac{\partial^2 \mathcal{L}}{\partial \alpha^2}\Big]}, \tag{1.10}$$

evaluated at $\beta = \widehat{\beta}$, $\gamma = \widehat{\gamma}$ and $\alpha = 0$. Under these conditions, P_{C_b} is the probability function of the Poisson with mean $\mu = \widehat{\mu}$ and

$$A_b = -1 + (\widehat{\mu}_b - y_b) + \frac{(y_b - 1)(y_b - \widehat{\mu}_b)}{\widehat{\mu}_b} \tag{1.11}$$

$$B_b = 1 + \frac{(y_b - 1)(y_b - \widehat{\mu}_b)^2}{\widehat{\mu}^2}. \tag{1.12}$$

If we consider the case in which there is no heaping ($m = 0$), the score test statistic simplifies to

$$\frac{1}{2n}\Big[\sum_{i=1}^{n} \frac{(y-\mu)^2 - y}{\mu}\Big]^2, \tag{1.13}$$

matching the result given in Yang et al and, thus, extending it to the case of heaping [43]. Full details of the score test statistic calculation are provided in the Appendix.

1.3 Simulation Study

The simulation study conducted in Stata seeks to compare the proposed score test with the Wald and likelihood ratio tests [38]. The **heapr** package in Stata was modified to include our derived score test [13]. Empirical powers of the score, Wald, and likelihood ratio tests for the dispersion parameter in the heaped generalized Poisson model are examined. Data are simulated based on a heaped generalized Poisson distribution (both over- and underdispersed) and a heaped negative binomial distribution (only overdispersed).

Data were simulated using the heaped generalized Poisson distribution and the heaped negative binomial distribution such that,

$$k_b \mathrm{E(Y)} = k_b \mu_b = \exp(1 + 0.25x_1 - 0.25x_2) \tag{1.14}$$

.

where x_{1i} and x_{2i} are generated from continuous uniform $[0, 1]$ distributions. We simulate two heaping behaviors $k_2 = 3$ and $k_3 = 4$. Recall $k_1 = 1$, corresponding to reports of true values.

Using a series of different values of the dispersion parameter, α, in the heaped generalized Poisson and the probability of success on a single trial, p, in the negative binomial, samples of size $n = 100, 250,$ and 500 are taken from both the heaped generalized Poisson and the heaped negative binomial distributions. Power estimations for each situation are based on 10,000 replications.

To test the presence of over- or underdispersion in the data, we fit a heaped Poisson and a heaped generalized Poisson regression model with the same covariates, x_1 and x_2, then conduct Wald and likelihood ratio tests as well as a score test based on equation (1.10). The significance level of each test is set as $\alpha = 0.05$. As we are testing for both over- and underdispersion (H_0 : vs. H_1 :), this is a two-sided test and the asymptotic distribution is χ^2_1. Thus, the critical value is $\chi^2_{1,1-\alpha} = 3.84$ under the nominal level. The empirical power of the tests can be calculated as,

$$\frac{\#(S > \chi^2_{1,1-\alpha})}{R},$$

which is the proportion of times S is greater than our critical value based on $R = 10,000$ replications. Results for the empirical power calculations for the data simulated based on a heaped generalized Poisson distribution are provided in Table A.1 and Figure A.1.

When the dispersion parameter $\alpha = 0$, that is the true model is a heaped Poisson model, we expect the power of the tests to be equal to the significance level, 0.05. Results suggest that the score test tends to reject more often than expected in this case, indicating a preference toward the heaped generalized Poisson relative to the Wald and likelihood ratio tests. Larger sample sizes dampen this preference, however.

As dispersion values increase, we expect the power of the tests to increase as well.

This is generally true, however, when sample sizes are small, we see slight decreases in power for all three tests when there are small increases in the dispersion parameter (e.g. $\alpha = 0.01$ relative to $\alpha = 0$). This effect is lessened by increases in the sample size. For all tests and sample sizes, $n = 100, 250$ and 500, power begins steadily increasing when $\alpha = 0.06$.

Note we see that the likelihood ratio test tends to outperform the Wald and score tests as the dispersion parameter, α, increases. The Wald and score tests perform comparably to each other, however the score test outperforms the Wald test when the sample size is small ($n = 100$). This may suggest that the likelihood ratio test is best for detecting overdispersion in a heaped generalized Poisson relative to the heaped Poisson. However, as noted above, the likelihood ratio test requires estimation of both the heaped Poisson and heaped generalized Poisson models making it more computationally expensive.

A similar pattern of increase in power was observed for increasingly negative values of dispersion (indicating underdispersion relative to the heaped Poisson). These data are not reported to avoid redundancy.

Table A.2 and Figure A.2 display empirical power results for the three tests conducted based on the data simulated using a heaped negative binomial distribution. The negative binomial distribution converges to the Poisson as the probability of success, p, approaches 1. Thus, we expect the power of the three tests to be approximately equal to the significance level, 0.05, when p is large. As we can see, when $p = 0.99$, the power of the tests under all sample sizes is near 5%. As p increases, the power of the tests also increase. The power of the likelihood ratio test, again, increases at a faster rate than that of the Wald and score tests. The Wald and score tests perform comparably to each other.

1.4 Case Study

A case study was conducted using the National Health and Examination Survey (NHANES) 2009-2010 data [18]. We model the average number of cigarettes smoked per day during the past 30 days as a function of the participant's age, gender, and race. The original variable **ridereth1** was recoded to be an indicator variable for non-Hispanic white race versus other races. The **riagendr** variable was also recoded to be an indicator variable for gender. Characteristics of the variables included in our model are provided in Table A.3.

To evaluate the presence of heaping, we created a histogram of the frequency for number of cigarettes smoked during the past 30 days. This plot is provided in Figure A.3. From the plot, we can see heaps occur on multiples of five. Therefore, we specify a heaping value of five in our models.

We began by estimating a heaped Poisson regression model. The results of this regression model are provided in Figure A.4. As this model is considered the restricted model, we conducted a score test of $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$. We then estimated a heaped generalized Poisson regression model, the results of which are provided in Figure A.5. Using this model, we conducted a Wald test for over- or underdispersion relative to the heaped Poisson (i.e. $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$). Finally, we compare the log-likelihood of each model in a likelihood ratio test to evaluate the same hypotheses. Results of all three tests are provided in Table A.4. All three tests indicated that the dispersion parameter was significantly different than 0, suggesting over- or underdispersion relative to the heaped Poisson with heaping behavior at multiples of five.

1.5 Discussion

In this paper, we established that the score test derived within performs comparably to the Wald and likelihood ratio tests. Using simulated data, we evaluated the performance of all three tests in terms of their power at different values of the dispersion parameter, α, for heaped generalized Poisson distributed data and different values of the probability of success on a single trial, p for heaped negative binomial distributed data. While the likelihood ratio tests consistently achieved higher power than both the Wald and score tests, it was only slightly higher and it has the downside of requiring estimation of both the full (heaped generalized Poisson) model and the restricted (heaped Poisson) model. Similarly, the Wald tests requires estimation of the full model.

Although all three tests are appropriate in situations in which the dispersion structure is assumed to be a scalar, the score tests requires only estimation of the restricted model, making it computationally inexpensive in comparison to the other two tests. When estimating a heaped Poisson model, the score test can and should be conducted to determine if the more complex heaped generalized Poisson would be more appropriate. In the case that the score test fails to reject the null, it can confidently be concluded that the data are equidispersed relative the mean and the heaped Poisson is an appropriate approach to modeling and there is no need to estimate that more complex model. Therefore, we believe our derived score test provides an attractive alternative to both the Wald and likelihood ratio tests in detecting the presence of over- and underdispersion relative the heaped Poisson distribution.

Chapter 2

Modeling the Dispersion Parameter in the Double Poisson Distribution

2.1 Introduction

When presented with count data, the first inclination to model the data is to use regression based on the Poisson distribution. Taking this approach requires the assumption of equidispersion, that is, the mean is assumed to be equal to the variance ($E(Y) = \text{Var}(Y)$). However, for observed data this is often not the case [7] [16]. In the case of over- or underdispersed data, this approach will under- or overestimate standard errors, respectively, and, therefore, lead to incorrect inference about regression parameters.

As this is a considerable limitation, a number of alternative approaches have been proposed allowing the variance to vary independent from the mean [11] [36] [10] [40]. One such distribution is the double Poisson [19]. In his paper detailing the double exponential family of distributions including double Poisson, Efron provides a generalized linear model framework relating a linear model to not only the count outcome but also the dispersion parameter.

In this paper, we explore the application and limitations of Efron's proposed model for the dispersion parameter. Via a simulation study using Stata software developed for this paper, we test the model for different structures of the dispersion parameter. This includes simulating data with constant dispersion, data that are over- or underdispersed and the value of the dispersion parameter changes with the

value of an independent variable, and data that goes from over- to under-dispersed dependent on the value of an independent variable. We hope to determine the role of the statistician selected maximum value of dispersion, M, and whether Efron's suggested approach to M is appropriate in practice.

2.2 Double Poisson Distribution

The double Poisson distribution was first proposed by Efron [19] as a member of the double exponential family of distributions. The distribution is a combination of two Poisson distributions, $P(\mu)$ and $P(y)$, so that

$$f_{\mu,\phi}(y) = c(\mu,\phi) f *_{\mu,\phi} (y) = c(\mu,\phi)\phi^{\frac{1}{2}}[P(\mu)]^{\phi}[P(y)]^{1-\phi} \tag{2.1}$$

where $c(\mu,\phi)$ is a normalizing constant that depends on μ and ϕ. This constant is nearly equal to 1 and is included to ensure the density sums to unity. Thus, the exact probability mass function is given by,

$$P(Y=y) = f_{\mu,\phi}(y) = c(\mu, \phi f *_{\mu,\phi} (y) = c(\mu,\phi)\left(\phi^{\frac{1}{2}} e^{-\phi\mu}\right)\left(\frac{e^{-y}y^{y}}{y!}\right)\left(\frac{e\mu}{y}\right)^{\phi y}, \tag{2.2}$$

where $y = 0, 1, 2, ...$ Because this constant is an infinite series, $\sum_{y=0}^{\infty} f *_{\mu,\phi} (y)$, an exact closed form is not available [19] [45].

Efron[19] proposed an approximation for this normalizing constant based on the a three-term Edgeworth expansion. The closed form approximation is then,

$$\frac{1}{c(\mu,\phi)} = \sum_{y=0}^{\infty} f *_{\mu,\phi} (y) \approx 1 + \frac{1-\phi}{12\mu\phi}\left(1 + \frac{1}{\mu\phi}\right) \tag{2.3}$$

This closed-form approximation performs well, however the estimations produced are not exact [22] [42]. The approximation provided by the Edgeworth expansion becomes particularly unreliable when μ is small. For example, when $\mu = .1$ and $\phi = 2$, the approximation is negative ($\frac{1}{c(\mu,\phi)} = -1.50$).

Due to this limitation, researchers have taken alternative approaches to the normalizing constant. One approach is to ignore the constant entirely. Although this provides a simplified PMF, the sum of the likelihoods will no longer equal unity. Zhu [44] evaluated the performance of the double Poisson without the normalizing constant and found that while it captured the mean well, it did not provide accurate estimates of the variance.

Another approach is to approximate $c(\mu, \phi)$ via a k-th partial sum,

$$\frac{1}{c(\mu, \phi)} = \sum_{y=0}^{\infty} f_{\mu,\phi}(y) \approx \sum_{y=0}^{k} f_{\mu,\phi}(y). \tag{2.4}$$

Zou, Geedipally, and Lord [45] recommend that k be no smaller than twice the sample mean. This method involves significantly more computing power than ignoring the constant or utilizing the closed from approximation provided by the Edgeworth expansion, however it provides much higher accuracy in estimations.

The mean and variance referring to the exact density are

$$\mathrm{E}(Y) = \mu + O(n^{-2}) \approx \mu = g(\mathbf{x}\boldsymbol{\beta})$$

$$\mathrm{Var}Y) = \frac{\mu}{\phi}[1 + O(n^{-2})] \approx \frac{\mu}{\phi},$$

where $\phi > 0$. Notice if ϕ is equal to one, the distribution reduces to the Poisson, signifying that the data are equidispersed. When ϕ is greater than one, the variance is less than the mean implying that the data are underdispersed. Conversely, when ϕ is less than one, the variance is greater than the mean and the data are overdispersed.

2.3 Regression Models

2.3.1 Modeling Count Outcomes

To utilize this distribution in a generalized linear model framework, suppose $Y = Y_1, Y_2, ..., Y_n$ represents an independent, identically distributed sample from a double Poisson distribution with parameters μ and ϕ. The log-likelihood function is then given by

$$\mathcal{L}(\mu, \phi|Y) = \sum_{i=1}^{n} \Big[\frac{1}{2}\ln(\phi) - \phi\mu - y_i + y_i \ln y_i - \ln\Gamma(y_i + 1) + \phi y_i(\ln\mu - \ln y_i + 1) - \ln(c(\mu, \phi))\Big] \tag{2.5}$$

The expected outcome can then be linked to the independent variables $\mathbf{x}$ via the link function

$$\mathrm{E}(Y) \approx \mu = \exp(\mathbf{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\beta}$ is the vector of coefficients to be estimated. This is the same link function used in traditional Poisson regression and allows for a familiar interpretation of the coefficients.

2.3.2 Modeling Dispersion

To model the dispersion parameter, ϕ, Efron suggests utilizing a logistic-like regression model using the link-function,

$$\phi = \frac{M}{1 + \exp(-\mathbf{s}\boldsymbol{\alpha})}, \tag{2.6}$$

where $\mathbf{s}$ are the covariates, $\boldsymbol{\alpha}$ is the vector of coefficients to be estimated, and $M > 0$ is the maximum value the dispersion parameter can attain [19]. Efron refers to this

link function as logistic-like because it closely resembles the link function used in logistic regression,

$$p = \frac{1}{1 + \exp(-\mathbf{s}\boldsymbol{\alpha})}.$$

In logistic regression this link function is motivated by a desire to ensure the probability of interest remains between zero and one. When modeling the dispersion parameter, Efron suggests we have a similar desire to ensure that the dispersion parameter remains positive but still small. Thus, Efron suggests keeping M above one but low, to avoid having the Poisson distribution $M = 1$, on the edge of the parameter space [19]. Although there is no straightforward interpretation of the coefficients, such a model allows for investigation into the relationship between covariates and dispersion.

2.4 STATA Syntax

The syntax for the **dpoisson** command is a familiar construction similar to other regression commands in Stata, including the **poisson** command [38]. The syntax is

dpoisson *depvar [indepvars] [if] [in] [weight] [,options]*

which is the command to fit a double Poisson regression of *depvar* on *indepvars*, where *depvar* is a non-negative count variable. There are also a set of qualifiers and options that are familiar to users of Stata. Including the *if* qualifier requires an expression and restricts the command to values for which the expression is true. Similarly, the in qualifier restricts the command to values within the provided range. The *weight* qualifier indicates the weight that should be attached to each observation.

2.4.1 Options

In addition to options that are familiar to users of the **poisson** command in Stata, there are also a set of options that are unique to the **dpoisson** command. A list of covariates to be included in the model for variable dispersion can be provided with the optional **vd(itvarlist)**. If the option is not included, the model estimates the dispersion, ϕ, as a constant.

The maximum value, M, of the dispersion parameter is set by **efron**(#). If no M value is provided (i.e. **efron**()), the user will receive an error message as the default value is set to -1. If **efron**(#) is not stated at all, the link function will be treated as a log-link.

Three options are provided for estimating the normalizing constant, c. The default approximation, **sca**, is the estimation provided in equation 2.3. The second, **sce**, uses the finite sum approximation suggested by Zou, Geedipally, and Lord [45]. Finally, the constant can be set equal to 1 using **sc1**.

2.5 Simulation Study

A simulation study was conducted to test the application and limitations of the proposed dispersion model. Each simulation was run with 10,000 repetitions for sample sizes of 100, 250, and 500 observations. To evaluate the performance of the model under a variety of dispersion structures, data were simulated to be overdispersed with ϕ dependent on an independent variable, underdispersed with ϕ dependent on an independent variable, over- to underdispersed with ϕ dependent on an independent variable, and with constant dispersion. The count data outcomes were simulated using the double Poisson distribution such that,

$$\mathrm{E}(Y) \approx \mu = \exp(\mathbf{X}\boldsymbol{\beta}) = \exp(1 - 1.5x_1 + 0.25x_2),$$

where $x_1 \sim$ Bernoulli(0.5) and $x_2 \sim$ Uniform(0,1). The different structures of ϕ used to simulate data are provided in Table B.1. Under each of these dispersion structures, four different models were estimated. Three of these models involved setting the maximum allowable value of the dispersion, M, to values 1.25, 2, and 2.5. The fourth model assumed constant dispersion. Each of these models was repeated for the three different approaches to calculating the normalizing constant using options **sc1**, **sca**, and **sce**.

2.5.1 Overdispersed Simulation

The first set of data was simulated using dispersion,

$$\phi = \frac{2}{1 + \exp(-(-1 + 0.5x_1))},$$

where $x_1 \sim$ Bernoulli(0.5) as above. Note that the true values of the coefficients, t_0 and t_1 are -1 and 0.5, respectively, when $M = 2$. Because x_1 can only take on values of 0 or 1, $\phi \approx 0.538$ when $x_1 = 0$ and $\phi \approx 0.755$ when $x_1 = 1$. Thus, dispersion is dependent on x_1 and data are consistently overdispersed relative to the Poisson distribution. The mean values and 95% confidence intervals for t_0 and t_1 or ϕ, where appropriate, for each of the models estimated for this set of simulated data are provided in Table B.2.

One notable result is the very large standard errors for t_1 under some model conditions, particularly when the maximum allowable value of dispersion, M, was small (1.25). This occurs for all sample sizes when calculating the normalizing constant using Efron's approximation or the k-th partial sum approach. When the normalizing constant was assumed equal to 1, the standard error for t_1 was very large only for sample size $n = 100$. Standard errors were also large for t_1 for all tested values of M when sample size was $n = 100$ and the normalizing constant was calculated using the k-th partial sum. Finally, the standard error was large for t_1 when $M = 1.25$, sample size

was $n = 100$, and the normalizing constant was calculated using the Efron's approximation. All models treating ϕ as a constant estimated $\phi < 1$, correctly indicating overdispersion relative to the Poisson.

For the models estimated setting the maximum allowable value of dispersion to be $M = 2$, we expect the values of the coefficients to be equal to the values used to simulate the data. This proved true for the models estimated using the k-th partial sum approach to calculating the normalizing constant. For these models, the mean value of t_0 and t_1 ranged from 0.513 and -0.989, respectively when $n = 500$ to 0.588 and -0.929, respectively when $n = 100$, slightly overestimating the true values. Despite the slight over-estimations, the true values of t_0 and t_1 were contained within the 95% confidence intervals. Models estimated assuming the normalizing constant was 1 also performed well, slightly underestimating the value of t_0 and overestimating the value of t_1 by approximately 0.21 for all sample sizes, however the true value of t_1 was not contained within the 95% confidence interval for larger sample sizes, $n = 250$ and $n = 500$.. Models estimated using Efron's estimation of the normalizing constant performed the worst. Such models consistently overestimated the value of t_0 by over 0.20 and overestimating the value of t_1 by over 1.25.

Due to the lack of a straightforward interpretation of the coefficients, the values of the dispersion parameter, ϕ, for the possible values of x_1 are provided in Table B.3. Dispersion when $x_1 = 0$ was estimated fairly accurately by all models, with models using the k-th partial sum approach for calculating the normalizing constant performing the best and models using Efron's estimation performing the worst. Similarly, the models estimated using the k-th partial sum approach for the normalizing constant also performed best in estimating dispersion when $x_1 = 1$ and the models using the Efron estimation performed the worst. In fact, all such models, on average, estimated the dispersion to be greater than one, incorrectly indicating that data would be underdispersed, when $x_1 = 1$. Despite the drawbacks of assuming the nor-

malizing constant equals one, models using this approach performed relatively well in estimating the dispersion under different values of x_1.

2.5.2 Underdispersed Simulation

The second set of data was simulated such that dispersion was dependent on x_1 and data were underdispersed for all values of x_1. Dispersion was simulated using,

$$\phi = \frac{2}{1 + \exp(-(1 + 0.5x_1))},$$

where $x_1 \sim$ Bernoulli(0.5) and the true values of the coefficients are $t_0 = 1$ and $t_1 = 0.5$ when $M = 2$. Thus, when $x_1 = 0$, $\phi \approx 1.462$ and when $x_1 = 1$, $\phi \approx 1.635$. The mean values and 95% confidence intervals for t_0 and t_1 or ϕ, where appropriate, for each of the models estimated for this set of simulated data are provided in Table B.4.

Standard errors were very large under some model conditions. Models estimated using smaller maximum allowable values of dispersion, $M = 1.25$, had large standard errors for all sample sizes. Standard errors remained large for models in which the normalizing constant was estimated by the k-th partial sum and the Efron estimation even when using the larger $M = 2$. Standard errors were more reasonable for all models estimated using $M = 2.5$ when sample sizes were $n = 500$. All models treating ϕ as a constant estimated $\phi > 1$, correctly indicating underdispersion relative to the Poisson.

For models estimated using $M = 2$, we expect the mean value of the parameters, t_0 and t_1, to be near the values used to simulate the data, however, this was not often the case. Unlike in the case of data with consistent overdispersion discussed above, the models estimated assuming the normalizing constant was equal to 1 did not provide mean values of the parameters approximately equal to their true value. Additionally, the mean value of t_1 was negative for all sample sizes, indicating that that the level

of underdispersion decreases as x_1 increases, which is contrary to the true dispersion structure. The models estimated using $M = 2$ and the k-th partial sum for the normalizing constant also did not accurately estimate the value of the parameters. Both t_0 and t_1 were consistently overestimated in these models. The model that performed the best in terms of the parameter estimates for this dispersion structure was the model using Efron's approximation for the normalizing constant when sample sizes were large $n = 500$. The true values of both t_0 and t_1 were contained within the appropriate 95% confidence intervals in this model.

The values of the dispersion parameter, ϕ, for the possible values of x_1 are provided in Table B.5. For all models estimated using $M = 1.25$, parameter estimates were large, leading to $\phi \approx 1.250$ for both $x_1 = 0$ and $x_1 = 1$. Recall that the true value of the ϕ is greater than 1.25 for both $x_1 = 0$ and $x_1 = 1$. As previously noted, the models estimated using $M = 2$ and assuming the normalizing constant was one, yielded estimates for t_1 indicating that the dispersion parameter decreases as x_1 increases. This pattern continued for such models using $M = 2.5$. Models using $M = 2$ and $M = 2.5$ and estimating the normalizing constant through either a k-th partial sum or the Efron estimation, all accurately determine that ϕ increases with x_1. The model that most accurately estimated the true values of dispersion for both $x_1 = 1$ and $x_1 = 1$ was the model using $M = 2.5$ and the k-th partial sum approach to the normalizing constant when sample sizes were large ($n = 500$), however the model using $M = 2$ and the Efron estimation of the normalizing constant also performed well for this sample size.

2.5.3 Over- to Underdispersed Simulation

The third set of data was simulated using,

$$\phi = \frac{2}{1 + \exp(-(-1 + 2x_1))},$$

where $x_1 \sim$ Bernoulli(0.5), $t_0 = -1$, and $t_1 = 2$ when $M = 2$. Thus, when $x_1 = 0$, $\phi \approx 0.538$ indicating that data are overdispersed relative to the Poisson distribution. Then, when $x_1 = 1, \phi \approx 1.462$ indicating that data are underdispersed. The mean values and 95% confidence intervals for t_0 and t_1 or ϕ, where appropriate, for each of the models estimated for this set of simulated data are provided in Table B.6.

As observed in previously discussed dispersion structures, the standard errors under some model conditions were very large. Again, this was particularly true for parameter t_1 in models estimated using $M = 1.25$ for all sample sizes. Standard errors for t_1 are also large in models estimated using the k-th partial sum approach to estimating the normalizing constant when using $M = 2$ for all sample sizes and when using $M = 2.5$ for sample sizes $n = 100$ and $n = 250$. They remained large in the model using $M = 2$ and treating the normalizing constant as equal to one when sample size was $n = 100$.

Models which treated ϕ as a constant and either assumed the normalizing constant equals one or utilized a k-th partial sum to estimate the normalizing constant, estimated $\phi < 1$ indicating overdispersion. Utilizing this approach would lead to incorrect conclusions about the pattern of dispersion in the data, as we would not recognize that data are underdispersed relative to the Poisson when $x_1 = 1$. Further, the model treating ϕ as a constant and utilizing the Efron estimation of the normalizing constant, estimated ϕ with one in its 95% confidence interval for sample size $n = 100$. This would lead to incorrectly concluding that the data are Poisson distributed.

All models estimated using the maximum allowable value of dispersion $M = 2$ provided reasonable approximations of t_0 and t_1 for sample size $n = 500$, however corresponding 95% confidence intervals often did not contain the true value of the parameters. The confidence intervals for the parameters in the model estimated using the k-th partial sum approach to the normalizing constant contained the true value

of the parameters for sample sizes $n = 250$ and $n = 500$. Note that the standard errors for t_1 were very large in both of these models. Models estimated assuming that the normalizing constant was one, consistently underestimated the value of t_1 while models using the Efron estimation consistently overestimated the value.

The values of the dispersion parameter, ϕ, when $x_1 = 0$ and $x_1 = 1$ are provided in Table B.7. The models estimated assuming the normalizing constant is one, underestimated the value of the dispersion parameter for both $x_1 = 0$ and $x_1 = 1$ for large sample sizes. Models using the Efron approximation overestimated the dispersion when $x_1 = 0$ for all sample sizes. The models that most accurately estimated the values of ϕ for the different values of x_1 were those estimated using the k-th partial sum approach to the normalizing constant.

2.5.4 Constant Dispersion Simulation

Finally, data were simulated to have constant dispersion, $\phi = 0.75$. The mean values and 95% confidence intervals for t_0 and t_1 or ϕ, where appropriate, for each of the models estimated for this set of simulated data are provided in Table B.8. For all models in which ϕ is not assumed to be a constant, we expect the coefficient $t_1 = 0$ and the value of t_0 to change depending on the value of M. When $M = 2$, we expect $t_0 \approx -0.511$. For models estimated using $M = 1.25$, we expect $t_0 \approx 0.405$. Finally, when models are estimated using $M = 2.5$, we expect $t_0 \approx -0.847$.

Models using the k-th partial sum approach to estimate the normalizing constant performed the best. For large sample sizes, the true values for the coefficients t_0 and t_1 were contained in their 95% confidence intervals, and this was often true for smaller sample sizes as well. Both models assuming the normalizing constant was one and models using the Efron approximation, the value of t_1 was consistently overestimated, such that the 95% confidence intervals did not include one. This would lead us to incorrectly conclude that dispersion is dependent on the value of x_1.

Among the models that correctly treated ϕ as a constant, both the models assuming the normalizing constant is one and the model using the k-th partial sum approach provided estimates of ϕ with 95% confidence intervals including the true value of $\phi = 0.75$. The model using the Efron approximation estimated ϕ to be near one for all sample sizes. When $n = 100$, this approach incorrectly estimated $\phi > 1$, which would indicate underdispersion relative to the Poisson distribution. The values of the dispersion parameter, ϕ, for the possible values of x_1 are provided in Table B.9.

2.6 Example

The following example demonstrates a regression model for the dispersion parameter in the double Poisson distribution. The data set, recorded between 1978 and 1985, consists of 126 observations of corporate takeover activity [24] [9] [34]. The dependent count variable is the number of takeover bids made on the company. There are nine explanatory variables in the dataset, including five indicator variables and four continuous variables. The five indicator variables indicate if there was a change in the ownership structure of the company proposed, whether the company launched a legal defense, if there was a proposed change in asset structure, if the company's management invited third party friendly bids, and whether the justice department intervened in the takeover process. The first of the continuous explanatory variables is the percentage of institutional holding. The second and third explanatory variables are the size of the company in billions of dollars and the size of the company squared, to account for non-linearity. Finally, the fourth continuous variable is the bid premium calculated by the bid price divided by the market price of stock for 14 business days prior to the takeover. This variable is mean-centered. Selected characteristics of the given variables are provided in Table B.10.

First, we estimated a model assuming constant dispersion. This yielded an estimated value of $\phi = 1.644$ with a 95% confidence interval $(1.321, 2.047)$. This indicates

that the number of bids is significantly underdispersed relative to the Poisson. Full results of this regression are provided in Figure B.1.

The second regression model was aimed at examining the relationship between the bid premium and the dispersion in the number of bids. The results of the regression are provided in Figure B.2. As the bid premium was mean-centered, we can use the constant value of 0.057 to estimate that the dispersion parameter of the number of bids at the mean value of the bid premium is 1.800. This indicates that the number of bids is underdispersed relative to the Poisson when the bid premium is at its mean. Further, based on the sign of the coefficient for the mean-centered bid premium, we can conclude that the dispersion parameter increases for bid premiums greater than the mean. For the minimum value of the mean-centered bid premium (-0.404), the dispersion parameter is estimated to be 0.542, indicating overdispersion. We can also determine that the number of bids is overdispersed for bid premiums lower than 0.224 below the mean and is underdispersed for any bid premiums greater than that.

2.7 Discussion

Based on the results of the simulation study and the case study, we can conclude that the regression model for the dispersion parameter proposed by Efron [19] performs well, particularly when using the k-th partial sum approximation of the normalizing constant. However, as utilizing this approximation is computationally expensive, the approaches assuming the normalizing constant equals one or the Efron approximation both provide adequate performance, especially in the case of overdispersion. Concerns arise when dealing with any underdispersion in the data, particularly with the assumption that the constant is one.

Beyond the performance of the model, one consideration this study aimed to address was the choice of the maximum allowable value for the dispersion parameter, M. Despite Efron's suggestion of choosing a value of M that is only slightly larger

than one, the results of our simulations indicate that utilizing a larger value will tend to lead to more stable estimates. An intuitive approach may be to use at least double the dispersion estimate from a model that treats the parameter as a constant. There seem to be few downsides to choosing an M that is much larger than necessary. When choosing a very large M, parameter estimates will be small in cases where data are overdispersed, but our simulations indicate that the estimated dispersion should remain relatively accurate.

Particular care should be taken with data that appear to be equidispersed relative to the Poisson when evaluated only through a model that assumes constant dispersion. If a consideration of the pattern of dispersion is of interest, such a result from a model assuming constant dispersion should not be taken to mean dispersion is constant. It may be the case that there is a change in direction of dispersion as the value of a covariate increases. In this case, we run the risk of incorrectly concluding that the data are Poisson distributed.

This model can be used in instances where the relationship between a covariate (or group of covariates) and the dispersion is of interest and/or when there is concern that the dispersion is not scalar. In the case that the covariates of interest are categorical, we can explore the possibility of a non-scalar structure by estimating the mean and variance of the data for each category and determining if their ratio remains constant across categories. In the case that the ratio is not constant, this would suggest that the dispersion parameter is not a scalar. Similarly, in the case of continuous covariates, we can create categories based on value groupings of the covariates and take the same approach.

Additionally, while there are other available approaches to modeling dispersion based on other distributions such as the generalized Poisson or negative binomial, the choice of which approach should be used comes down to the variance structure. The double Poisson approach will be most appropriate in data which has a variance

structure that closely resembles that of the model.

Chapter 3

Evaluating and Improving Identification of Tic Disorders in Children

3.1 Introduction

Tics are defined as "sudden, rapid, recurrent, nonrhythmic motor movement[s] or vocalization[s]" and are a defining feature of three types of recognized tic disorders: Tourette syndrome (TS), persistent motor or vocal tic disorder, and provisional tic disorder [3]. The prevalence of isolated tic behaviors in children, which may lead to a diagnosis of provisional tic disorder, has been estimated to be as high as 20% [35]. TS and persistent motor or vocal tic disorder, while less common, have estimated prevalence of 3 to 8 and 4 to 9 cases per 1000 school-aged children, respectively [35].

Two recently developed tools aimed at improved identification of potential or existing tic disorders in children are the Motor or Vocal Inventory of Tics (MOVeIT) provided in Table C.2, developed as a screening tool, and the Description of Tic Symptoms (DoTS) provided in Table C.1, a diagnostic measure. The MOVeIT questionnaire consists of 10 items with ordinal response categories of "never", "sometimes", and "often". The screening portion of the DoTS questionnaire includes 4 multi-part questions (totaling 10 items). Four of these items have "yes"/"no" responses and six have ordinal responses of "never", "sometimes", "often", and "always". Both tools can be completed by the child, their parent, or the child's teacher. Due to data availability, this paper focuses only on child and parent responses.

We aim to utilize logistic regression and random forest modeling to evaluate the

performance of these existing questionnaires as well as to identify individual items within the existing questionnaires that play an important role in the prediction of tic disorders in children. Utilizing both methods will allow us to compare their results and develop a more comprehensive picture of the items' contribution to accurate identification.

3.2 Data and Methods

3.2.1 Participants

Data were collected from 1,307 participants in five separate studies conducted at the University of South Florida (USF), the University of Florida College of Medicine (UF), and the University of Rochester Medical Center (URMC). Each study used at least one gold standard measure for identifying tic disorder status. These measures include the Yale Global Tic Severity Scale (YGTSS) [28], the Schedule for Affective Disorder and Schizophrenia for School-Age Children - Present and Lifetime tic disorder module (K-SADS) [26], and/or tic expert evaluation. Tic disorder status and a MOVeIT or DoTS questionnaire fully or partially completed by the child and/or parent was available for 1,205 children. Details on sample size, source population, and gold standards used in each study are available in Table C.3.

The first study, the Project to Learn about Youth Mental Health (PLAY-MH), aimed to better understand prevalence and treatment for child mental health disorders [15]. Among the data collected in this school-based multistage study was parent and child completed DoTS questionnaires. Children with potential tic disorders as indicated by the completed DoTS questionnaire were then invited to UF for assessment using the K-SADS.

The second study, conducted through USF, was focused on development of the MOVeIT tool and included both parent and child responses [29]. Participants were recruited at a specialty clinic, however investigators were unaware of tic disorder

status prior to recruitment. To determine whether a child had a tic disorder, the YGTSS was administered to all participants.

The third study was a case-control study to evaluate the sensitivity and specificity of both the DoTS and MOVeIT tools, including both parent and child responses, conducted by URMC [1]. This was done by comparing the questionnaire results to an evaluation conducted by a clinician with expertise in diagnosis and treatment of tic disorders. Participants were recruited at a tic disorder specialty clinic while controls were community-based.

The fourth study aimed to determine if the MOVeIT could be used to identify tics in children with stereotypy [39]. Thus, participants were recruited at a developmental and behavioral pediatrics clinic. It included only parent responses to the questionnaire and, similar to the third study, tic expert evaluation was used to determine the presence of a tic disorder.

The fifth and largest study recruited participants from a primary care pediatric clinic and aimed to evaluate the performance of various screening measures, including both the MOVeIT and DoTS in a pediatric care setting. A secondary aim was to determine if identification of tics could be used as a marker for symptoms of other conditions. Following child and/or parent completion of the screening tools, the YGTSS and K-SADS were administered as the gold standard. This study's data is yet unpublished.

The dataset created by combining the data collected in these five studies includes demographic information about the participants (age, sex, and race (non-Hispanic White, non-Hispanic Black, and other or multiple races and/or ethnicities)), child- and/or parent-reported responses to the MOVeIT and/or DoTS questionnaires (where available), and tic disorder status.

3.2.2 Logistic Regression

Logistic regression models the probability of an event occurring by utilizing the logit-link function giving the probability of the i-th child as,

$$p_i = \frac{\exp[\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}]}{1 + \exp[\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}]} \tag{3.1}$$

$$= \frac{1}{1 + \exp[-(\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij})]}, \tag{3.2}$$

thus the log-odds of that child having a tic disorder is given by

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}, \tag{3.3}$$

where β_0 is the intercept, m is the number of predictor variables, β_j are the derived coefficients, and x_{ij} are the covariates (demographic variables and/or responses to questionnaire items).

Often, a predicted probability above .5 will result in a positive prediction (i.e. that the child has a tic disorder) while any value below .5 will result in a negative prediction (i.e. that the child does not have a tic disorder). However, the cutoff value can be adjusted to make predictions more sensitive (lower cutoff values) or more specific (higher cutoff values).

To identify the individual items that have a significant relationship with tic disorder diagnosis, we utilize a Wald test of significance for each individual coefficient at a significance level of $\alpha = 0.10$. We chose a significance level higher than the commonly used $\alpha = 0.05$ because we are interested in being more inclusive as to which items should be considered for inclusion in future tools. The strength and direction of this relationship is determined by the size and sign of the coefficient, respectively. Coefficients are transformed to odds ratios (OR) for interpretability.

3.2.3 Random Forest

Building upon ideas of bootstrap aggregation (bagging) [5] and random feature selection [23] [2], random forest is an ensemble machine learning algorithm that trains multiple decision trees, the weak learners, to create a forest of trees, the strong learner [6]. By aggregating the predictions of multiple classification trees, random forests produce more accurate results than an individual tree [17] and avoid overfitting without the need for pruning [21].

The algorithm to build a random forest consisting of J classification trees is as follows.

Random Forest Algorithm

Suppose we have a training dataset $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_i = (x_{i,1}, ..., x_{i,p})^T$ are the predictors and y_i is the associated binary response.

1. Select a bootstrap sample, $\mathcal{D}_j$, of size n from $\mathcal{D}$.
2. Construct a classification tree using $\mathcal{D}_j$.
    a. Begin with all observations in $\mathcal{D}_j$ in a single root node.
    b. Select m predictors randomly from the p predictors.
    c. Identify the best binary split among the m selected predictors.
    d. Split the node into two subsequent nodes using the identified best split.
    e. Repeat steps b. through d. recursively until the stopping criterion is met.
3. Repeat the first two steps until J classification trees have been constructed.

The predicted classification for an observation classified by the random forest model

is most often determined by the majority vote, though this can be adjusted to allow for lower or higher cutoff values. Thus predictions are given by,

$$\widehat{f}(x) = arg\max_{y} \sum_{j=1}^{J} I(\widehat{h}_j(x) = y), \tag{3.4}$$

where $\widehat{h}_j(x)$ is the prediction at x using the jth tree [14].

The best split at each node is typically determined by a purity measure. If there are K classes and p_k is the proportion of observations in class k, one such purity measure, the Gini index, is given by

$$Q = \sum_{k \neq k'}^{K} p_k p_{k'}. \tag{3.5}$$

Large values of Q indicate poor classification. As each node is split into two descendant nodes, two values of Q will be calculated for each split, Q_L and Q_R. These can be multiplied by the number of observations in the descendant nodes and summed to give $Q_S = n_L Q_L + n_R Q_R$. Note that Q_S is calculated only for splits based on the m predictors randomly selected at the node.

As random forests utilize bagging, the observations not selected in the bootstrap sample for building an individual tree in the forest are referred to as out-of-bag (OOB). Each observation has a probability of being selected in each sample of $(1-\frac{1}{n})^n$ and this probability tends towards $\frac{1}{e}$ as n increases toward infinity. Thus, we can conclude that there will be a set of approximately one-third the number of observations in $\mathcal{D}$ that is not included in the jth sample, $\mathcal{D}_j$ and are, therefore, OOB.

This set of observations can be used to estimate the generalization error by determining the proportion of OOB observations that are incorrectly classified by the trees for which they are out-of-bag. That is, if the prediction for an OOB observation, x_i, is given by

$$\widehat{f}_{oob}(x_i) = arg\max_{y} \sum_{j \notin \mathcal{D}_j} I(\widehat{h}_j(x_i) = y), \tag{3.6}$$

then the generalization error rate can be estimated by the OOB error rate, given by

$$E_{oob} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \widehat{f}_{oob}(x_i)). \tag{3.7}$$

OOB data can also be used to determine variable importance. This is done by first determining predictions for the OOB observations and calculating the OOB error rate. Values of a particular predictor are then permuted while leaving all other predictor values fixed. The permuted OOB data are then used to compute new predictions and a new OOB error rate is calculated. The difference between the two error rates provides a measure of variable importance. Thus, if changing the values for a particular predictor greatly decreases the error rate, that predictor will be considered important.

Unlike many other methods utilizing decision tress, random forests are not sensitive to a large number of trees, N_{tree}, or the size of the trees. If the random forest is too small, however, the OOB error rate will be biased upward [8]. Because of this and the tendency of random forests to avoid overfitting, the chosen N_{tree} should be large but does not require tuning. One parameter that random forest may be sensitive to is the number of predictors randomly selected to determine the best split at each node, m [14]. We tune for this parameter using the area under the ROC curve.

3.2.4 Statistical Analysis

Six models were estimated based on three non-mutually exclusive subsets of data. These subsets were determined by completion of the MOVeIT or DoTS questionnaire by a parent, the child, or both a parent and the child. A random forest and a logistic regression model were estimated based on each of the three subsets. To estimate each model, 70% of the subset was randomly selected to be the training data set while the

remaining 30% was used as a validation set to assess the performance of the model. Statistical analysis was conducted in R [31] [27] [30] [37].

The predictive ability of the model was assessed using sensitivity, specificity, negative predictive value (NPV), and a weighted Youden's index.

$$\begin{aligned}
\text{sensitivity} &= \frac{TP}{TP+FN}, \\
\text{specificity} &= \frac{TN}{TN+FP}, \\
\text{NPV} &= \frac{TN}{TN+FN}, \\
\text{weighted Youden's} &= 3(\text{sensitivity}) + \text{specificity},
\end{aligned}$$

where TP, TN, FP, and FN are true positives, true negatives, false positives, and false negatives, respectively. The final cutoff values were determine using the weighted Youden's index. As we aim to correctly identify children that require further evaluation for tic disorders, avoiding false negatives is more important than incurring false positives. Thus, sensitivity and NPV are the most important indicators to consider. The weighted Youden's index was also chosen with this goal in mind. Weighting sensitivity more heavily than specificity placed more importance on limiting false negatives but still ensured that the number of false positives would be taken into consideration.

To identify the items in the questionnaires that should be considered in future tools, we considered the results of both the logistic regression and random forest models. As previously mentioned, a Wald test of significance at significance level $\alpha = 0.10$ was used to assess whether individual predictors have a significant relationship with the presence of tic disorders. Further, variable importance (as measured by the decrease in accuracy when OOB values of predictors are permuted) was calculated for each item in the random forest models.

3.2.5 Sensitivity Analysis

Of the five studies, three (the USF MOVeIT development study, the UR validation study, and the UR AUCD Project) recruited children from tic disorder or developmental and behavioral specialty clinics. Thus, sensitivity analyses were conducted to assess model performance when including only children recruited from general population settings. All logistic regression and random forest models were re-assessed excluding these three studies. Wald tests of significance were conducted and variable importance measures were calculated for each logistic regression and random forest model, respectively.

3.3 Results

3.3.1 Population Characteristics

Of the 1,307 children recruited across the six studies, 1,205 (92.20%) had known tic disorder status as determined by at least one gold standard as well as partial or complete responses from a parent and/or the child to either or both of the DoTS and MOVeIT questionnaires. The median age of the 1,205 included children was 9, with a range of ages from 2 to 20-years-old. Although there were 441 (36.51%) children under the age of 8, all self-reported responses to either questionnaire were provided by children 8-years-old and older. There were 705 (58.51%) male children and 500 (41.49%) female children. The racial and ethnic background of the children included 537 (44.56%) children identifying as non-Hispanic White, 335 (27.80%) identifying as non-Hispanic Black, and 325 (26.97%) identifying as other or multiple races and/or ethnicities. Eight (0.07%) children did not have a provided racial/ethnic identification. Detailed breakdown of demographic information is available in Table C.4.

3.3.2 Prediction of Tic Disorder

Prediction results for models estimated based on DoTS questionnaire responses are provided in Table C.5. As we are most interested in correctly identifying children with tic disorders, achieving high sensitivity and NPV is important. Logistic regression and random forest models based on DoTS responses all achieved high sensitivity and NPV (greater than or equal to 85% and 96%, respectively). They also maintained high specificity, with all three logistic models having specificity greater than 94% and the three random forest models having specificity greater than 83%.

Table C.6 provides prediction results for the logistic regression and random forest models estimated using responses to MOVeIT questionnaires. All logistic regression models based on MOVeIT data achieve sensitivity above 76% and NPV above 87% while maintaining specificity above 55%. The random forest models consistently outperform their corresponding LR models in terms of sensitivity and NPV. All random forest models achieved sensitivity and NPV greater than 84% and 94%, respectively. However, they under-performed logistic models in terms of specificity (minimum specificity achieved was 43.5%).

3.3.3 Logistic Regression Significant Items

Logistic regression results for the three DoTS models are summarized in Table C.7. The effects of parent responses to DoTS items 1a and 4a and child responses to DoTS item 4a were significant and positive in all logistic models in which the responses were included, indicating that increasingly affirmative responses were associated with greater odds of tic disorder. Other significant and positive effects include parent responses to DoTS item 2 and child responses to DoTS item 1c. Each of these were significant in only one model in which they were included.

DoTS responses with significant and negative effects in at least one model include parents responses to item 1e and child responses to item 1b, indicating that increas-

ingly affirmative responses to these items is associated with lower odds of tic disorder. Additionally, children who were neither non-Hispanic Black nor non-Hispanic White were indicated to have significantly lower odds of having a tic disorder diagnosis relative to non-Hispanic White children in all three DoTS logistic models. Non-Hispanic black children had significantly lower odds of tic disorder relative to non-Hispanic White children in the DoTS logisitc model including only child responses. Females also had significantly lower odds of tic disorder relative to males in this model.

A summary of logistic regression results for MOVeIT models are provided in Table C.8. Child responses to MOVeIT item 10 had significant positive effects in all logistic MOVeIT models in which they were included. Other MOVeIT item responses with significant positive effects in at least one model include parent responses to items 5 and 7 and child responses to items 1 and 9.

MOVeIT item responses with significant negative effects in at least one model include child responses to items 7 and 8. Additionally, the effect of sex was significant and negative in two of the MOVeIT logistic regression models, indicating females had lower odds of having a tic disorder than males. Non-Hispanic Both Black children and children who identified as neither non-Hispanic Black nor non-Hispanic White had significantly lower odds of having a tic disorder relative to non-Hispanic White children in two of the MOVeIT logistic models.

3.3.4 Variable Importance Results

Table C.9 includes variable importance measures from the DoTS random forest models. Parent responses to DoTS items 1a, 1c, 2, and 4a were among the five most important features in all models in which they were included. Child responses to DoTS item 4a was also among the most important features in all models including that variable. Features identified as among the most important in only one model include parent responses to DoTS item 4b and child responses to DoTS items 1c,

1d, and 4b. Race was also among the most important features in the DoTS random forest model including only child responses.

Features with negative variable importance were also identified in the DoTS random forest models, indicating that prediction accuracy increased when these values were permuted. One such feature was parent responses to item 1e, which had negative variable importance in all models in which it was included. Sex had negative variable importance in two of the DoTS random forest models in which it was included. Child responses to items 1b, 1f, and 3 had negative variable importance in the DoTS random forest model including only child responses. This suggests that removing these features may improve prediction of tic disorder status.

Variable importance results for the MOVeIT random forest models are in Table C.10. Features that were identified as among the most important in all models in which they were included were parent responses to MOVeIT items 1, 4, 5, and 6 as well as child responses to item 9. Child responses to items 1, 2, 4, and 6 were among the most important items in the MOVeIT random forest models including only child responses. Age was identified as among the most important features in the model including on parent responses.

Features with negative variable importance in at least one MOVeIT random forest model include sex and age. Sex had negative variable importance in the random forest model including only parent responses while age had negative variable importance in the model including both parent and child responses.

3.3.5 Sensitivity Analysis

Sensitivity analyses were conducted excluding data collected in studies recruiting participants from tic specialty or developmental and behavioral specialty clinics. Results from logistic regression and random forest models including DoTS questionnaire data are provided in Tables C.11 and C.12,respectively. Parent responses to DoTS item

4a were no longer significant in the logistic regression model including only parent responses. However, this item remained important in the corresponding random forest model. Child responses to item 4a was no longer significant in the logistic model including both parent and child responses. However, child responses to item 4b were significant in that model. Similarly, child responses to item 1b was no longer significant in the full model but was significant in the model including only child responses. There were no changes in the sign of variable importance for any of these items. The sign of variable importance changed for child responses to item 1a in the model including only child responses and parent responses to item 1d in all models in which it was included. The sign change indicates that permuting the OOB values of these responses actually improves the accuracy of the random forest models.

Logistic regression and random forest results from models including MOVeIT questionnaire data are provided in Tables C.13 and C.14,respectively. In the original random forest models, all features had positive variable importance in all random forest models. In the sensitivity analysis, parent responses to items 2 and 3 were no longer positive in the model including all responses and child responses to items 5, 6, 7, and 8 were no longer positive in the models including only child responses. Additionally, parent responses to item 5 was no longer significant in the logistic model including only parent responses while parent responses to item 6 became significant. Similarly, child responses to item 1 was no longer significant in the model including only child responses while child responses to item 2 became significant in this model. Child responses to items 7, 9, and 10 were also no longer significant in the model including all responses. Child responses to item 10 was also no longer significant in the logistic model including only child responses.

The role of demographic variables were also affected by the exclusion of this data. In the DoTS models, the sign of variable importance for race changed from positive to negative in all models. The coefficient associated with other/multiple races were also

no longer significant in the model including only parent only responses. The same was true for the coefficient associated with the non-Hispanic Black racial category in the model including only child responses. The coefficient for age became significant in the model including parent only responses.

In the MOVeIT models, the sign of the variable importance measure for race changed from positive to negative in all models and coefficients for racial categories were no longer significant. Additionally, the sign of the variable importance for sex changed from positive to negative in the model including only parent responses and the coefficient associated with sex was no longer significant in the model including all responses. The coefficient associated with age was also no longer significant in any models. Variable importance for this feature also changed signs in the random forest model including all responses and the model including only child responses.

3.4 Discussion

The random forest model is a popular alternative to logistic regression because the algorithm favors prediction over explanation [12] and this is reflected in the comparison of prediction results in our data. As expected, random forest tended to outperform logistic regression in the prediction of tic disorders, particularly in the MOVeIT models. All random forest models achieved sensitivity equal to or greater than their corresponding logistic regression models. While the accuracy of these models is determined by adjusting the cutoff values, the adjustment is on a finer scale for the random forest models due to the larger number of unique predicted probabilities produced by this type of model. By estimating across random subgroups, random forest models produce empirical distributions of probabilities that better reflect the variability in the data while the logistic regression models produce only one set of predicted probabilities. This results in random forest models being more effective at reducing Type II errors (false negatives).

Logistic regression, on the other hand, has the benefit of not only providing the strength of an item's association with tic disorder presence, but also the direction of the relationship. Variable importance calculated from random forest models can not provide information about the direction of the relationship without the assistance of partial dependence plots [12]. Logistic regression, however, suffers in this context because of the existence of multicollinearity between independent variables. Variable importance in random forest models may also be affected by dependence between features. These is evidence that the measure may be overinflated when there is correlation between features [4], however this is of less concern in this context because we are seeking to eliminate non-informative items rather than exclude correlated items.

Combining the results of the Wald tests of significance for the coefficients estimated in the logistic regression models and the calculated variable importance results from the random forest models, several items from the DoTS questionnaire were determined to provide useful information in predicting tic disorders. Parent responses to items 1a, 2, and 4a and child responses to items 1b and 4a were significant in at least one of the estimated models in which they were included and remained significant in a least one model in the sensitivity analysis. Parent responses to items 1a, 2, and 4a and child responses to item 4a also had high variable importance in random forest models although parent responses to item 2 and child responses to item 4a had lower variable importance in the sensitivity analysis.

There were conflicting results between the Wald tests of significance in the logistic regression models and the variable importance measure in the random forest models for some items. For example, parent responses to DoTS item 1e was significant in one logistic regression model but had negative variable importance in both random forest models in which it was included. Similarly, child responses to DoTS item 1b was significant in one logistic regression model while having negative variable impor-

tance in one random forest model. These items had negative variable importance in sensitivity analyses as well. Additionally, child responses to DoTS items 1f and 3 had negative variable importance in both the full and sensitivity analyses. Including results with negative variable importance may reduce the predictive ability of the tools and, thus, questions with negative importance should be considered for removal in future screening tools.

The performance of the MOVeIT models was less consistent. For example, parent responses to items 5 and 7 and child responses to items 1, 7, 9, and 10 were significant in at least one model in which they were included, however only parent responses to item 7 remained significant in at least one model in the sensitivity analysis. Similarly, all questionnaire item responses had positive variable importance in the original analysis but parent responses to items 2 and 3 and child responses to items 5, 6, 7, and 8 were negative in the sensitivity analysis.

The results of these analyses can be used to inform the development of future screening and diagnostic tools for identifying tic disordered in children. Utilizing the items identified as important in the logistic and/or random forest models in future tools may allow for a screening measure with higher sensitivity and has the potential to improve identification of tic disorders in a clinical setting and future studies.

Bibliography

[1] H.R. Adams et al. "Evaluation of new instruments for screening and diagnosis of tics and tic disorders inn a well characterized sample of children with tics and recruited controls." In: *Evidence-Based Practice in Child and Adolescent Mental Health* (in press).

[2] Y. Amit and D. Geman. "Shape quantization and recognition with randomized trees". In: *Natural Computation* 9 (1997), pp. 1545–1588.

[3] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*. 5th ed. Arlington, VA: American Psychiatric Association, 2013.

[4] C. Bernard, S. Da Veiga, and E. Scornet. "Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA". In: *Biometrika* 109.4 (2022), pp. 881–900.

[5] L. Breiman. "Bagging predictors". In: *Machine Learning* 24 (1996), pp. 123–140.

[6] L. Breiman. "Random forests". In: *Machine Learning* 45 (2001), pp. 5–32.

[7] N. Breslow. "Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models". In: *Journal of the American Statistical Association* 85.410 (1990), pp. 565–571.

[8] T. Bylander. "Estimating generalization error on two-class datasets using out-of-bag estimates". In: *Machine Learning* 48 (2002), pp. 287–297.

[9] A.C. Cameron and P.K. Trivedi. *Regression analysis of count data*. Second. Cambridge University Press, 2013.

[10] P.C. Consul and G.C. Jain. "A generalizaton of the Poisson distribution". In: *Technometrics* 15.4 (Nov. 1973), pp. 791–799.

[11] R.W. Conway and W.L. Maxwell. "A queuing model with state dependent service rates". In: *Journal of Industrial Engineering* 12 (1962), pp. 132–136.

[12] R. Couronne, P. Probst, and A. Boulesteix. “Random forest versus logistic regression: a large-scale benchmark experiment”. In: *BMC Bioinformatics* 19 (2018), p. 270.

[13] T.H. Cummings et al. “Modeling heaped count data”. In: *The Stata Journal* 15.2 (2015), pp. 457–479.

[14] A. Cutler, D. Cutler, and J. Stevens. “Machine Learning”. In: 2011. Chap. Random Forests.

[15] M.L. Danielson et al. “Community-based prevalence of externalizing and internalizing disorders among school-aged chilhdrena and adolescents in four geographically dispersed school districts in the United States”. In: *Child Psychiatry & Human Development* 52.3 (2021), pp. 500–514.

[16] C.B. Dean. “Testing for overdispersion in Poisson and binomial regression models”. In: *Journal of the American Statistical Association* 87.418 (June 1992), pp. 451–457.

[17] T.G. Dietterich. “Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science”. In: vol. 1857. Springer, Berlin, Hedielberg, 2000. Chap. Ensemble methods in machine learning.

[18] Centers for Disease Control and Prevention. URL: `http://www.cdc.gov/nchs/nhanes/nhanes2009-%202010/nhanes09%2010.htm`.

[19] B. Efron. “Double exponential families and their use in generalized linear regression”. In: *Journal of the American Statistical Association* 81.395 (Sept. 1986), pp. 709–721.

[20] E. García-Portugués. *Notes for Predictive Modeling*. Version 5.9.12. ISBN 978-84-09-29679-8. 2023. URL: `https://bookdown.org/egarpor/PM-UC3M/`.

[21] T. Hastie, R. Tibshirani, and J. Friedman. “The elements of statistical learning”. In: Second. Springer-Verlag, 2009. Chap. Random Forest, pp. 587–604.

[22] J.M. Hilbe. *Negative binomial regression*. 2nd ed. Cambridge University Press, 2011.

[23] T.K. Ho. “Random Decision Forests”. In: *Proceedings of the 3rd International Conference on Document Analysis and Recognition*. Montreal, QC, Aug. 1995, pp. 278–282.

[24] S. Jaggia and S. Thosar. "Multiple Bids as a Consequence of Target Management Resistance". In: *Review of Quantitative Finance and Accounting* (1993), pp. 447–457.

[25] Alboukadel Kassambara. *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.6.0. 2023. URL: https://CRAN.R-project.org/package=ggpubr.

[26] J. Kaufman et al. "Schedule for Affective Disorders and Schizophrenia for School-Age Children-Present and Lifetime Version (K-SADS-PL): initial reliability and validity data". In: *Journal of the American Academy of Child and Adolescent Psychiatry* 36.7 (1997), pp. 980–988.

[27] Max Kuhn. *caret: Classification and Regression Training*. R package version 6.0-93. 2022. URL: https://CRAN.R-project.org/package=caret.

[28] J.F. Leckman et al. "The Yale Global Tic Severity Scale: initial testing of a clinician-rated scale of tic severity". In: *Journal of the American Acadmey of Child and Adolescent Psychiatry* 288.4 (19898), pp. 566–573.

[29] A.B. Lewin et al. "Brief youth self-report screener for tics: Can a subscale of the Motor tic, Obsession and compulsion, and Vocal tic Evaluation Survey (MOVES) identify tic disorders in youth?" In: *Evidence-Based Practice in Child and Adolescent Mental Health* (under review).

[30] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[32] C.R Rao. "Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation". In: *Mathemtical Proceedings of the Cambridge Philosophical Society* 44.1 (1948), pp. 50–57.

[33] C.R. Rao. "Advances in ranking and selection, multiple comparisons, and reliability". In: Birkhauser, 2005. Chap. Score Test: Historicl Review and Recent Developments, pp. 3–15.

[34] A.J. Saez-Castillo and A. Conde-Sanchez. "A hyper-Poisson regression model for overdispersed and underdispersed count data". In: *Computational Statistics and Data Analysis* 61 (2013), pp. 148–157.

[35] L. Scahill, M. Specht, and C. Page. "The prevalence of tic disorders and clinical characteristics of children". In: *Journal of Obsessive Compulsive Related Disorders* 3.4 (Oct. 2014), pp. 394–400.

[36] G. Shmueli et al. "A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution". In: *Journal of the Royal Statistical Society Part C* 54 (2005), pp. 127–142.

[37] Julia Silge et al. *rsample: General Resampling Infrastructure*. R package version 1.1.0. 2022. URL: `https://CRAN.R-project.org/package=rsample`.

[38] *Stata Statistical Software: Release 17*. StataCorp. College Station, TX: StataCorp LP, 2021.

[39] J. Vermilion et al. "Performance of a tic screening tool (MOVeIT-14) for tics in comparison to expert clinician assessment in a developmental-behavioral pediatrics clinic sample". In: (in preparation).

[40] R.W.M. Wedderburn. "Quasi-likelihood functions, generalized linear models, and the Guass–Newton method". In: *Biometrika* 61.3 (Dec. 1974), pp. 439–447.

[41] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

[42] R. Winkelmann. *Econometric analysis of count data*. 5th ed. Springer Berlin, Heidelberg, 2008.

[43] Z. Yang et al. "Testing approaches for overdispersion in Poisson regression versus the generalized Poisson model". In: *Biometrical Journal* 49 (2007), pp. 565–584.

[44] F. Zhu. "Modeling time series of counts with COM-Poisson INGARCH Models". In: *Mathematical and Computer Modelling* 56.9 (2012), pp. 191–203.

[45] Y. Zou, S.R. Geedipally, and D. Lord. "Evaulating the double Poisson generalied linear model". In: *Accident Analysis and Prevention* 59 (2013), pp. 497–505.

Appendix A

Appendix - Derivation of a Score Test for Over- and Underdispersion Based on the Heaped Generalized Poisson Distribution

A.0.1 Derivation of the Score Test

Given the log-likelihood function of the heaped generalied Poisson model, the linear form first derivatives of the components of the regression model are given by the following equations,

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \eta_{\boldsymbol{\beta}}^{C}} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} I_{k_b} \frac{\partial P_{C_b}}{\partial \mu_b} \frac{\partial \mu_b}{\partial \eta_{\boldsymbol{\beta}}^{C}} = \sum_{i=1}^{n} g_{\boldsymbol{\beta} i}^{C} \\
\frac{\partial \mathcal{L}}{\partial \eta_{\gamma_u}^{M}} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} \frac{\partial P_{M_b}}{\partial \theta_u} \frac{\partial \theta_u}{\partial \eta_{\gamma_u}^{M}} P_{C_b} I_{k_b} = \sum_{i=1}^{n} g_{\gamma_u i}^{M} \\
\frac{\partial \mathcal{L}}{\partial \alpha} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} I_{k_b} \frac{\partial P_{C_b}}{\partial \alpha} = \sum_{i=1}^{n} g_{\alpha i}^{C}.
\end{aligned}$$

When evaluated under the null hypothesis, $\mathrm{H}_0 : \alpha = 0$, the score vector is then given by,

$$U(\eta_{\boldsymbol{\beta}}^{C}, \eta_{\gamma_u}^{M}, \alpha)' \Big|_{(\widehat{\eta_{\boldsymbol{\beta}}^{C}}, \widehat{\eta_{\gamma_u}^{M}}, 0)} = \left(\frac{\partial \mathcal{L}}{\partial \eta_{\boldsymbol{\beta}}^{C}}, \frac{\partial \mathcal{L}}{\partial \eta_{\gamma_u}^{M}}, \frac{\partial \mathcal{L}}{\partial \alpha} \right) \Big|_{(\widehat{\eta_{\boldsymbol{\beta}}^{C}}, \widehat{\eta_{\gamma_u}^{M}}, 0)} = (\mathbf{0}, \mathbf{0}, \widehat{U_\alpha}),$$

where

$$\widehat{U_\alpha} = \frac{\partial \mathcal{L}}{\partial \alpha} \Big|_{(\widehat{\eta_{\boldsymbol{\beta}}^{C}}, \widehat{\eta_{\gamma_u}^{M}}, 0)}.$$

The negative expected values of the second derivatives of the log-likelihood form the Fisher's Information Matrix. The seccond derivatives are given by,

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \eta_{\beta}^{C^2}} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} I_{k_b} \left[\frac{\partial^2 P_{C_b}}{\partial \mu_b^2} \left(\frac{\partial \mu}{\partial \eta_{\beta}^{C}} \right)^2 + \frac{\partial P_{C_b}}{\partial \mu_b} \frac{\partial^2 \mu_b}{\partial \eta_{\beta}^{C^2}} \right] - \sum_{i=1}^{n} \left(g_{\beta i}^{C} \right)^2 \\
\frac{\partial^2 \mathcal{L}}{\partial \eta_{\beta}^{C} \partial \eta_{\gamma_u}^{M}} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} \frac{\partial P_{M_b}}{\partial \theta_u} \frac{\partial \theta_u}{\partial \eta_{\gamma_u}^{M}} I_{k_b} \frac{\partial P_{C_b}}{\partial \mu_b} \frac{\partial \mu_b}{\partial \eta_{\beta}^{C}} - \sum_{i=1}^{n} \left(g_{\gamma_u i}^{M} \right) \left(g_{\beta i}^{C} \right) \\
\frac{\partial^2 \mathcal{L}}{\partial \eta_{\beta}^{C} \partial \alpha} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} I_{k_b} \frac{\partial^2 P_{C_b}}{\partial \mu_b \partial \alpha} \frac{\partial \mu_b}{\partial \eta_{\beta}^{C}} - \sum_{i=1}^{n} \left(g_{\beta i}^{C} \right) \left(g_{\alpha i}^{C} \right) \\
\frac{\partial^2 \mathcal{L}}{\partial \eta_{\gamma_u}^{M} \partial \eta_{\gamma_v}^{M}} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{C_b} I_{k_b} \frac{\partial^2 P_{M_b}}{\partial \theta_u \partial \theta_v} \frac{\partial \theta_u}{\partial \eta_{\gamma_u}^{M}} \frac{\partial \theta_v}{\partial \eta_{\gamma_v}^{M}} - \sum_{i=1}^{n} \left(g_{\gamma_u i}^{M} \right) \left(g_{\gamma_v i}^{M} \right) \\
\frac{\partial^2 \mathcal{L}}{\partial \eta_{\gamma_u}^{M} \partial \alpha} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} \frac{\partial P_{M_b}}{\partial \theta_u} \frac{\partial \theta_u}{\partial \eta_{\gamma_u}^{M}} I_{k_b} \frac{\partial P_{C_b}}{\partial \alpha} - \sum_{i=1}^{n} \left(g_{\gamma_u i}^{M} \right) \left(g_{\alpha i}^{C} \right) \\
\frac{\partial^2 \mathcal{L}}{\partial \alpha^2} &= \sum_{i=1}^{n} \frac{1}{L_i} \sum_{b=1}^{m+1} P_{M_b} I_{k_b} \frac{\partial^2 P_{C_b}}{\partial \alpha^2} - \sum_{i=1}^{n} \left(g_{\alpha i}^{C} \right)^2 .
\end{aligned}$$

Note that,

$$\frac{\partial P_{C_b}}{\partial \eta_{\beta}^{C}} = \frac{\partial P_{C_b}}{\partial \mu_b} \frac{\partial \mu_b}{\partial \eta_{\beta}^{C}}$$

and since $\mu_b = \exp(\eta_{\beta}^{C})$,

$$\frac{\partial \mu_b}{\partial \eta_{\beta}^{C}} = \mu_b.$$

As we are dealing with the heaped generalized Poisson distribution, P_{C_b} is given by,

$$P_{C_b} = \left\{ \mu_b(1-\delta) \left[\mu_b(1-\delta) + \delta y_b \right]^{y_b - 1} \right\} \exp\left\{ - \left[\mu_b(1-\delta) + \delta y_b \right] \right\} / y_b!$$

and

$$
\begin{aligned}
\frac{\partial P_{C_b}}{\partial \mu_b} &= P_{C_b}\left[\frac{1}{\mu_b} + \frac{(y_b-1)(1-\delta)}{\mu_b(1-\delta)+\delta y_b} - (1-\delta)\right] \\
\frac{\partial^2 P_{C_b}}{\partial^2 \mu_b} &= P_{C_b}\left[\frac{1}{\mu_b} + \frac{(y_b-1)(1-\delta)}{\mu_b(1-\delta)+\delta y_b} - (1-\delta)\right]^2 \\
&\quad -P_{C_b}\left\{-\frac{1}{\mu_b^2} + \frac{(y_b-1)(1-\delta)^2}{[\mu_b(1-\delta)+\delta y_b]^2}\right\} \\
\frac{\partial P_{C_b}}{\partial \delta} &= P_{C_b}\left[\frac{-1}{1-\delta} + (\mu_b - y_b) + \frac{(y_b-1)(y_b-\mu_b)}{\mu_b(1-\delta)+\delta y_b}\right] \\
\frac{\partial^2 P_{C_b}}{\partial \delta^2} &= P_{C_b}\left[\frac{-1}{1-\delta} + (\mu_b - y_b) + \frac{(y_b-1)(y_b-\mu_b)}{\mu_b(1-\delta)+\delta y_b}\right]^2 \\
&\quad -P_{C_b}\left\{\frac{1}{(1-\delta)^2} + \frac{(y_b-1)(y_b-\mu_b)^2}{[\mu_b(1-\delta)+\delta y_b]^2}\right\}
\end{aligned}
$$

Further, as we assume m heaping values other than 1, the multinomial probabilities are given by,

$$
P_{M_1}(B=1|\mathbf{Z},\boldsymbol{\gamma}_b) = \frac{1}{1+\exp(\mathbf{Z}\boldsymbol{\gamma}_2)+...+\exp(\mathbf{Z}\boldsymbol{\gamma}_{m+1})}, \text{and}
$$

and

$$
P_{M_b}(B=b|\mathbf{Z},\boldsymbol{\gamma}_b) = \frac{\exp(\mathbf{Z}\boldsymbol{\gamma}_b)}{1+\exp(\mathbf{Z}\boldsymbol{\gamma}_2)+...+\exp(\mathbf{Z}\boldsymbol{\gamma}_{m+1})}, b=2,...,m+1.
$$

where we define $p_j = z\gamma_j$ and $\theta_j = \exp(p_j)$. Then by $b = 1, ..., m+1$, we have,

$$
\frac{\partial P_{M_b}}{\partial \theta_u} = \frac{1}{1+\sum_{i=1}^m \theta_i} I(b=u) + \frac{-\theta_b}{(1+\sum_{i=1}^m \theta_i)^2},
$$

and we note the following,

$$
\frac{\partial P_{M_b}}{\partial p_u} = \frac{\partial P_{M_b}}{\partial \theta_u}\frac{\partial \theta_u}{\partial p_u} = \frac{\partial P_{M_b}}{\partial \theta_u}\theta_u
$$

$$\frac{\partial^2 P_{M_b}}{\partial\theta_u\partial\theta_v} = \frac{-1}{\left(1+\sum_{i=1}^m \theta_i\right)^2} I(b=u) + \frac{2\theta_b\left(1+\sum_{i=1}^m \theta_i\right)}{\left(1+\sum_{i=1}^m \theta_i\right)^3} + \frac{-1}{\left(1+\sum_{i=1}^m \theta_i\right)^2} I(b=v).$$

Based on the above and the structure of $U(\eta_{\boldsymbol{\beta}}^C, \eta_{\boldsymbol{\gamma}_u}^M, \alpha)' \mid_{(\widehat{\eta_{\boldsymbol{\beta}}^C}, \widehat{\eta_{\boldsymbol{\gamma}_u}^M}, 0)}$, we obtain the score statistic as in (1.10).

Table A.1 Power (%) of the Wald, likelihood ratio (SSR-LRT), and score tests for heaped generalized Poisson distributed simulated data at different values of the dispersion parameter, α.

n	Method	Power (%)										
	α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.10
	Wald	6.901	6.193	5.760	5.586	5.887	6.590	7.858	9.362	11.825	14.453	17.652
100	SSR-LRT	6.305	6.102	6.003	6.233	7.180	8.652	10.417	12.874	15.608	18.779	22.687
	Score	7.265	6.587	6.094	5.839	6.160	6.954	8.161	9.747	12.240	14.928	18.188
	Wald	5.830	5.301	5.701	7.011	9.891	13.841	18.700	24.580	31.380	38.590	46.160
250	SSR-LRT	5.410	5.411	6.311	8.561	12.041	16.542	22.110	28.180	35.610	42.930	50.940
	Score	5.890	5.441	5.801	7.101	9.981	13.921	18.920	24.790	31.550	38.820	46.440
	Wald	5.701	5.651	7.760	10.701	16.832	25.753	36.544	47.700	58.880	69.517	78.440
500	SSR-LRT	5.631	5.911	8.330	12.201	19.112	28.673	39.694	51.000	62.230	72.367	80.910
	Score	5.731	5.661	7.770	10.721	16.862	25.843	36.584	47.780	58.970	69.527	78.500

n	Method	Power (%)									
	α	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60
	Wald	37.898	62.132	79.972	91.096	96.369	98.816	99.627	99.854	99.905	100.000
100	SSR-LRT	45.050	68.046	84.587	93.626	97.699	99.259	99.762	99.906	99.996	100.000
	Score	38.831	62.792	80.754	91.535	96.604	98.898	99.679	99.864	99.994	100.000
	Wald	81.428	96.559	99.640	99.970	100.000					
250	SSR-LRT	84.458	97.269	99.750	99.980	100.000					
	Score	81.658	96.599	99.640	99.970	100.000					
	Wald	98.300	99.970	100.000							
500	SSR-LRT	98.620	99.980	100.000							
	Score	98.310	99.970	100.000							

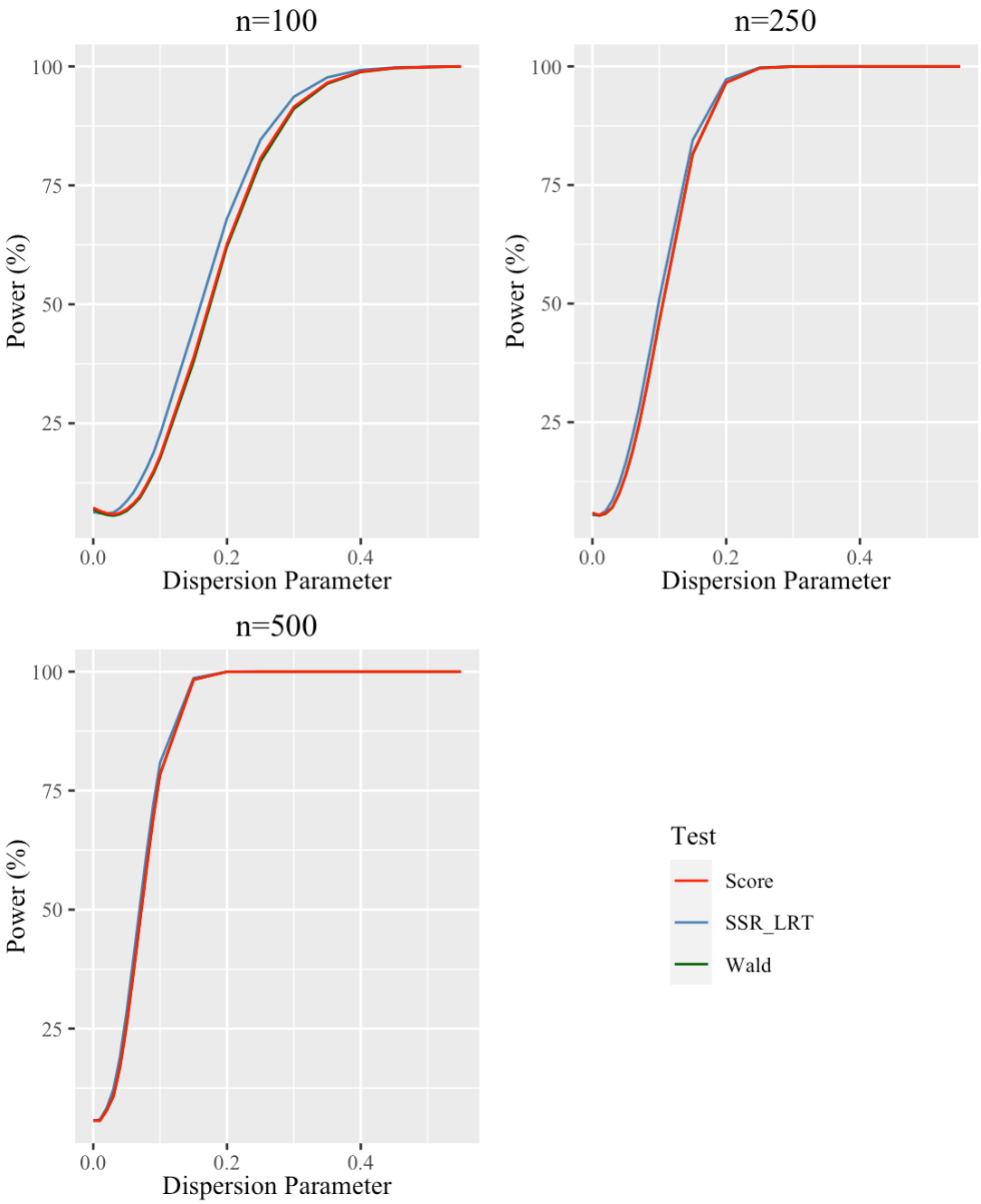

Figure A.1 Plots of the power (%) of the Wald, likelihood ratio (SSR-LRT), and score tests for heaped generalized Poisson distributed simulated data at different values of the dispersion parameter, α. Plots were constructed using R [31] [41] [25].

Table A.2 Power (%) of the Wald, likelihood ratio (SSR-LRT), and score tests for heaped negative binomial distributed simulated data at different values of the probability of success on a single trial, p.

n	Method	Power (%)								
	p	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91
	Wald	6.594	6.193	5.863	5.750	5.605	5.515	5.604	5.876	6.491
100	SSR-LRT	6.271	6.102	5.954	6.003	6.009	6.232	6.674	7.239	8.058
	Score	6.927	6.587	6.307	6.084	5.919	5.767	5.826	6.118	6.875
	Wald	5.530	5.310	5.460	5.671	6.190	7.140	8.480	10.180	11.790
250	SSR-LRT	5.290	5.420	5.820	6.361	7.330	8.650	10.290	12.330	14.130
	Score	5.620	5.450	5.570	5.771	6.260	7.210	8.600	10.260	11.850
	Wald	5.470	5.660	6.320	7.770	9.290	11.390	14.690	18.340	23.270
500	SSR-LRT	5.670	5.900	6.980	8.420	10.250	13.070	16.600	20.910	25.820
	Score	5.520	5.670	6.360	7.770	9.300	11.410	14.740	18.400	23.310

n	Method	Power (%)							
	p	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
	Wald	6.712	19.405	44.764	73.283	91.653	98.439	99.813	100.000
100	SSR-LRT	8.765	24.694	52.314	79.001	94.051	99.168	99.875	100.000
	Score	7.097	20.073	45.778	74.239	92.010	98.603	99.834	100.000
	Wald	14.340	50.980	88.199	99.090	99.990	100.000		
250	SSR-LRT	17.160	55.580	90.379	99.350	100.000	100.000		
	Score	14.390	51.360	88.309	99.100	99.990	100.000		
	Wald	27.460	83.510	99.480	100.000				
500	SSR-LRT	30.430	85.210	99.600	100.000				
	Score	27.480	83.570	99.510	100.000				

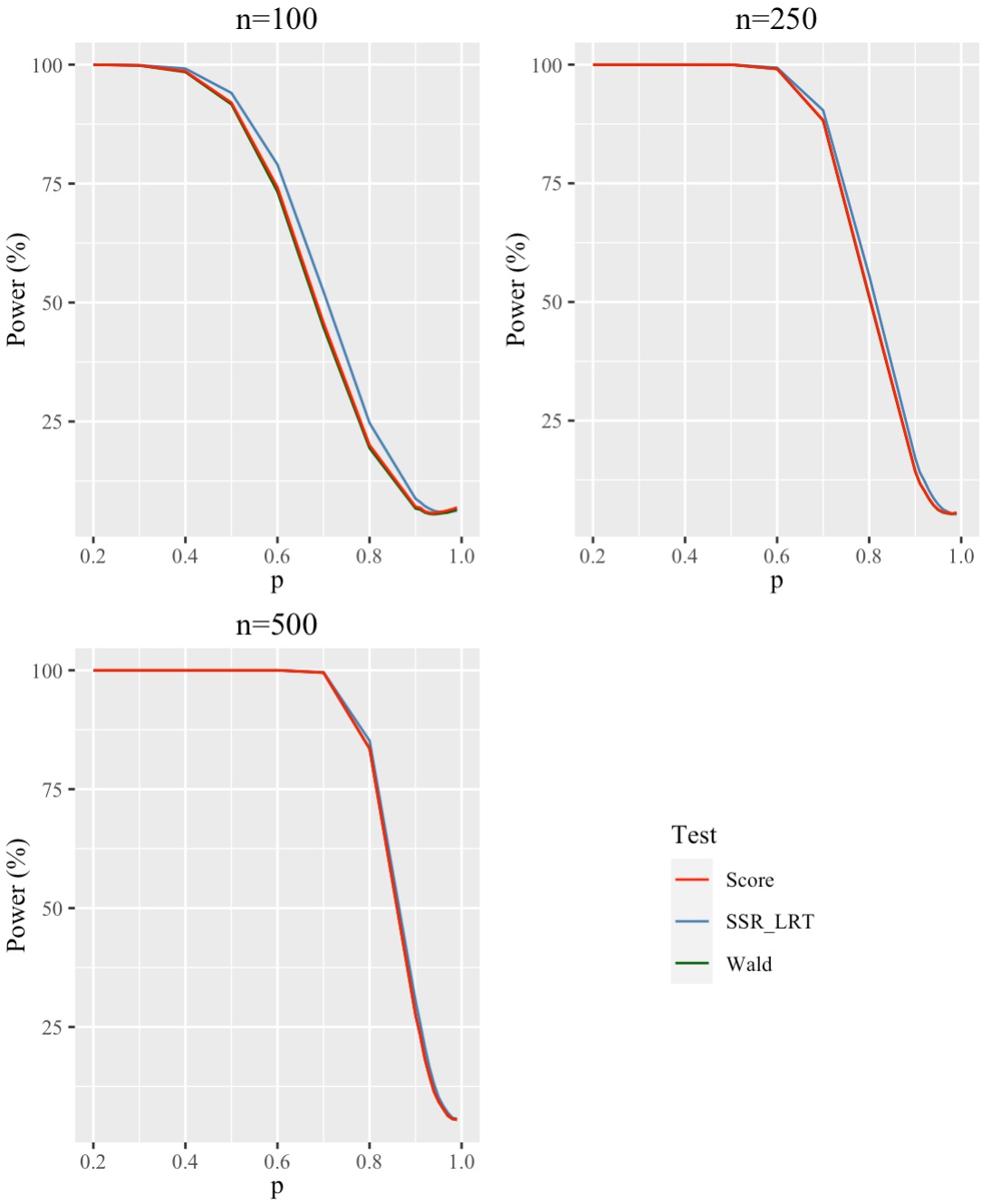

Figure A.2 Plots of the power (%) of the Wald, likelihood ratio (SSR-LRT), and score tests for heaped negative binomial distributed simulated data at different values of the probability of success on a single trial, p. Plots were constructed using R [31] [41] [25].

Table A.3 Selected characteristics of the variables included in the models estimated in the case study ($n = 1,504$).

		Variable Name	Mean (Std. Dev.)	Range (Min., Max.)
Cigarettes smoked per day in the past 30 days		**smd650**	11.549 (9.982)	(1, 95)
Age		**ridageyr**	40.735 (16.644)	(13, 80)
			Freq.	%
Gender		**gendernew**		
	Females		669	44.481
	Males		835	55.519
Race		**racenew**		
	Non-Hispanic white		749	49.801
	Other races		755	50.199

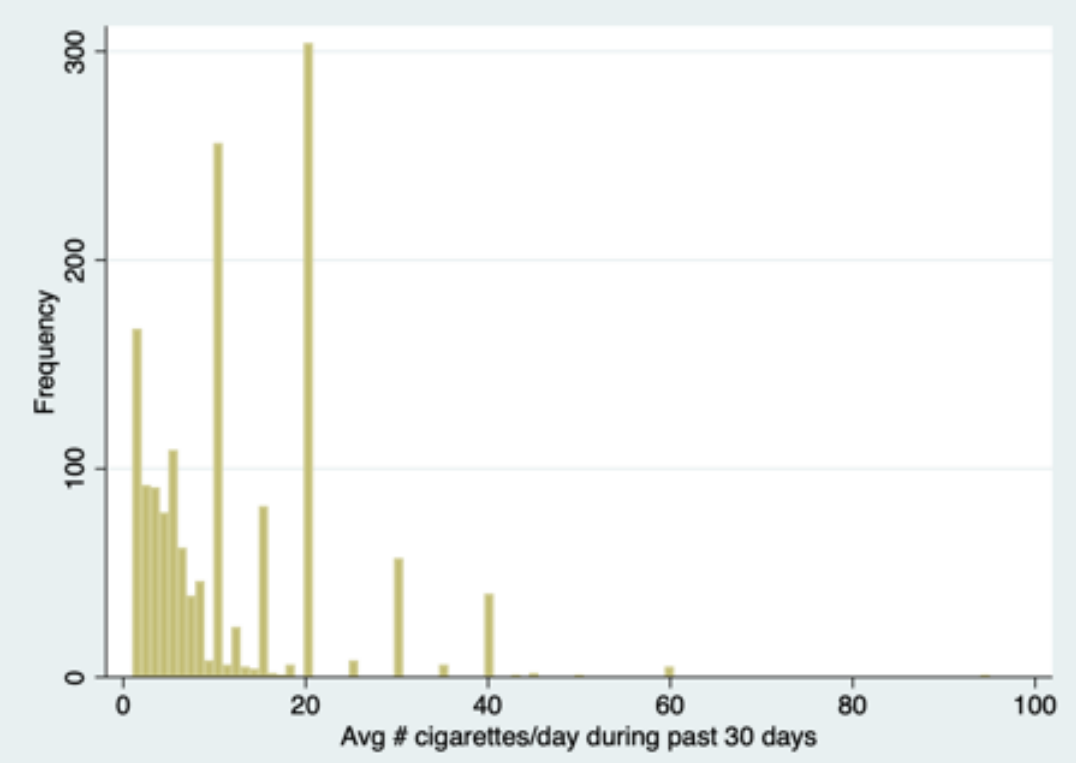

Figure A.3 Histogram of the number of cigarettes participants smoked per day for the past 30 days.

Heaped Poisson regression — Number of obs = 1,504
LR chi2(3) = 799.97
Prob > chi2 = 0.0000
Log likelihood = -5241.3788 — Pseudo R2 = 0.0709

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
smd650						
gendernew	-.117404	.0270364	-4.34	0.000	-.1703944	-.0644136
racenew	.6523586	.0269814	24.18	0.000	.5994759	.7052412
ridageyr	.0115673	.0007373	15.69	0.000	.0101222	.0130124
_cons	1.240218	.0390411	31.77	0.000	1.163699	1.316737
modulo_5						
gendernew	-.1070762	.1254271	-0.85	0.393	-.3529088	.1387563
racenew	1.305365	.126273	10.34	0.000	1.057874	1.552855
ridageyr	.0271538	.003793	7.16	0.000	.0197197	.0345878
_cons	-1.618054	.191724	-8.44	0.000	-1.993826	-1.242281

Figure A.4 Results of the estimated heaped Poisson regression model for number of cigarettes smoked in the past 30 days.

Heaped Gen. Poisson regression

Number of obs = 1,504
LR chi2(3) = 292.90
Prob > chi2 = 0.0000
Pseudo R2 = 0.0301

Log likelihood = -4722.9238

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
smd650						
gendernew	-.0621945	.0390535	-1.59	0.111	-.1387378	.0143489
racenew	.576981	.0390636	14.77	0.000	.5004177	.6535444
ridageyr	.0107667	.0010346	10.41	0.000	.0087389	.0127945
_cons	1.369426	.0568109	24.11	0.000	1.258079	1.480774
modulo_5						
gendernew	-.1823013	.1490652	-1.22	0.221	-.4744638	.1098611
racenew	1.5208	.154008	9.87	0.000	1.21895	1.82265
ridageyr	.0300395	.0043496	6.91	0.000	.0215143	.0385646
_cons	-2.129924	.2373677	-8.97	0.000	-2.595156	-1.664692
/atanhdelta	.4463878	.0190762	23.40	0.000	.4089991	.4837765
delta	.4189253	.0157284			.3876225	.449263

Figure A.5 Results of the estimated heaped generalized Poisson regression model for number of cigarettes smoked in the past 30 days.

Table A.4 Results of the Wald, likelihood ratio (SSR-LRT), and derived score test for $H_0 : \alpha = 0$ and $H_1 : \alpha \neq 0$.

Test	χ^2 value	p-value
Wald	547.570	< 0.001
SSR_LRT	1036.910	< 0.001
Score	649.311	< 0.001

Appendix B

Appendix - Modelling the Dispersion Parameter in the Double Poisson Distribution

Table B.1 Structures of the dispersion parameter used to simulate data.

Structure of ϕ	Value of x_1	ϕ	Description
$\frac{2}{1+\exp(-(-1+0.5x_1))}$	0	0.538	Dispersion depends on x_1 and data are overdispersed for all values of x_1.
	1	0.755	
$\frac{2}{1+\exp(-(1+0.5x_1))}$	0	1.462	Dispersion depends on x_1 and data are underdispersed for all values of x_1.
	1	1.635	
$\frac{2}{1+\exp(-(-1+2x_1))}$	0	0.538	Data are overdispersed when $x_1 = 0$ and underdispersed when $x_1 = 1$.
	1	1.462	
0.75	0	0.75	Dispersion is constant.
	1	0.75	

Table B.2 Mean values and 95% confidence intervals for the parameters of dispersion models estimated based on data simulated to be dependent on x_1 and consistently overdispersed. The dispersion structure used to simulate data was $\frac{2}{1+\exp(-(-1+0.5x_1))}$.

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sc1	1.25	t_0	-0.288 (-0.558, -0.018)	-0.354 (-0.448, -0.261)	-0.376 (-0.421, -0.331)
		t_1	1.192 (-1.714E+06, 1.714E+06)	1.04 (0.636, 1.444)	1.031 (0.845, 1.216)
	2	t_0	-1.016 (-1.168, -0.865)	-1.061 (-1.119, -1.003)	-1.076 (-1.36, -0.792)
		t_1	0.711 (0.305, 1.116)	0.712 (0.561, 0.863)	0.714 (0.64, 0.788)
	2.5	t_0	-1.311 (-1.442, -1.181)	-1.352 (-1.402, -1.302)	-1.366 (-1.391, -1.341)
		t_1	0.642 (0.322, 0.961)	0.646 (0.525, 0.768)	0.649 (0.589, 0.709)
		ϕ	0.645 (0.621, 0.671)	0.631 (0.621, 0.641)	0.627 (0.622, 0.631)
sca	1.25	t_0	0.054 (-9.591E+06, 9.591E+06)	0.003 (-0.051, 0.056)	-0.004 (-0.029, 0.021)
		t_1	13.76 (-1.215E+07, 1.215E+07)	7.310 (-3.158E+06, 3.158E+06)	5.413 (-7.871E+05, 7.871E+05)
	2	t_0	-0.757 (-0.825, -0.689)	-0.79 (-0.817, -0.763)	-0.802 (-0.816, -0.787)
		t_1	1.756 (1.555, 1.956)	1.782 (1.7, 1.863)	1.762 (1.719, 1.805)
	2.5	t_0	-1.082 (-1.141, -1.024)	-1.113 (-1.137, -1.089)	-1.122 (-1.135, -1.11)
		t_1	1.406 (1.267, 1.544)	1.412 (1.355, 1.468)	1.139 (1.11, 1.169)
		ϕ	0.889 (0.868, 0.911)	0.858 (0.849, 0.866)	0.850 (0.846, 0.854)

continued on next page

continued from previous page

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sce	1.25	t_0	-0.166 (-0.546, 0.215)	-0.238 (-0.357, -0.119)	-0.263 (-0.32, -0.205)
		t_1	3.058 (-8.367E+06, 8.367E+06)	2.81 (-3.496E+05, 3.496E+05)	0.761 (-1.733E+02, 1.748E+02)
	2	t_0	-0.929 (-1.112, -0.746)	-0.973 (-1.042, -0.903)	-0.989 (-1.023, -0.955)
		t_1	0.588 (-1.687E+05, 1.687E+05)	0.520 (0.18, 0.859)	0.513 (0.351, 0.675)
	2.5	t_0	-1.231 (-1.387, -1.075)	-1.27 (-1.33, -1.21)	-1.284 (-1.314, -1.255)
		t_1	0.478 (-1.017E+02, 1.026E+02)	0.467 (0.197, 0.737)	0.465 (0.335, 0.595)
	ϕ		0.603 (0.564, 0.646)	0.588 (0.447, 0.772)	0.582 (0.574, 0.59)

Table B.3 Estimated values of the dispersion parameter, ϕ, at different values of x_1 calculated using the mean values of the parameters for dispersion models estimated based on data simulated to be dependent on x_1 and consistently overdispersed. The dispersion structure used to simulate data was $\frac{2}{1+\exp(-(-1+0.5x_1))}$.

		sample size					
		n=100		n=250		n=500	
		x_1=0	x_1=1	x_1=0	x_1=1	x_1=0	x_1=1
sc1	M=1.25	0.536	0.890	0.515	0.831	0.509	0.823
	M=2	0.532	0.848	0.514	0.827	0.508	0.821
	M=2.5	0.531	0.847	0.514	0.826	0.508	0.820
sca	M=1.25	0.642	1.250	0.626	1.249	0.624	1.244
	M=2	0.639	1.462	0.624	1.459	0.619	1.446
	M=2.5	0.633	1.450	0.618	1.435	0.614	1.261
sce	M=1.25	0.573	1.184	0.551	1.161	0.543	0.778
	M=2	0.566	0.831	0.549	0.777	0.542	0.766
	M=2.5	0.565	0.800	0.548	0.773	0.542	0.765

Table B.4 Mean values and 95% confidence intervals for the parameters of the dispersion model estimated based on data simulated to be dependent on x_1 and consistently underdispersed. The dispersion structure used to simulate data was $\frac{2}{1+\exp(-(1+0.5x_1))}$.

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sc1	1.25	t_0	19.274 (-2.470E+14, 2.470E+14)	18.519 (-6.566E+12, 6.566E+12)	19.069 (-1.593E+12, 1.593E+12)
		t_1	4.54E+10 (-2.489E+14, 2.490E+14)	4.84E+07 (-6.566E+12, 6.566E+12)	-10.450 (-1.627E+12, 1.627E+12)
	2	t_0	2.560 (-3.856+E06, 3.856+E06)	1.160 (-4.428+04, 4.428E+04)	0.936 (-2.709E+02, 2.728E+02)
		t_1	-2.137 (-3.857E+06, 3.857E+06)	-0.810 (-4.428E+04, 4.428E+04)	-0.606 (-2.725E+02, 2.713E+02)
	2.5	t_0	0.643) (-2.888E+05, 2.888E+05	0.334 (0.117, 0.551)	0.289 (0.199, 0.378)
		t_1	-0.730 (-2.888E+05, 2.888E+05)	-0.461 (-0.792, -0.13)	-0.427 (-0.572, -0.281)
	ϕ		1.311 (1.26, 1.363)	1.285 (1.265, 1.305)	1.275 (1.265, 1.285)
sca	1.25	t_0	17.949 (-1.895E+14, 1.895E+14)	17.977 (-1.529E+15, 1.529E+15)	19.535 (-1.096E+14, 1.096E+14)
		t_1	-5.111 (-1.895E+14, 1.895E+14)	-5.694 (-1.727E+15, 1.727E+15)	1.990E+04 (-1.096E+14, 1.096E+14)
	2	t_0	2.198 (-1.047E+06, 1.047E07)	1.255 (-4.931E+03, 4.934E+03) (-4.931E+03, 4.934E+03)	1.068 (0.933, 1.203)
		t_1	2.094E+05 (-1.368E06, 1.787E+06)	0.363 (-4.933E+03, 4.933E+03)	0.557 (0.365, 0.749)

continued on next page

continued from previous page

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sca	2.5	t_0	0.675 (-2.506E+04, 2.506E+04)	0.420 (0.310, 0.529)	0.346 (0.304, 0.388)
		t_1	0.125 (-2.506E+04, 2.506E+04)	0.376 (0.221, 0.531)	0.441 (0.377, 0.505)
		ϕ	1.617 (1.592, 1.642)	1.568 (1.558, 1.579)	1.549 (1.543, 1.554)
sce	1.25	t_0	20.779 (-4.312E+13, 4.312E+13)	22.388 (-4.822E+13, 4.822E+13)	24.616 (-3.332E+13, 3.332E+13)
		t_1	49.249 (-4.312E+13, 42312E+13)	-1.554 (-4.822E+13, 4.822E+13)	-2.473 (-3.332E+13, 3.332E+13)
	2	t_0	2.861 (-1.178E+10, 1.178E+10)	1.382 (-1.464E+05, 1.464E+05)	1.112 (-5.227E+02, 5.249E+02)
		t_1	4.04 (-1.186E+10, 1.186E+10)	4.216 (-5.676E+06, 5.676E+06)	1.055 (-1.064E+06, 1.064E+06)
	2.5	t_0	0.768 (-3.219E+04, 3.219E+04)	0.432 (0.045, 0.818)	0.380 (0.293, 0.467)
		t_1	1.654E+06 (-2.773E+06, 6.081E+06)	0.363 (-1.636E+04, 1.636E+04)	0.312 (0.005, 0.619)
		ϕ	1.586 (1.52, 1.655)	1.549 (1.523, 1.576)	1.535 (1.522, 1.548)

Table B.5 Estimated values of the dispersion parameter, ϕ, at different values of x_1 calculated using the mean values of the parameters for dispersion models estimated based on data simulated to be dependent on x_1 and consistently underdispersed. The dispersion structure used to simulate data was $\frac{2}{1+\exp(-(1+0.5x_1))}$.

		sample size					
		n=100		n=250		n=500	
		x_1=0	x_1=1	x_1=0	x_1=1	x_1=0	x_1=1
sc1	M=1.25	1.250	1.250	1.250	1.250	1.250	1.250
	M=2	1.856	1.208	1.523	1.173	1.437	1.163
	M=2.5	1.639	1.196	1.457	1.171	1.429	1.164
sca	M=1.25	1.250	1.250	1.250	1.250	1.250	1.250
	M=2	1.800	2.000	1.556	1.669	1.488	1.671
	M=2.5	1.656	1.725	1.508	1.722	1.464	1.718
sce	M=1.25	1.250	1.250	1.250	1.250	1.250	1.250
	M=2	1.892	1.998	1.599	1.993	1.505	1.795
	M=2.5	1.708	2.500	1.516	1.722	1.485	1.666

Table B.6 Mean values and 95% confidence intervals for the parameters of the dispersion models estimated based on data simulated to be overdispersed when $x_1 = 0$ and underdispersed when $x_1 = 1$. The dispersion structure used to simulate data was $\frac{2}{1+\exp(-(-1+2x_1))}$.

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sc1	1.25	t_0	-0.289 (-0.558, -0.019)	-0.354 (-0.448, -0.261)	-0.376 (-0.421, -0.331)
		t_1	9.47E+10 (9.463E+10, 9.477E+10)	21.364 (-5.292E+06, 5.292E+06)	2.615 (-2.894E+05, 2.894E+05)
	2	t_0	-1.016 (-1.168, -0.865)	-1.061 (-1.119, -1.004)	-1.076 (-1.105, -1.048)
		t_1	1.298 (-2.116E+02, 2.142E+02)	1.281 (1.059, 1.503)	1.278 (1.17, 1.385)
	2.5	t_0	-1.131 (-1.262, -1.001)	-1.352 (-1.402, -1.302)	-1.366 (-1.391, -1.341)
		t_1	1.125 (0.712, 1.538)	1.124 (0.97, 1.278)	1.124 (1.049, 1.199)
		ϕ	0.716 (0.688, 0.744)	0.700 (0.689, 0.711)	0.694 (0.689, 0.7)
sca	1.25	t_0	0.022 (-2.165E+05, 2.165E+05)	-0.043 (-0.09, 0.005)	-0.070 (-0.094, -0.045)
		t_1	2.413E+06 (-1.758E+07, 2.24E+07)	12.16 (-1.522E+07, 1.522E+07)	12.418 (-1.435E+06, 1.435E+06)
	2	t_0	-0.760 (-0.805, -0.714)	-0.810 (-0.828, -0.792)	-0.827 (-0.836, -0.817)
		t_1	6.520E+07 (6.520E+07, 6.520E+07)	2.438 (2.331, 2.545)	2.377 (2.326, 2.428)
	2.5	t_0	-1.082 (-1.121, -1.042)	-1.126 (-1.142, -1.11)	-1.142 (-1.15, -1.134)
		t_1	1.904 (1.765, 2.042)	1.891 (1.839, 1.943)	1.187 (0.935, 1.439)
		ϕ	1.004 (0.982, 1.027)	0.961 (0.952, 0.97)	0.949 (0.944, 0.953)

continued on next page

continued from previous page

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sce	1.25	t_0	-0.167 (-1.905E+03, 1.904E+03)	-0.239 (-0.359, -0.12)	-0.266 (-0.324, -0.209)
		t_1	6.460E+07 (8.936E+06, 1.203E+08)	4.330E+08 (4.112E+08, 4.548E+08)	7.555E+04 (-8.523E+06, 8.674E+06)
	2	t_0	-0.933 (-1.116, -0.75)	-0.975 (-1.044, -0.905)	-0.990 (-1.024, -0.955)
		t_1	7.480E+07 (4.834E+07, 1.013E+08)	3.252 (-1.940E+06, 1.940E+06)	2.285 (-6.633E+05, 6.633E+05)
	2.5	t_0	-1.236 (-1.396, -1.076)	-1.272 (-1.332, -1.212)	-1.285 (-1.582, -0.988)
		t_1	2.125 (-2.436E+06, 2.436E+06)	1.716 (-2.419E+04, 2.419E+04)	1.669 (1.483, 1.855)
		ϕ	0.693 (0.652, 0.738)	0.676 (0.66, 0.693)	0.669 (0.661, 0.677)

Table B.7 Estimated values of the dispersion parameter, ϕ, at different values of x_1 calculated using the mean values of the parameters for dispersion modelsestimated based on data simulated to be overdispersed when $x_1 = 0$ and underdispersed when $x_1 = 1$. The dispersion structure used to simulate data was $\frac{2}{1+\exp(-(-1+2x_1))}$.

		sample size					
		n=100		n=250		n=500	
		x_1=0	x_1=1	x_1=0	x_1=1	x_1=0	x_1=1
sc1	M=1.25	0.535	1.250	0.515	1.250	0.509	1.130
	M=2	0.531	1.140	0.514	1.109	0.508	1.100
	M=2.5	0.610	1.246	0.514	1.108	0.508	1.100
sca	M=1.25	0.632	1.250	0.612	1.250	0.603	1.250
	M=2	0.637	2.000	0.616	1.672	0.609	1.650
	M=2.5	0.633	1.737	0.612	1.706	0.605	1.278
sce	M=1.25	0.573	1.250	0.551	1.250	0.542	1.250
	M=2	0.565	2.000	0.548	1.814	0.542	1.570
	M=2.5	0.563	1.772	0.547	1.523	0.542	1.487

Table B.8 Mean values and 95% confidence intervals for the parameters of the dispersion models estimated based on data simulated to have a scalar dispersion structure, $\phi = 0.75$.

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sc1	2	t_0	-0.569 (-0.777, -0.361)	-0.631 (-0.706, -0.555)	-0.650 (-0.687, -0.613)
		t_1	0.256 (-0.204, 0.716)	0.275 (0.107, 0.444)	0.282 (0.2, 0.364)
	1.25	t_0	0.433 (-8.456E+04, 8.456E+04)	0.239 (0.06, 0.419)	0.202 (0.121, 0.282)
		t_1	0.51 (-1.556E+06, 1.556E+06)	0.436 (-0.049, 0.921)	0.444 (0.225, 0.663)
	2.5	t_0	-0.904 (-1.067, -0.742)	-0.956 (-1.018, -0.895)	-0.973 (-1.003, -0.943)
		t_1	0.229 (-0.122, 0.579)	0.246 (0.114, 0.378)	0.252 (0.187, 0.317)
		phi	0.766 (0.737, 0.797)	0.749 (0.738, 0.761)	0.744 (0.738, 0.75)
sca	1.25	t_0	0.826 (-3.410E+07, 3.410E+07)	0.618 (0.492, 0.744)	0.587 (0.531, 0.643)
		t_1	5.142 (-3.665E+07, 3.665E+07)	4.927 (-5.095E+05, 5.095E+05)	4.494) (-4.147E+05, 4.147E+05
	2	t_0	-0.343 (-0.448, -0.237)	-0.395 (-0.436, -0.353)	-0.413 (-0.434, -0.391)
		t_1	1.308 (1.068, 1.548)	1.310 (1.212, 1.408)	1.279 (1.227, 1.33)
	2.5	t_0	-0.717 (-0.799, -0.635)	-0.760 (-1.091, -0.43)	-0.776 (-0.793, -0.759)
		t_1	1.019 (0.856, 1.182)	1.003 (0.935, 1.07)	0.989 (0.954, 1.024)
		phi	1.036 (1.014, 1.06)	0.993 (0.984, 1.003)	0.982 (0.977, 0.987)

continued on next page

continued from previous page

			sample size		
	M		n=100	n=250	n=500
			mean (95% CI)	mean (95% CI)	mean (95% CI)
sce	2	t_0	-0.417 (-0.632, -0.201)	-0.476 (-0.553, -0.398)	-0.495 (-0.533, -0.457)
		t_1	0.050 (-1.073E+05, 1.073E+05)	0.010 (-0.336, 0.357)	0.007 (-0.157, 0.172)
	1.25	t_0	0.778 (-234901.106, 234902.662)	0.484 (-19.39, 20.358)	0.440 (0.341, 0.539)
		t_1	2.534E+06 (-1.563E+06, 6.630E+06)	0.202 (-1.270E+06, 1.270E+06)	0.038 (-101.784, 101.86)
	2.5	t_0	-0.770 (-0.932, -0.607)	-0.882 (-0.943, -0.821)	-0.834 (-0.864, -0.804)
		t_1	0.003 (-249.109, 249.115)	0.004 (-0.268, 0.276)	0.004 (-0.126, 0.135)
		phi	0.781 (0.739, 0.826)	0.763 (0.746, 0.78)	0.756 (0.748, 0.764)

Table B.9 Estimated values of the dispersion parameter, ϕ, at different values of x_1 calculated using the mean values of the parameters for dispersion models estimated based on data simulated to have a scalar dispersion structure, $\phi = 0.75$.

		sample size					
		n=100		n=250		n=500	
		x_1=0	x_1=1	x_1=0	x_1=1	x_1=0	x_1=1
sc1	M=2	0.723	0.845	0.695	0.824	0.686	0.818
	M=1.25	0.758	0.900	0.699	0.828	0.688	0.820
	M=2.5	0.720	0.843	0.694	0.824	0.686	0.818
sca	M=2	0.830	1.448	0.805	1.428	0.797	1.408
	M=1.25	0.869	1.247	0.812	1.245	0.803	1.242
	M=2.5	0.820	1.437	0.796	1.401	0.788	1.383
sce	M=2	0.795	0.819	0.767	0.771	0.757	0.761
	M=1.25	0.857	1.250	0.773	0.831	0.760	0.772
	M=2.5	0.791	0.793	0.732	0.734	0.757	0.759

Table B.10 Selected characteristics of the variables included in the models estimated in the case study ($n = 126$).

		Variable Name	Mean (Std. Dev)	Range (Min., Max.)
Number of Bids		**numbids**	1.738 (1.432)	(0, 10)
Size		**size**	1.219	(0.018, 22.169)
Size^2		**sizesq**	10.999	(0.000, 491.465)
% Institutional Holding		**insthold**	0.252 (0.186)	(0, 0.904)
Centered Bid Premium		**cbidprem**	0	(-0.404, 0.720)
			Freq.	%
Legal Defense		**leglrest**		
	Yes		72	57.143
	No		54	42.857
Change in Asset		**realrest**		
Structure	Yes		23	18.254
	No		103	81.746
Change in		**finrest**		
Ownership	Yes		13	10.317
	No		113	89.683
3rd Party Bid		**whtknght**		
	Yes		75	59.524
	No		51	40.476
Intervention by		**regulatn**		
Fed	Yes		34	26.984
	No		92	73.016

Double Poisson regression — Number of obs = 126
phi>1 : underdispersion — LR chi2(9) = 46.68
phi<1 : overdispersion — Prob > chi2 = 0.0000
Log likelihood = -177.13726 — Pseudo R2 = 0.1164

numbids	IRR	Std. err.	z	P>|z|	[95% conf.	interval]
leglrest	1.301914	.1541564	2.23	0.026	1.03227	1.641991
realrest	.8202227	.123712	-1.31	0.189	.6103055	1.102342
finrest	1.078113	.1834059	0.44	0.658	.7724313	1.504764
whtknght	1.625216	.2012664	3.92	0.000	1.274966	2.071685
cbidprem	.5044857	.148506	-2.32	0.020	.2833216	.8982927
insthold	.6942976	.2307799	-1.10	0.272	.3619194	1.331924
size	1.198548	.0563618	3.85	0.000	1.093019	1.314266
sizesq	.9923538	.0024226	-3.14	0.002	.9876169	.9971133
regulatn	.9707207	.1220397	-0.24	0.813	.7587185	1.241961
_cons	1.054636	.1324082	0.42	0.672	.8245845	1.348869
/lnphi	.497378	.1118416			.2781725	.7165836
phi	1.644404	.1839128			1.320714	2.047426

Note: Estimates are transformed only in the first equation to incidence-rate ratios.
Note: **_cons** estimates baseline incidence rate.

Figure B.1 Results of the double Poisson regression model of the number of takeover bids assuming a scalar dispersion structure.

```
Double Poisson (variable dispersion) regression   Number of obs   =       126
phi>1 : underdispersion                           LR chi2(9)      =     49.42
phi<1 : overdispersion                            Prob > chi2     =    0.0000
Log likelihood = -172.77451                       Pseudo R2       =    0.1251
```

numbids	IRR	Std. err.	z	P>|z|	[95% conf.	interval]
numbids						
leglrest	1.215362	.1468786	1.61	0.107	.9590399	1.540192
realrest	.7379683	.1227757	-1.83	0.068	.5326273	1.022473
finrest	1.220229	.2089811	1.16	0.245	.872291	1.706952
whtknght	1.646994	.2048835	4.01	0.000	1.290636	2.101746
cbidprem	.4521896	.1334818	-2.69	0.007	.2535447	.8064669
insthold	.5687703	.196459	-1.63	0.102	.289017	1.11931
size	1.233312	.0605021	4.27	0.000	1.120252	1.357782
sizesq	.9910425	.0024945	-3.57	0.000	.9861653	.9959438
regulatn	1.056664	.1338633	0.44	0.664	.8243333	1.354476
_cons	1.082883	.1377198	0.63	0.531	.8439694	1.389429
logitphi						
cbidprem	4.339814	1.765413	2.46	0.014	.8796679	7.799961
_cons	.0566661	.2574851	0.22	0.826	-.4479955	.5613276

```
Note: Estimates are transformed only in the first equation to incidence-rate ratios.
Note: _cons estimates baseline incidence rate.
Note: Dispersion parameter <= specified value (3.5)
Likelihood ratio test of phi=1:  chi2(1) =    24.35    Prob>=chi2 =     0.0000
```

Figure B.2 Results of the double Poisson regression model of the number of takeover bids. Dispersion is modeled as a function of mean-centered bid premium.

Appendix C

Appendix - Evaluating and Improving Identification of Tic Disorders in Children

Table C.1 DoTS items included in the parent report. Items in the self report are similar.

DoTS - Parent Report					
1a	My child makes noises (like grunts) that he/she can't stop.	Never	Sometimes	Often	Always
1b	Parts of my child's body jerk again and again, that he/she can't control.	Never	Sometimes	Often	Always
1c	At times my child has the same jerk or twitch over and over.	Never	Sometimes	Often	Always
1d	My child can't control all of his/her movements.	Never	Sometimes	Often	Always
1e	My child seems to feel pressure to talk, shout, or scream.	Never	Sometimes	Often	Always
1f	My child has habits or movements that come out more when he/she is nervous.	Never	Sometimes	Often	Always
2	Does your child make short movements over and over?	No	Yes		
3	Does your child make short sounds over and over?	No	Yes		
4a	Do you think that your child ever had tics?	No	Yes		
4b	Do you think that your child currently has tics?	No	Yes		

Table C.2 MOVeIT-10 items included in the parent report. Items in the self report are similar.

MOVeIT 10 – Parent Report				
Please answer the questions below by circling your response. Some of the questions may sound similar, but please answer each the best you can.				
1	My child makes the same twitches, movements, noises, words or sounds over and over	Never	Sometimes	Often
2	My child feels like they have to make a noise or say a word even if they don't want to	Never	Sometimes	Often
3	My child feels like they have to move parts of their body even if they don't want to - like constant blinking, twitching the nose, moving mouth or jaw, shrugging the shoulders, jerking arms or legs	Never	Sometimes	Often
4	My child makes the same twitches, movements, noises, words or sounds over and over that are hard to keep from doing – like grunts, coughs, blinking, shrugging the shoulders	Never	Sometimes	Often
5	My child has the same jerk or twitch over and over – like constant blinking, twitching the nose, moving mouth or jaw, shrugging the shoulders, jerking arms or legs	Never	Sometimes	Often
6	My child makes the same twitches, movements, noises, words or sounds over and over - like grunts, coughs, blinking, shrugging shoulders	Never	Sometimes	Often
7	My child makes the same noises or sounds over and over that are hard to keep from doing	Never	Sometimes	Often
8	My child feels like they have to make a noise or say a word, or move parts of my body even if they don't want to	Never	Sometimes	Often
9	My child makes the same noises or says the same words over and over	Never	Sometimes	Often
10	My child makes the same movements over and over that are hard to keep from doing	Never	Sometimes	Often

Table C.3 Data source information including study name, gold standard used to determine tic disorder status, original sample size, number of participants with known tic disorder status, and the source population

Study	Gold Standard	Sample Size	Observations with Known Tic Disorder Status	Population
PLAY-MH DoTS	KSADS	100	93	Population-based epidemiologic study
USF MOVeIT development	YGTSS	42	37	Tic disorder specialty clinic and general pediatric clinic
UR Validation Study MOVeIT & UR Validation Study DoTS	Expert Determination	100	100	Tic disorder specialty clinic and community controls
UR AUCD Project MOVeIT	Expert Determination	266	199	Developmental and behavioral pediatrics clinic
Tics as a Marker USF MOVeIT & USF DoTS	YGTSS, KSADS	799	776	Primary care pediatric clinic

Table C.4 Demographic information for each study including sex, race, and age.

Study	Sex - Female (%)	Race - Non-Hispanic White (%)	Race - Non-Hispanic Black (%)	Race - Other/Multiple (%)	Age Range (Median)
PLAY-MH DoTS	40 (43%)	46 (49.5%)	28 (30.1%)	19 (20.4%)	8-20 (12)
USF MOVeIT development	15 (40.5%)	9 (24.3%)	16 (43.2%)	12 (32.4%)	4-17 (10)
UR Validation Study MOVeIT & UR Validation Study DoTS	37 (37.0%)	87 (87.0%)	2 (2.0%)	11 (11.0%)	6-17 (11)
UR AUCD Project MOVeIT	46 (22.8%)	163 (80.7%)	16 (7.9%)	13 (6.4%)	2-16 (7)
Tics as a Marker USF MOVeIT & USF DoTS	362 (46.6%)	232 (29.9%)	273 (35.2%)	270 (34.8%)	4-17 (9)

Table C.5 Prediction results for DoTS logistic regression (LR) and random forest (RF) models estimated using both child and parent responses, only parent responses, and only child responses. Sensitivity, NPV, Youden's indicies, weighted Youden's indices, and specificity reported at the cutoff value that maximizes weighted Youden's index Y3.

	LR DoTS	LR DoTS Parent	LR DoTS Self
Cutoff	0.48	0.28	0.34
Sensitivty	0.909	0.850	0.938
NPV	0.974	0.961	0.976
Youden's Index	0.859	0.799	0.810
Y2	2.768	2.649	2.747
Y3	3.677	3.499	3.685
Y4	4.586	4.349	4.622
Y5	5.495	5.199	5.560
Specificity	0.950	0.949	0.872
	RF DoTS	RF DoTS Parent	RF DoTS Self
Cutoff	0.500	0.500	0.350
Sensitivty	0.909	0.850	0.938
NPV	0.974	0.961	0.975
Youden's Index	0.834	0.786	0.767
Y2	2.743	2.636	2.705
Y3	3.652	3.486	3.642
Y4	4.561	4.336	4.580
Y5	5.470	5.186	5.517
Specificity	0.925	0.936	0.830

Table C.6 Prediction results for MOVeIT logistic regression (LR) and random forest (RF) models estimated using both child and parent responses, only parent responses, and only child responses. Sensitivity, NPV, Youden's indicies, weighted Youden's indices, and specificity reported at the cutoff value that maximizes weighted Youden's index Y3.

	LR MOVeIT	LR MOVeIT Parent	LR MOVeIT Self
Cutoff	0.04	0.12	0.08
Sensitivity	0.767	0.804	0.846
NPV	0.873	0.933	0.907
Youden's Index	0.554	0.547	0.411
Y2	2.320	2.351	2.258
Y3	3.087	3.155	3.104
Y4	3.854	3.959	3.950
Y5	4.620	4.763	4.796
Specificity	0.787	0.743	0.565
	RF MOVeIT	RF MOVeIT Parent	RF MOVeIT Self
Cutoff	0.02	0.10	0.04
Sensitivity	0.967	0.843	0.962
NPV	0.969	0.943	0.968
Youden's Index	0.475	0.554	0.396
Y2	2.442	2.398	2.358
Y3	3.408	3.241	3.319
Y4	4.375	4.084	4.281
Y5	5.342	4.927	5.242
Specificity	0.508	0.711	0.435

Table C.7 Logistic regression coefficients and odds ratios with 95% confidence intervals for Description of Tic Symptoms (DoTS) models. Race - Black indicates non-Hispanic Black and Race - Other indicates other or multiple races and/or ethnicities.

	DoTS		DoTS Parent		DoTS Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Sex (female)	0.88 (0.31, 2.45)	0.80	0.840 (0.348, 2.029)	0.70	0.40** (0.16, 0.99)	0.05**
Race - Black	0.39 (0.13, 1.22)	0.11	0.657 (0.242, 1.785)	0.41	0.40* (0.15, 1.09)	0.07*
Race - Other	0.19** (0.05, 0.78)	0.02**	0.379* (0.132, 1.085)	0.07*	0.32* (0.11, 1.00)	0.05*
Age	1.05 (0.88, 1.25)	0.61	1.080 (0.971, 1.200)	0.16	1.03 (0.88, 1.21)	0.69
Parent 1a	2.83** (1.12, 7.18)	0.03**	1.75* (0.91, 3.34)	0.09*		
Parent 1b	0.88 (0.18, 4.41)	0.88	0.83 (0.31, 2.20)	0.71		
Parent 1c	3.20 (0.53, 19.23)	0.20	2.28 (0.85, 6.07)	0.10		
Parent 1d	1.35 (0.42, 4.29)	0.62	1.25 (0.66, 2.35)	0.49		
Parent 1e	0.78 (0.25, 2.40)	0.66	0.48* (0.21, 1.07)	0.07*		
Parent 1f	0.82 (0.35, 1.90)	0.64	1.15 (0.66, 2.00)	0.63		
Parent 2	1.55 (0.16, 15.19)	0.71	2.67* (0.93, 7.70)	0.07*		
Parent 3	0.82 (0.13, 5.26)	0.83	1.92 (0.66, 5.60)	0.24		

continued on next page

continued from previous page

	DoTS		DoTS Parent		DoTS Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Parent 4a	4.68* (0.78, 28.07)	0.09*	4.66** (1.11, 19.50)	0.04**		
Parent 4b	1.442 (0.10, 20.13)	0.79	1.90 (0.38, 9.31)	0.43		
Self 1a	0.76 (0.31, 1.88)	0.56			1.35 (0.71, 2.57)	0.36
Self 1b	0.34** (0.12, 0.99)	0.05**			0.64 (0.30, 1.38)	0.25
Self 1c	1.53 (0.63, 3.72)	0.35			1.82* (0.92, 3.57)	0.08*
Self 1d	1.41 (0.69, 2.89)	0.35			1.67 (0.87, 3.21)	0.12
Self 1e	1.23 (0.66, 2.30)	0.52			0.64 (0.36, 1.12)	0.12
Self 1f	0.85 (0.44, 1.65)	0.64			0.75 (0.44, 1.28)	0.29
Self 2	2.02 (0.70, 5.84)	0.20			1.55 (0.62, 3.85)	0.35
Self 3	1.60 (0.50, 5.13)	0.43			1.78 (0.65, 4.83)	0.26
Self 4a	8.55*** (2.13, 34.37)	0.00***			9.84*** (3.11, 31.14)	0.00***
Self 4b	0.35 (0.08, 1.62)	0.18			0.81 (0.25, 2.64)	0.73

Table C.8 Logistic regression coefficients and odds ratios for 10 item Motor or Vocal Inventory of Tics (MOVeIT-10) models. Race - Black indicates non-Hispanic Black and Race - Other indicates other or multiple races and/or ethnicities.

	MOVeIT		MOVeIT Parent		MOVeIT Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Sex (female)	0.24* (0.05, 1.10)	0.07*	0.68 (0.38, 1.23)	0.20	0.19*** (0.07, 0.53)	0.00***
Race - Black	0.22* (0.04, 1.33)	0.10*	0.72 (0.38, 1.40)	0.33	0.40* (0.14, 1.18)	0.10*
Race - Other	0.24 (0.04, 1.47)	0.12	0.31*** (0.13, 0.72)	0.01***	0.36* (0.11, 1.17)	0.09*
Age	0.96 (0.72, 1.27)	0.75	1.07* (0.99, 1.15)	0.05*	1.05 (0.87, 1.26)	0.61
Parent 1	1.04 (0.20, 5.55)	0.96	1.42 (0.79, 2.58)	0.24		
Parent 2	0.18 (0.02, 1.52)	0.12	1.14 (0.63, 2.08)	0.67		
Parent 3	0.89 (0.19, 4.20)	0.89	1.04 (0.57, 1.91)	0.90		
Parent 4	2.92 (0.54, 15.69)	0.21	1.33 (0.66, 2.67)	0.42		
Parent 5	2.79 (0.53, 14.61)	0.22	2.35*** (1.23, 4.49)	0.01***		
Parent 6	0.67 (0.13, 3.52)	0.63	1.66 (0.77, 3.61)	0.20		
Parent 7	11.91*** (2.34, 60.54)	0.00***	1.09 (0.54, 2.17)	0.82		
Parent 8	0.90 (0.17, 4.84)	0.90	1.50 (0.77, 2.91)	0.24		

continued on next page

continued from previous page

	MOVeIT		MOVeIT Parent		MOVeIT Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Parent 9	4.65 (0.71, 30.37)	0.11	0.73 (0.39, 1.36)	0.32		
Parent 10	0.85 (0.22, 3.19)	0.80	0.98 (0.54, 1.79)	0.94		
Self 1	2.38 (0.68, 8.30)	0.17			2.51** (1.11, 5.67)	0.03**
Self 2	0.44 (0.11, 1.81)	0.26			1.96 (0.73, 5.28)	0.18
Self 3	0.95 (0.29, 3.15)	0.94			1.05 (0.47, 2.38)	0.90
Self 4	1.60 (0.34, 7.55)	0.55			1.25 (0.52, 2.97)	0.62
Self 5	0.78 (0.22, 2.76)	0.70			1.23 (0.56, 2.71)	0.61
Self 6	0.92 (0.24, 3.55)	0.90			1.27 (0.49, 3.31)	0.62
Self 7	0.22** (0.05, 0.94)	0.04**			0.94 (0.34, 2.59)	0.90
Self 8	0.55 (0.16, 1.90)	0.35			0.42* (0.17, 1.06)	0.07*
Self 9	4.63*** (1.55, 13.87)	0.01***			1.72 (0.66, 4.52)	0.27
Self 10	4.73*** (1.54, 14.55)	0.01***			2.39* (0.97, 5.90)	0.06

Table C.9 Random forest variable importance results for Description of Tic Symptoms (DoTS) models. All responses to questionnaire items are treated as categorical.

Item	DoTS	DoTS Parent	DoTS self
Sex	-1.05	-0.68	3.92
Race	4.95	8.28	12.35
Age	5.91	8.13	1.03
Parent 1a	23.61	15.80	
Parent 1b	8.89	11.90	
Parent 1c	22.35	23.32	
Parent 1d	8.35	9.30	
Parent 1e	-2.26	-1.04	
Parent 1f	1.23	9.48	
Parent 2	10.42	12.93	
Parent 3	6.14	6.24	
Parent 4a	17.29	32.00	
Parent 4b	9.80	21.70	
Self 1a	0.68		4.01
Self 1b	1.08		-0.07
Self 1c	4.30		9.93
Self 1d	5.11		5.50
Self 1e	8.29		1.41
Self 1f	0.67		-0.55
Self 2	2.57		0.68
Self 3	2.94		-0.38
Self 4a	13.04		30.43
Self 4b	5.83		20.99

Table C.10 Random forest variable importance results for 10 item Motor or Vocal Inventory of Tics (MOVeIT-10) models All responses to questionnaire items are treated as categorical.

Item	MOVeIT	MOVeIT Parent	MOVeIT self
Sex	2.571	-0.332	12.814
Race	8.04	12.79	13.435
Age	-0.517	18.202	0.159
Parent 1	17.834	16.614	
Parent 2	2.205	7.433	
Parent 3	5.133	9.865	
Parent 4	14.612	18.816	
Parent 5	15.301	27.224	
Parent 6	15.53	19.081	
Parent 7	13.026	4.779	
Parent 8	5.361	9.102	
Parent 9	9.455	5.88	
Parent 10	8.556	16.107	
Self 1	9.033		18.184
Self 2	4.968		14.953
Self 3	0.591		6.323
Self 4	8.317		16.822
Self 5	0.751		8.734
Self 6	5.492		14.282
Self 7	1.876		11.206
Self 8	1.297		4.531
Self 9	17.18		14.887
Self 10	9.009		11.387

Table C.11 Sensitivity analysis logistic regression coefficients and odds ratios with 95% confidence intervals for Description of Tic Symptoms (DoTS) models excluding data from studies recruiting from tic specialty or developmental and behavioral specialty clinics. Race - Black indicates non-Hispanic Black and Race - Other indicates other or multiple races and/or ethnicities.

	DoTS		DoTS Parent		DoTS Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Sex (female)	1.10 (0.37, 3.27)	0.86	0.60 (0.26, 1.39)	0.23	0.46 (0.18, 1.15)	0.10*
Race - Black	0.60 (0.20, 1.8)	0.36	0.63 (0.25, 1.59)	0.32	0.58 (0.22, 1.54)	0.27
Race - Other	0.15** (0.03, 0.69)	0.02**	0.46 (0.17, 1.24)	0.12	0.26** (0.08, 0.84)	0.02**
Age	0.94 (0.80, 1.10)	0.43	1.13** (1.02, 1.24)	0.02**	0.87 (0.74, 1.03)	0.10
Parent 1a	2.22* (0.91, 5.41)	0.08*	3.06*** (1.4, 6.69)	0.01***		
Parent 1b	1.46 (0.3, 7.03)	0.63	0.78 (0.25, 2.40)	0.66		
Parent 1c	2.72 (0.57, 13.08)	0.21	1.45 (0.39, 5.40)	0.58		
Parent 1d	0.84 (0.13, 5.24)	0.85	0.60 (0.25, 1.45)	0.26		
Parent 1e	0.61 (0.16, 2.31)	0.47	0.54 (0.23, 1.27)	0.16		
Parent 1f	0.76 (0.32, 1.8)	0.53	1.20 (0.69, 2.09)	0.52		
Parent 2	3.03 (0.41, 22.2)	0.28	2.85* (0.91, 8.96)	0.07*		
Parent 3	0.55 (0.07, 4.13)	0.56	0.92 (0.28, 3.01)	0.89		

continued on next page

continued from previous page

	DoTS		DoTS Parent		DoTS Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Parent 4a	4.61* (0.81, 26.26)	0.09*	2.36 (0.60, 9.37)	0.22		
Parent 4b	1.27 (0.09, 17.13)	0.86	1.16 (0.21, 6.52)	0.87		
Self 1a	0.98 (0.31, 3.11)	0.97			0.74 (0.26, 2.09)	0.57
Self 1b	0.61 (0.23, 1.62)	0.32			0.4 0* (0.15, 1.04)	0.06*
Self 1c	1.84 (0.74, 4.56)	0.19			2.38 (1.12, 5.07)	0.02
Self 1d	1.11 (0.51, 2.41)	0.8			0.91 (0.46, 1.79)	0.78
Self 1e	1.28 (0.69, 2.37)	0.43			1.13 (0.66, 1.95)	0.65
Self 1f	0.67 (0.33, 1.33)	0.25			1.10 (0.62, 1.95)	0.74
Self 2	3.36 (1.10, 10.31)	0.2			1.61 (0.59, 4.38)	0.35
Self 3	0.58 (0.14, 2.34)	0.45			1.08 (0.36, 3.22)	0.89
Self 4a	4.23 (1.11, 16.16)	0.03			5.21*** (1.5, 18.16)	0.01***
Self 4b	0.21* (0.04, 1.15)	0.07*			0.30 (0.07, 1.28)	0.10

Table C.12 Sensitivity analysis random forest variable importance results for Description of Tic Symptoms (DoTS) models excluding data from studies recruiting from tic specialty or developmental and behavioral specialty clinics. All responses to questionnaire items are treated as categorical.

Item	DoTS	DoTS Parent	DoTS self
Sex	-2.10	3.65	-1.11
Race	-0.79	-4.06	-0.09
Age	5.79	10.56	2.81
Parent 1a	15.88	18.15	
Parent 1b	1.62	0.04	
Parent 1c	8.46	1.52	
Parent 1d	-2.54	-8.32	
Parent 1e	-4.07	-1.97	
Parent 1f	1.83	-9.38	
Parent 2	4.24	0.94	
Parent 3	0.40	-3.87	
Parent 4a	5.74	10.50	
Parent 4b	2.21	6.68	
Self 1a	4.06		-3.61
Self 1b	-0.90		-0.01
Self 1c	7.07		4.92
Self 1d	1.54		-0.20
Self 1e	4.83		2.25
Self 1f	3.00		-0.94
Self 2	1.39		-0.83
Self 3	1.33		-0.23
Self 4a	5.51		5.10
Self 4b	0.27		2.41

Table C.13 Sensitivity analysis logistic regression coefficients and odds ratios with 95% confidence intervals for 10 item Motor or Vocal Inventory of Tics (MOVeIT-10) models excluding data from studies recruiting from tic specialty or developmental and behavioral specialty clinics. Race - Black indicates non-Hispanic Black and Race - Other indicates other or multiple races and/or ethnicities.

	MOVeIT		MOVeIT Parent		MOVeIT Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Sex (female)	0.49 (0.09, 2.62)	0.41	0.61 (0.31, 1.19)	0.15	0.34* (0.09, 1.21)	0.10*
Race - Black	0.60 (0.10, 3.68)	0.58	0.85 (0.40, 1.80)	0.67	0.91 (0.25, 3.38)	0.89
Race - Other	0.88 (0.14, 5.30)	0.89	1.00 (0.50, 2.01)	0.99	0.69 (0.16, 3.07)	0.63
Age	0.80 (0.53, 1.21)	0.29	1.05 (0.97, 1.14)	0.20	1.02 (0.80, 1.30)	0.87
Parent 1	1.49 (0.21, 10.84)	0.69	1.49 (0.78, 2.85)	0.23		
Parent 2	0.61 (0.06, 5.95)	0.67	0.89 (0.48, 1.65)	0.72		
Parent 3	1.25 (0.24, 6.46)	0.79	1.08 (0.57, 2.07)	0.81		
Parent 4	5.93 (0.41, 85.25)	0.19	1.40 (0.70, 2.80)	0.35		
Parent 5	1.96 (0.19, 20.03)	0.57	1.41 (0.72, 2.76)	0.31		
Parent 6	1.24 (0.06, 24.57)	0.89	2.13** (1.01, 4.48)	0.05**		
Parent 7	7.90** (1.28, 48.78)	0.03**	0.80 (0.39, 1.66)	0.55		
Parent 8	0.66 (0.09, 5.05)	0.69	0.97 (0.46, 2.06)	0.94		

continued on next page

continued from previous page

	MOVeIT		MOVeIT Parent		MOVeIT Self	
	OR (95% CI)	p-value	OR (95% CI)	p-value	OR (95% CI)	p-value
Parent 9	0.09 (0.00, 2.19)	0.14	1.05 (0.57, 1.92)	0.89		
Parent 10	0.44 (0.08, 2.32)	0.33	1.07 (0.57, 2)	0.84		
Self 1	1.11 (0.26, 4.64)	0.89			1.33 (0.47, 3.72)	0.59
Self 2	1.93 (0.49, 7.5)	0.35			2.69* (0.95, 7.56)	0.06*
Self 3	3.53* (0.88, 14.1)	0.07*			1.01 (0.35, 2.94)	0.98
Self 4	0.75 (0.10, 5.41)	0.78			0.93 (0.30, 2.93)	0.90
Self 5	1.02 (0.23, 4.54)	0.98			0.80 (0.26, 2.47)	0.70
Self 6	1.88 (0.33, 10.75)	0.48			1.57 (0.44, 5.61)	0.49
Self 7	0.35 (0.07, 1.72)	0.20			0.55 (0.16, 1.87)	0.34
Self 8	0.67 (0.15, 2.97)	0.60			0.87 (0.27, 2.81)	0.81
Self 9	1.29 (0.34, 4.87)	0.71			1.72 (0.56, 5.26)	0.34
Self 10	2.71 (0.65, 11.34)	0.17			2.08 (0.79, 5.47)	0.14

Table C.14 Sensitivity analysis random forest variable importance results for 10 item Motor or Vocal Inventory of Tics (MOVeIT-10) models excluding data from studies recruiting from tic specialty or developmental and behavioral specialty clinics. All responses to questionnaire items are treated as categorical.

Item	MOVeIT	MOVeIT Parent	MOVeIT self
Sex	-3.51	-1.43	1.10
Race	-2.30	-1.14	-2.78
Age	2.30	3.12	-1.51
Parent 1	7.52	3.66	
Parent 2	-2.06	1.01	
Parent 3	-0.56	1.67	
Parent 4	8.32	9.16	
Parent 5	4.62	11.88	
Parent 6	5.23	10.02	
Parent 7	7.69	4.75	
Parent 8	1.10	0.43	
Parent 9	0.58	2.09	
Parent 10	2.27	5.30	
Self 1	2.00		0.57
Self 2	2.16		4.80
Self 3	5.44		0.90
Self 4	3.04		2.73
Self 5	4.79		-0.86
Self 6	2.25		-1.92
Self 7	1.68		-2.18
Self 8	1.50		-0.34
Self 9	0.82		0.09
Self 10	2.75		3.27