Promoting Data Science

Volume 5 2025 Number 1

Journal of Behavioral Data Science V5N1 (2025)

https://isdsa.org

JOURNAL OF BEHAVIORAL DATA SCIENCE

ISBN: 2575-8306 (Print) 2574-1284 (Online)

https://jbds.isdsa.org

JOURNAL OF BEHAVIORAL DATA SCIENCE

No Publication Charge and Open Access

jbds@isdsa.org

List of Articles

Journal of Behavioral Data Science, 2025, *5* (1), 1–22.
DOI: https://doi.org/10.35566/jbds/moulder

A Data Permutation Method for Testing Random Slopes of Bayesian Growth Curves

Robert G. Moulder[1][0000−0001−7504−9560] and Xin Tong[2][0000−0003−3050−1554]

[1] Institute of Cognitive Science, University of Colorado Boulder
[2] Department of Psychology, University of Virginia
xt8b@virginia.edu

Abstract. Growth curve analysis is a popular method for modeling individual development across time. Specifying growth curve models in a Bayesian framework affords researchers the flexibility of including previous information as prior distributions of parameters. However, common choices of prior distribution for modeling slope variance in a Bayesian growth curve framework make determining the existence of meaningful interindividual differences in intraindividual change across time difficult due to boundary values of these priors. Additionally, many current methods are either technically difficult to implement or are sensitive to model specification. We present a simple data permutation method that reliably distinguishes between longitudinal data with individual slope variation and those without slope variation. We show situations in that the proposed data permutation testing outperforms DIC based model comparison through Monte Carlo simulations and apply this data permutation method to data derived from the National Longitudinal Study of Adolescent to Adult Health.

Keywords: Bayesian Growth Curve Modeling · Random Slope Testing · Longitudinal Data Analysis · Permutation Testing.

1 Introduction

Longitudinal research design is a powerful framework for testing psychological hypotheses regarding change. In such a framework, researchers measure the same construct from multiple participants across multiple time points so as to study how a given psychological process changes over time (Baltes & Nesselroade, 1979). Due to the versatility and statistical power afforded by longitudinal research designs, researchers have been able to study time-varying phenomena such as patterns and outcomes of drug use among adolescents, trajectories of public reaction to large-scale disasters, and stability of personality traits across time (Roberts, Walton, & Viechtbauer, 2006; Shedler & Block, 1990; Silver, Holman, McIntosh, Poulin, & Gil-Rivas, 2002). By collecting data in a longitudinal

manner, researchers are able to simultaneously study how a given psychological construct changes within an individual and what factors influence the varying trajectories of said construct among different individuals.

Although collecting data in a longitudinal manner may be more difficult than collecting data in a single wave, advances in data collection technologies have made longitudinal research designs accessible to many researchers. As such, in recent years longitudinal research designs have become commonplace in psychological research. A Google Scholar search for the terms "longitudinal", "research", and "psychology" shows an increase in number of related works from about 140,000 results in the 1990s to more than 1,200,000 related works between 2010 and 2020. With this increase in popularity of longitudinal research designs there has also come an increase in the number and quality of statistical methods for analysing longitudinal data. Although varied, each method provides researchers some insight into how psychological processes change over time.

1.1 Statistical Methods for Longitudinal Research

Statistical methods for longitudinal data analysis help researchers to understand both intraindividual change and interindividual differences in intraindividual change across time. That is, researchers may use statistical methods for longitudinal data analysis in order to gain a deeper understanding of how individuals change over time with respect to a variable of interest and how different individuals may show different patterns of change. Growth curves modeling is one popular way of assessing these qualities given a longitudinal sample of participants (Grimm, Ram, & Estabrook, 2016; Hertzog & Nesselroade, 2003; Oravecz & Muth, 2018). Growth curve models have been used by researchers to study a wide variety of phenomena such as academic trajectories of children, the development of individuals' self-esteem, and changes in depressive symptoms of adolescents over time (Baldwin & Hoffmann, 2002; Gomez-Baya, Mendoza, Paino, Sanchez, & Romero, 2016; Gutman, Sameroff, & Cole, 2003). Due to the simplicity and flexibility of growth curve models, different researchers may use different statistical frameworks for estimating growth curve models. Such statistical frameworks for conducting growth curve analyses include mixed-effects modeling/multilevel modeling and structural equation modeling.

Across statistical frameworks, growth curve models generally take the form:

$$Y_{ij} = \beta_0 + \beta_1 T_j + u_{0i} + u_{1i} T_j + \epsilon_{ij}, \tag{1}$$

where Y_{ij} is the realizatio of an outcome variable from person i at time j, $i = 1, \ldots, N$, $j = 1, \ldots, K$, where N is the sample size and K is the total number of measurement occasions, β_0 is a fixed effect representing the average intercept value at time $T_j = 0$ for all participants, β_1 is a fixed effect representing the average slope over time, u_{0i} is a random component of intercept for each individual i with variance $\sigma^2_{u_0}$, at time $T_j = 0$, u_{1i} is a random component of slope for each individual i with variance $\sigma^2_{u_1}$, and ϵ is an error term with variance σ^2_ϵ. Specific and meaningful interpretation of these parameters have allowed for

growth curve modeling to become a common tool for studying change (McArdle & Nesselroade, 2003). Fixed-effect parameters relate to general trends across all participants, while random-effect parameters relate to individual participant variation from this overall group level behavior. Multiple statistical software packages are capable of estimating parameters of growth curve models using various techniques.

1.2 Bayesian Growth Curve Modeling

Bayesian analysis is one way of estimating growth curve models for a given longitudinal data set (Fearn, 1975; Oravecz & Muth, 2018; Zhang, Hamagami, Wang, Nesselroade, & Grimm, 2007). Compared to other analysis frameworks, Bayesian analysis allows researchers a high degree of flexibility in modeling complex longitudinal patterns of change. While many modern analysis methods have strict assumptions of normality and other asymptotic assumptions, researchers using Bayesian analyses are generally not limited by these concerns asprior distributions of all variables can be explicitly and flexibly modeled (Bayarri & Berger, 2004). Thus common longitudinal data analysis problems such as sample size restrictions, non-normal data distributions, and missing data patterns due to attrition are more easily handled in a Bayesian framework than in a frequentist framework. Additionally, advancement in computational efficiency and Bayesian analysis software has helped ease the burden of conducting Bayesian analysis put on researchers new to Bayesian modeling (e.g., JAGS, STAN, BUGS).

In a Bayesian framework, parameters of a growth curve model are treated as random variables whose realizations are modeled using some form of a Markov chain Monte Carlo (MCMC) process such as Gibbs sampling to sample from constantly updated distributions (Carlin & Chib, 1995; Gilks, Wang, Yvonnet, & Coursaget, 1993). Equation (1) can also be expressed as:

$$\begin{aligned} Y_{ij} &\sim N(\bar{Y}_{ij}, \sigma^2_\epsilon) \\ \bar{Y}_{ij} &= b_{0i} + b_{1i}T_j \\ b_{0i} &\sim N(\beta_0, \sigma^2_{u_0}) \\ b_{1i} &\sim N(\beta_1, \sigma^2_{u_1}), \end{aligned} \tag{2}$$

where $\bar{Y}_{ij}$ is the expected value of Y_{ij}. This Bayesian parameterization of a growth curve model allows researchers to use previous knowledge to hypothesize the prior distributions of the parameters $\beta_0, \beta_1, \sigma^2_{u_0}, \sigma^2_{u_1}$, and σ^2_ϵ. Parameters b_{0i} and b_{1i} may also be correlated. In such a case an additional parameter, $\sigma_{u_0u_1}$, is also modeled. Typically researchers set priors for β_0 and β_1 as either normal or uniform distributions, while setting priors of the variance components $\sigma^2_{u_0}$, $\sigma^2_{u_1}$, and σ^2_ϵ as inverse gamma distributions, although other distributions have been assessed (Gelman, 2006; Zhang, 2016; Zhang et al., 2007). These priors are then iteratively updated into posterior distributions using data. After a large number of iterations, a Bayesian model will converge, parameter estimates will remain stable, and researchers may draw statistical inference.

Substantive researchers routinely need to determine the statistical significance of each parameter. Credible intervals are a commonly used in Bayesian growth curve modeling (Zhang et al., 2007). A $100 \times (1-\alpha)\%$ credible interval for a parameter is as an interval for which there is at least a $100 \times (1-\alpha)\%$ chance said interval contains the true value of a given parameter, conditional on a given data set. Similar to a frequentist confidence interval, a parameter is considered significant at the α-level when a $100 \times (1-\alpha)\%$ credible interval for said parameter does not include 0. While versatile, credible intervals are not useful for testing variance components of Bayesian growth curves. This is because the gamma/inverse gamma distributions used to model such variance components are bounded $(0, \infty)$. Also, parameters with gamma/inverse gamma distributed priors tend to also have gamma/inverse gamma distributed posteriors. In such a case, a Bayesian credible interval at any α-level will never include a 0 value (Gelman, 2006). This boundary problem makes Bayesian hypothesis testing using credible intervals completely ineffective for testing variance parameters, thus making statistical inference on the existence of significant individual differences in interindividual change impossible. Fortunately, there are ways to overcome this problem. In this article, we review alternative methods to credibility intervals for testing for the existence of interindividual differences in intraindividual change in growth curve models and propose a new test based upon data permutations.

1.3 Testing for the Existence of Interindividual Differences in Intraindividual Change

This problem of determining the existence of interindividual differences in intraindividual change can be viewed as a problem of model comparison and selection. That is, determining if a model which includes a parameter indicative of interindividual differences in intraindividual change fits data better than a model without such a parameter. In determining how to specify such a model, Barr, Levy, Scheepers, and Tily (2013) argued for using the most complex structure admissible for a given data set; see also Barr (2013). Other researchers such as Bates, Kliegl, Vasishth, and Baayen (2015) and Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017), urged caution when using such an approach as more complex models may lead to convergence issues, as well as a a loss of statistical power. Model selection is key for accurately assessing all important effects, while minimizing estimation issues. Many methods currently exist for testing for significant random slope parameters within a frequentist framework by determining an optimal model structure. These include likelihood based comparison methods, penalty functions, and information criterion (Fan & Li, 2012; Peng & Lu, 2012; Stram & Lee, 1994; Vaida & Blanchard, 2005). There are currently fewer methods for testing for significant random slope paramters within a Bayesian growth curve context. Perhaps the most common methods for Bayesian model comparison are using deviance information criterion (DIC) values and Bayes factors.

Deviance information criterion Deviance information criterion is an information metric derived from the posterior distribution of the log-likelihood of a given data set and a penalization value based on the complexity of a given model (Spiegelhalter, Best, & Carlin, 1998; Spiegelhalter, Best, Carlin, & van der Linde, 2002). DIC is calculated as:

$$\begin{aligned} DIC &= E_{\theta|y}[D(\theta)] + p_D \\ D(\theta) &= -2log(L(\theta|y)) \\ p_D &= E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta]), \end{aligned} \tag{3}$$

where θ is the parameterization of a given model, $L(\theta|y)$ is the likelihood of θ given some data, y, $E_{\theta|y}[D(\theta)]$ is a the expectation of $D(\theta)$ conditional on y, and $E_{\theta|y}[\theta]$ is the expectation of θ conditional on y.

As a model's likelihood increases, $D(\theta)$ tends to 0. Conversely, as the number of parameters in a model increase, so does p_D. In this way DIC simultaneously incorporates model fit and penalizes overly complex models. For model comparison purposes on a given data set, model selection by DIC is conducted by selecting the model with a lower DIC value by at least 10 points, otherwise selecting the model with fewer parameters (Spiegelhalter et al., 1998). Thus, a researcher interested in testing for the existence of interindividual differences in intraindividual change across time within his/her own data would compare the DIC values of two competing growth cruve models. One model would allow the slope parameter to vary by participant, and another model would fix this value to be the same for all participants. Assuming a DIC difference of more than 10 points, the model with a lower DIC value would then be considered more appropriate for these data than the model with a higher DIC value (Lunn, Jackson, Best, Spiegelhalter, & Thomas, 2012).

Although DIC is a relatively reliable metric for model selection it is not without its criticisms. According to a review by Spiegelhalter et al. (2014), some of the most common criticisms of DIC is its lack of consistency and its weak theoretical justification. As an alternative to model comparison using DIC, some researchers argue for the use of Bayes factors (Ward, 2008).

Bayes factor The Bayes factor is another common measure for model comparison within a Bayesian framework (Kass & Raftery, 1995; Lodewyckx et al., 2011; Saville & Herring, 2009). Bayes factors can be thought of as a ratio of evidence for one model over another, which is evident in its calculation:

$$B = \frac{p(y|M_1)}{p(y|M_2)} = \frac{p(M_1|y)}{p(M_2|y)} \frac{p(M_2)}{p(M_1)}, \tag{4}$$

where M_1 and M_2 are different models used on the same data, y. The Bayes factor, B, can then be used for model selection. For $B > 3$, one would say that there is substantial evidence for M_2 over M_1 and thus a researcher would select M_2 as the more probable model. If however $B < \frac{1}{3}$, a researcher would select M_1

as the more probable model for the generation of y (Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019).

Although intuitive, Bayes factors can be difficult to obtain analytically and calculations for their numeric approximations can be computationally intensive for some models or require hyper-parameters to be set by a researcher. Additionally there are methods for numerically approximating Bayes factors including so called default Bayes factors, approximate Bayes factors, and Bayes factors estimated through the product space method (Lodewyckx et al., 2011; Rouder & Morey, 2012; Saville & Herring, 2009). Each of these methods for estimating Bayes factors require time and energy for a researcher to understand each method's intricacies well enough to properly implement each method. Bayes factor calculations may also be sensitive to a researcher's specification of priors (Ward, 2008). Additionally, implementations of Bayes factors have been shown to be inappropriate for many data sources and Bayes factors themselves have been argued as having frequentist properties, making many numerically approximated Bayes factors uninformative (Hoijtink, van Kooten, & Hulsker, 2016; Morey, Wagenmakers, & Rouder, 2016). Such difficulties make estimation of Bayes factors using for more complex models, such as growth curves, intractable. Indeed the authors of this article could find no reliable method for estimating Bayes factors for growth curve models as most numerical methods are either not able to take into account random effect structures or require overly sensitive hyper-parameter settings to initiate jumping behaviors between models needed to obtain proper Bayes factor approximations (Lodewyckx et al., 2011; Rouder & Morey, 2012; Saville & Herring, 2009). Many current methods that do offer Bayes factors for random effects models do not give Bayes factor values for the random effects parameters of interest in this aritcle. Thus, a researcher would find difficulty in using Bayes factors for testing for the existence of interindividual differences in intraindividual change across time. Although the DIC and Bayes factor methods are not the only methods used to assesses the random effects structure of growth models, these are perhaps the most common (Cai & Dunson, 2006; Chen & Dunson, 2003; Piironen & Vehtari, 2017; Ward, 2008).

1.4 The Proposed Method: A Data Permutation Algorithm for Testing Random Slopes

The DIC and Bayes factor methods share a common quality, each are model driven approaches. With either method, a researcher must specify two separate models that are then compared to one another. Thus, in order to test for the existence of a quality of interest within a data set, the models themselves are modified and the associated data are left alone. In contrast, data driven methods such as bootstrap analyses, randomization tests, and surrogate data analyses have been shown to also be effective at establishing existence of a specific quality of interest within a given data set (Efron, 1979; Moulder, Boker, Ramseyer, & Tschacher, 2018; Theiler, Eubank, Longtin, Galdrikian, & Farmer, 1992). These methods rely on modifying data sets through some randomized approach such

as sampling with replacement or data shuffling, to destroy qualities of order and structure within a given data set.

With this in mind, we propose a simple and relatively uncomplicated data driven method for determining the existence of interindividual differences in intraindividual change in a Bayesian growth curve framework. Namely, we propose a data permutation algorithm which effectively tests if a random slope parameter is reliably distinguishable from random noise. In terms of model selection, this would be similar to determining if the model in Equation (2) fits the data better than a simpler model with a fixed slope:

$$\begin{aligned} Y_{ij} &\sim N(\bar{Y}_{ij}, \sigma^2_\epsilon), \\ \bar{Y}_{ij} &= b_{0i} + \beta_1 T_j, \\ b_{0i} &\sim N(\beta_0, \sigma^2_{u_0}). \end{aligned} \tag{5}$$

Our proposed data permutation algorithm is as follows:

i) Create a fully specified Bayesian growth curve model (Equation 2) including a random slope term, using unaltered/original data, denoted as y_0, and store the posterior samples of $\sigma^2_{u_1}|y_0$ obtained from a MCMC procedure after a burn-in period.
ii) Consistently sort data either descending or ascending at each time point to create a second data set, y_{sort}.
iii) Rerun step i) using y_{sort}, and store the samples of $\sigma^2_{u_1}|y_{sort}$.
iv) Randomly shuffle y_0 within each time point to create a third data set, y_{shuff}.
v) Rerun step i) using y_{shuff}, and store the samples of $\sigma^2_{u_1}|y_{shuff}$.
vi) Compare the mean of the samples from $\sigma^2_{u_{1i}}|y_0$, μ_0, with the mean of the samples of $\sigma^2_{u_1}|y_{sort}$, μ_{sort}, and the mean of samples of $\sigma^2_{u_1}|y_{shuff}$, μ_{shuff}. If $|\mu_0 - \mu_{sort}| < |\mu_0 - \mu_{shuff}|$ then slope term of the model can be said to reliably vary between individuals. Else the slope term can not reliably be said to vary between participants.

To understand how this algorithm works, consider Figure 1. Across all three plots, the parameters β_0 (fixed intercept) and β_1 (fixed slope) from equation (2) are all the same. Figure 1(b) represents the kind of data one might expect to find from a given research study, with $\sigma^2_{u_0}$, $\sigma^2_{u_1}$, and σ^2_ϵ all greater than 0. We will consider this data as y_0 for this example. Figure 1(a) is the sorted version of y_0, which we call y_{sort}. Notice a few interesting qualities of y_{sort}. Firstly, no line of y_{sort} crosses another. Also, the error variance about each individually modeled line is minimized. Thus the ratio of $\sigma^2_{u_{1i}}$ to $\sigma^2_{\epsilon_{ij}}$ for y_{sort} is larger than the same ratio for y_0, assuming y_0 is not already in a sorted state. The opposite is true for y_{shuff}. Assuming that y_0 had some intrinsic structure to itself due to some true and natural underlying growth phenomenon, the ratio of $\sigma^2_{u_1}$ to σ^2_ϵ for y_{shuff} should be smaller smaller than the same ratio for y_0, Figure 1(c). The difference between the means of the posterior sampling distributions of $p(\sigma^2_{u_1}|y_{sort})$, $p(\sigma^2_{u_1}|y_0)$, and $p(\sigma^2_{u_1}|y_{shuff})$ then give a measure of how similar the three distributions of $p(\sigma^2_{u_1}|y)$ are. Thus if $|\mu_0 - \mu_{sort}| < |\mu_0 - \mu_{shuff}|$, then

the sampling distribution of the posterior distribution of $\sigma^2_{u_1}$ for the unedited data is more like the posterior distribution of $\sigma^2_{u_1}$ for data which has noticeable slope variation. If $|\mu_0 - \mu_{sort}| > |\mu_0 - \mu_{shuff}|$, then the sampling distribution of the posterior distribution of $\sigma^2_{u_1}$ for the unedited data is more like the posterior distribution of $\sigma^2_{u_1}$ for data which has been randomly shuffled and has slope variation that is difficult to distinguish from random noise.

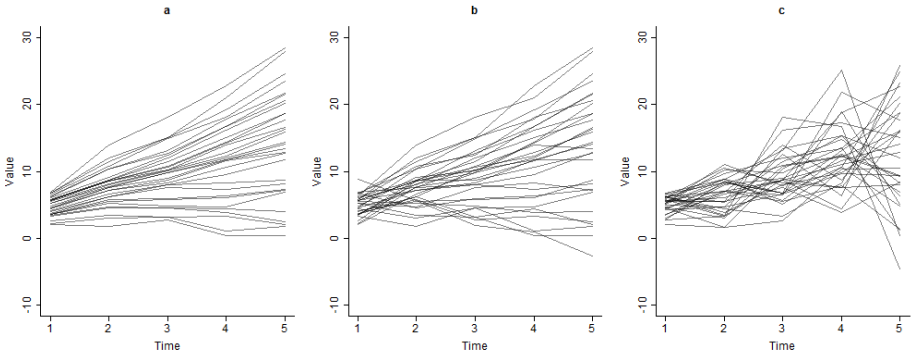


Figure 1. Example growth curve plots for (a) sorted data, (b) non-sorted data, and (c) randomly shuffled data. Each plot shares the same data values, only the order of the data at each time point has changed. As such, each plot has the same average intercept and average slope.

This method may be considered a form of a permutation test. Permutation tests are a class of tests for comparing a given test statistic to a distribution of these test statistics obtained from a random ordering of the data (Collingridge, 2013; Golland & Fischl, 2003; Pesarin & Salmaso, 2010; Theiler et al., 1992). This random ordering builds a test statistic under the distribution of a null-hypothesis that there is no natural order to the data. Any test statistic outside of a set α-level, based on the permutation distribution, is then considered to be highly unlikely given random chance and thus must contain some meaningful and non-random structure. Our method differs from traditional permutation methods in that we propose the use of only a single random shuffle. This is because of the bounds set by 0 and $\sigma^2_{u_1}$ for y_{sort}. Over multiple different parameterizations, we found that on a scale of 0 to $\sigma^2_{u_1}$ for y_{sort}, the distribution of multiple random y_{shuff} is small in comparison ($< 5\%$ of the overall space). As such, one random shuffle should give a good approximation of the distribution of multiple random y_{shuff}. However, should a researcher need more precision, taking an average of multiple random y_{shuff} values will give a more accurate result.

In order to gain an intuitive understanding of this algorithm, consider this analogy of an individual with messy hair who wants a new hair style from a barber. A customer (data) with messy hair walks into a barber shop and asks the barber (researcher) for a haircut fitting for said customer's natural hair

style. The barber accepts this request and beings work, but is unable to visually determine if the customer has naturally curly hair (variable slopes) or naturally straight hair (constant slopes) due to the current messy state of the customer's hair. The barber knows however, that a natural property of hair is that curly hair is naturally difficult to straighten and straight hair is naturally difficult to curl. So the barber first attempts to straighten (sort) the customer's hair and finds that the hair changed very little. The barber then attempts to curl (shuffle) the customer's hair and finds the customer's hair curled with ease and had changed much from its original messy state. Thus, the barber concludes that the customer had naturally curly hair as the messy state of the customer's hair was most easily and most dramatically changed by curling (i.e., a reliable variation in slopes was found because $|\mu_0 - \mu_{sort}| < |\mu_0 - \mu_{shuff}|$).

The remainder of this article is structured as follows: First, a simulation is presented of the proposed permutation method compared to using DIC values for determining the existence of slope variation. Then an application of this method to data from the National Longitudinal Study of Adolescent to Adult Health is presented. Finally this article concludes with a discussion regarding the proposed method's usefulness, an introduction to an analysis tool which facilitates the application of this method, limitations, and future directions.

2 A Simulation Study

2.1 Data Generation

In order to determine the effectiveness of the proposed data permutation method and to compare our method with a common model comparison procedure (i.e., DIC), a simulation study was conducted using the R programming language and OpenBUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009; R Core Team, 2013). Each simulation generated data from one of two models: model A or model B. Data simulated from models A and B were also used to study the effectiveness of DIC values relative to the proposed data permutation method.

Model A is a model including a random slope term and is parameterized as:

$$Y_{ij} = 5 + 2T_j + u_{0i} + u_{1i}T_j + \epsilon_{ij},$$

$$u_{0i} \sim Gaussian(0, 1),$$

$$u_{1i} \sim Gaussian(0, \sigma^2_{u_1}),$$

$$Cov(u_{0i}, u_{1i}, \epsilon_{ij}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma^2_{u_1} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Parameter values of 5 and 2 for β_0 and β_1, respectively, and $T_j = j - 1, j = 1, \ldots, 5$ were chosen as simple examples of positive linear growth. The variance of parameter u_{0i} was set to 1 for all simulated data sets. As the proposed permutation method is a test of random slopes and not random intercepts, the variance

of parameter u_{0i} is arbitrary. The covariance between parameters u_{0i} and u_{1i} was set to 0 as any covariance between u_{0i} and u_{1i} would necessitate variance of u_{1i}, thus increasing the effectiveness of the proposed permutation method[3]. The variance of the error term was held constant at 1 across all time points (Grimm & Widaman, 2010). Finally $\sigma^2_{u_1}$ was varied across simulations, $\sigma^2_{u_1} = .1, .2, \ldots, 2$. Model B is simply model A without a random slope term where $u_{1i} = 0$:

$$Y_{ij} = 5 + 2T_j + u_{0i} + \epsilon_{ij},$$

$$u_{0i} \sim Gaussian(0, 1),$$

$$u_{1i} \sim Gaussian(0, \sigma^2_{u_1}),$$

$$Cov(u_{0i}, \epsilon_{ij}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

For data generated from model A and B, $\sigma^2_{u_1} \in [0, .1, .2, \ldots, 2]$. This creates $\sigma^2_{u_1}$ (signal) to σ^2_ϵ (noise) ratios ranging between 0% and 200% across both models A and B. Thus, in total 21 data generating models were used.

The choice of specific values for this simulation are mostly arbitrary as u_1 is independent from β_0 and β_1, u_0, and ϵ_{ij}. Thus any value choices for these terms should have no effect on the validity of this method as the proposed method is a comparison of only the similarity of $\sigma^2_{u_1}$ estimated from the observed data to only $\sigma^2_{u_1}$ of the same data organized in such a way that minimizes σ^2_ϵ versus the same data organized in a way to increase σ^2_ϵ. Theoretically no other parameter values should influence our proposed method in the case that there is no covariance between random intercept and random slope terms.

2.2 Simulation Methods

For each round of simulation, $N \in [50, 200, 500]$ individuals data were simulated from each of the 21 data generating models, $\sigma^2_{u_1} \in [0, .1, \ldots, 2]$. Each round of simulation generated 1,000 instances giving a total of 63,000 (3 x 21 x 1000) data sets. Using Equation (2), Bayesian growth curves were fit to data generated by models A and B. Model A represents data which has individual slope variation and thus can be used to compute statistical power and type-II error rates. Similarly, model B represents data with no individual slope variation and

[3] A smaller simulation was conducted with data simulated from a model with a meaningful covariance between u_{0i} and u_{1i}. This smaller simulation showed an increase in both statistical power and specificity, and a decrease in type-I and type-II error rates. This increase made detection of random slope variation nearly perfect for all DIC and permutation methods as any covariance between u_{0i} and u_{1i} would imply meaningful variation of u_{1i} as covariance is conditional on variance. As such, this simulation is not reported.

thus can be used to compute specificity and type-I error rates[4]. Using DIC and the proposed data permutation method, guesses were made at each simulation step to determine if data were simulated from a process with fixed slope growth trajectory across individuals or with a growth trajectory whose slope varied per individual. These guesses were compared to known random effect structures to determine statistical power and specificity rates. Each model was run with 20,000 MCMC iterations and a burn-in period of 15,000 iterations using OpenBUGS and the R2OpenBUGS package in R (Lunn et al., 2009; Sturtz, Ligges, & Gelman, 2005). All models were checked for convergence with a Kolmogorov-Smirnov test (Brooks, Giudici, & Philippe, 2003). To ensure this method was not statistical package specific, we ran a similar simulation study using the MCMCglmm R package and found identical results (Hadfield, 2010).

For DIC comparison, two models were conducted at each simulation, one with a fixed slope growth trajectory across individuals and one with a growth trajectory whose slope varied per individual. If the model with a growth trajectory whose slope varied per individual had a DIC value 10 points lower than the model with a constant rate of change, data from this simulation were considered to have a growth trajectory whose slope varied per individual. Otherwise the simulated data for said simulation were considered to have a trajectory with constant rate of change across individuals. We compared two criterion for DIC selection: DIC > 10 and minimum DIC value(Spiegelhalter et al., 1998).

For data permutation comparison, at each simulation step a model with a growth trajectory whose slope varied per individual was run on the data for that simulation step and the average value of $\sigma^2_{u_1}$ was recorded. Data was then sorted by column in descending order and a second model was run on the sorted data, storing $\sigma^2_{u_1}$ for this model. Finally data were randomly shuffled per column and a third model was run on this shuffled data, again storing $\sigma^2_{u_1}$ for this model. The three $\sigma^2_{u_1}$ values were then compared using the proposed data permutation algorithm. We compared two criterion for our permutation method: only one shuffle and the average of 10 shuffles.

[4] For this simulation study, statistical power is defined as the proportion of simulations in which the proposed data permutation method determined the existence of meaningful slope variation when data was generated from a model that included a variable slope. Similarly, specificity is defined as the proportion of simulations in which the proposed data permutation method was unable to determine the existence of meaningful slope variation when data was generated from a model that did not include a variable slope. Type-I and type-II error rates are defined as the proportion of simulations in which the proposed data permutation method detected the existence of meaningful slope variation when data was generated from a model that did not include a variable slope, and the proportion of simulations in which the proposed data permutation method was unable to determine the existence of meaningful slope variation when data was generated from a model that included a variable slope, respectively.

2.3 Simulation Results

Table 1 shows the statistical power and specificity for both the DIC methods and the proposed data permutation algorithm for all sample sizes studied. For signal:noise ratios less that 1:1, DIC outperforms our proposed permutation method in terms of statistical power. However as sample size increases and/or signal:noise ratio increases these two methods quickly become equal in their statistical power. When comparing specificity, our proposed permutation method shows an improvement of approximately 10 percentage points over the DIC method across all sample sizes. Thus, in situations where signal:noise ratios are at least equal, our permutation method performs just as well as DIC based model comparison in terms of statistical power, but has a substantially reduced type-I error rate.

Table 1. Permutation Test vs. DIC Simulation Comparing Statistical Power and Specificity

	Statistical Power											
	DIC 10			DIC Min			Permutation Test			Permutation Test 10		
Effect:Error Ratio	N = 50	N = 200	N = 500	N = 50	N = 200	N = 500	N = 50	N = 200	N = 500	N = 50	N = 200	N = 500
1:10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
2:10	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
3:10	0%	0%	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%
4:10	0%	4%	5%	1%	7%	10%	0%	0%	3%	1%	4%	4%
5:10	45%	82%	88%	61%	88%	92%	14%	15%	20%	16%	19%	22%
6:10	86%	100%	100%	92%	97%	100%	22%	38%	41%	25%	44%	46%
7:10	100%	100%	100%	99%	100%	100%	34%	51%	63%	34%	58%	66%
8:10	100%	100%	100%	100%	100%	100%	81%	90%	84%	88%	93%	89%
9:10	100%	100%	100%	100%	100%	100%	85%	91%	92%	94%	99%	99%
10:10	100%	100%	100%	100%	100%	100%	92%	95%	97%	99%	100%	100%
11:10	100%	100%	100%	100%	100%	100%	97%	98%	100%	100%	100%	100%
12:10	100%	100%	100%	100%	100%	100%	98%	100%	100%	100%	100%	100%
13:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
14:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
15:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
16:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
17:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
18:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
19:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
20:10	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Specificity	89%	87%	93%	82%	83%	83%	99%	100%	100%	100%	100%	100%

Note. Results of a simulation study comparing statistical power and specificity for DIC and the proposed permutation testing algorithm across three sample sizes. Effect:error ratio is a measure of true population variance in slope to error variance added at each time point. Each percentage is based on 1000 simulations.

3 Application to Real Data

As an example of our proposed method on a real data set, a Bayesian growth curve modeling was conducted on a sample of 185 individuals (90 Male, 95 Female) from the

National Longitudinal Study of Adolescent to Adult Health (Add Health) who were age 17 and had reported drinking in the past 12 months[5]. At each wave of measurement, participants were asked "Think of all the times you have had a drink during the past 12 months. How many drinks did you usually have each time? (A "drink" is a glass of wine, a can of beer, a wine cooler, a shot glass of liquor, or a mixed drink.)." This value was recorded in 1994-95, 1996, 2001-02, and 2008. If a participant reported drinking more than 20 drinks, his/her data was dropped from this analysis to remove individuals who might have been excessive drinkers or may not have properly understood the question. The proposed data permutation method was then applied to this data in order to test for the presence of meaningful interindividual differences in intraindividual change across time in drinking behavior, table 2. All models used uninformative Poisson priors for all mean components and uninformative inverse gamma priors for all variance components.

Table 2. Bayesian Growth Curve Analysis of Add Health Drinking Behaviors

Parameter	Effect	Estimate	95% CI - Lower	95% CI - Upper
Intercept	Mean	5.06	4.61	5.50
	Variance	6.06	4.30	8.24
Slope	Mean	-0.10	-0.15	-0.04
	Variance	0.07	0.05	0.10
Permutation Test Results: No Significant Variance for Slope				

Note. Results of a Bayesian growth curve analysis of the average number of alcoholic drinks individuals reported drinking each time he/she drank alcohol. A permutation test showed no significant slope variation between individuals indicating a common downward trend across all individuals.

Significant fixed-effects for both the intercept and slope term were found for this model. At age 17, on average individuals reported drinking 5.06 alcoholic drinks with a standard deviation of 2.46. Each year after, individuals reported drinking 0.10 fewer drinks with a standard deviation of 0.26. When individuals reached age 31, on average they reported drinking 3.66 drinks. These results align with previous findings on alcohol consumption trajectories for the general population (Fillmore et al., 1991; Hartika et al., 1991). A permutation test found no meaningful interindividual differences in

[5] From the National Longitudinal Study of Adolescent to Adult Health website: This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (http://www.cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

intraindividual change across time in drinking behavior for these individuals. This indicates that the slopes of individuals' growth trajectories in alcohol use behavior did not reliably vary at the individual level, figure 2. That is, a single general downward trend is sufficient to describe how individuals drinking behaviors change across time, given that we can not reject the null-hypothesis that there is no variation between individuals in slope values. Although table 2 shows a 95% credible interval with positive values for the variance of the random slope term, this may be due to the boundary problem induced by utilizing gamma distributed priors used for the variance term. Additionally, the Effect:Error ratio for this data as assessed by our model was approximately 4:10. This indicates that our proposed method would have low statistical power in this case to pick up meaningful slope variation if it existed (as with using the DIC). As such these findings should be taken as only an example of our proposed method used on a real data set.

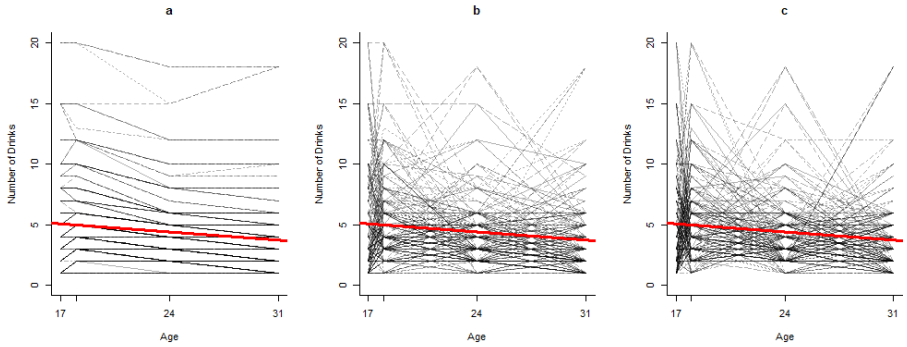


Figure 2. Permutation test for random slopes parameter of a Bayesian growth curve model modeling the average number of alcoholic drinks individuals reported drinking at each drinking occasion. The red/bold line is the result of each model. Plot (a) displays data in a sorted form. In this form the downward trajectory in drinking across time is evident. Plot (b) displays data in its original form. Plot (c) displays data in its shuffled form. Notice that (b) appears more similar to (c) than to (a), indicative of a random effect that may be indistinguishable from noise.

3.1 A Web Tool Implementation

In order to facilitate the use of our proposed data permutation method, we have developed a web application for Bayesian analyses of unconditional growth curve models. See Figure 3 for the interface of the web application. This web application incorporates our proposed data permutation method and is made available for free at `https://robertgm111.shinyapps.io/bayesiangrowthcurveapp/`.

This web tool was made to give researchers a simple to use interface for conducting Bayesian analyses of unconditional growth curve models. A researcher interested in using this tool would need to have data in a 3-column long format with column 1 being participant ID, column 2 being measurement occasion, and column 3 being the

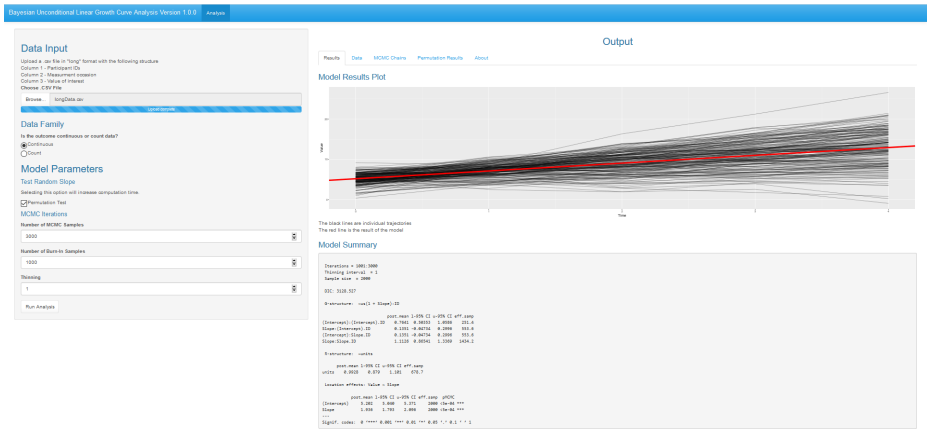


Figure 3. Screenshot of a web tool implementing the proposed data permutation test. This tool generates parameter estimates for unconditional bayesian growth curve models for data in the "long" format. Different tabs are available for model results, data viewing, MCMC chain veiwing, permutation test results, and citing information.

quantity of interest. Researchers can then select if the outcome is continuous (modeled by an uninformative normal prior) or count (modeled by an uninformative Poisson). Additionally researchers may select to run the proposed permutation test for the existence of variable slopes at the cost of increased computation time. After specifying the number of MCMC samples, burn-in period, and thinning, a researcher can obtain parameter estimates under the "Model Summary" heading and a plot of the data with a fitted line based on the model under the "Model Results Plot" heading. Additional tabs in this web tool allow for data viewing, MCMC chain viewing and download, results of the proposed permutation test (if selected), and citing information in an "About" section. Missing data points are sampled from posterior distributions during the MCMC updating process.

4 Discussion

The data permutation method shown in this article is a simple to use and widely applicable method for testing for the existence of interindividual differences in intraindividual change across time when these differences are modeled as gamma distributed variance components. Although itself not a Bayesian derived test, our method was able to perform on par with the DIC metric for most cases. Unlike more complicated methods such as DIC calculation and Bayes factors our permutation method requires little technical ability to implement, save for initial model specification. If a researcher is analysing data with a signal:noise ratio that is at least 1:1 then our method preforms just as well as common DIC comparison methods in terms of statistical power and outperforms DIC in terms of specificity. We do not believe this is an unreasonable ratio for many areas of psychological/behavioral sciences (Cooper & Findley, 1982; Wilson & Sherrell, 1993). Although other methods have been proposed for testing the existence

of meaningful random slope variation, our proposed method is simple to use and we offer a direct software implementation (Saville & Herring, 2009).

Beyond ease of use, the permutation method displayed in this article represents an alternative method for model comparison in a Bayesian framework that is data driven. Many methods such as DIC and Bayes factors are manipulations of a model such that plausible models are pitted against one another so as to determine a model best fitting to a given data set. In such a model comparison framework, a given model is typically compared to a constrained version of itself (Kruschke, 2011; Spiegelhalter et al., 2002). These constrains represent a researcher's qualities of interest, or unique hypotheses, regarding a specific data set. As opposed to constraining a specific parameter to test for the existence of a specific effect, our data driven method targets a quality if interest within the data itself. Instead of comparing a model with a given effect to a model without a given effect, our permutation method compares an estimated parameter (slope variation) from a given data set with the same parameter from both a modified data set in which this parameter has been destroyed and a second modified data set in which this parameter was amplified. That is, while model comparison asks "Which model was more likely to generate this data?", our proposed permutation method asks "Is the parameter I am interested in modeling in this data different from data in which this parameter is just noise?". Framing hypothesis testing in this manner is then a stepping stone to further data driven analyses in which a targeted permutation method is used to study a specific quality of interest.

4.1 Limitations

Firstly, the support for our proposed method comes from our simulation study. Although we have attempted to model realistic circumstances given our specific random effects structure, our results can not be generalized outside of simulated parameterizations. Future work should focus on understanding the analytical properties of our test given that our test works on a bounded classification framework. This includes extending the results of this simulation to more time points, however we see no reason this method would not work on more than four time points.

Although simulation showed our method to have exceptional statistical power and specificity under conditions of relatively equal signal:noise ratios, there are still limitations to this method. One such limitation is that our proposed method showed inadequate statistical power of signal:noise ratios of 7:10 or less. Thus, our proposed method should not be used in situations in which variation in individual slopes is substantially less than error variance. In such a case DIC based model comparison is more appropriate. However, we believe that most longitudinal studies will easily be able to exceed this threshold, reducing the impact of this limitation. In situations in which significant covariance exists between intercept and slope values, our proposed method performs as well as DIC based model comparison. This is due to the necessity of the existence of slope variation prior to the existence of covariance between intercepts and slopes. In many realistic data sets, if significant slope exists then a significant covariance between intercept and slope is also likely to exist. This is due to ceiling effects, floor effects, regression to the mean, and other phenomenon common in behavioural data.

Our proposed method may also be limited in its usefulness beyond testing for significant slope variation. That is, our proposed method capitalizes on the fact that sorting data and shuffling data preserves intercept values and only changes error variance about slope estimates. Due to this capitalization, this permutation method is

not applicable for testing for the existence of meaningful intercept variation and more research is needed to discover such a test. In practice however, researchers interested in longitudinal processes are generally more concerned with slope parameters as slope parameters represent change over time.

Additionally, this method only works for cases in which all participants have been sampled at the same discrete bins. That is, this method is not applicable for continuous sampling designs (Bolger & Laurenceau, 2013) In this case, alternative sorting and shuffling strategies must be employed so as to maintain the same structural changes int he data as would have occurred if the data was in discretely sampled bins at from the same time points. This also extends to cases of missing data. Missing data is common in longitudinal research and must be expected to occur more in studies over longer periods of time. In this case, multiple imputation may be used as a method for creating multiple possible tests using our algorithm. The most selected state (i.e., random effect or no random effect) across these imputations would then be chose as the best state to describe the data given the model.

4.2 Future Directions

One possible extension of the proposed data permutation method would be to test for nonlinear effects. Growth curve models are not limited to modeling solely linear growth, but may be extended to model curves of higher order polynomials (McArdle & Nesselroade, 2003). We do not see any reason for permutation testing to be ineffective for polynomial growth curve models, however this testing should still be conducted for purposes of understanding statistical power and specificity.

We also note the usefulness of plots of sorted data for understanding trajectories over time. Figures 1(a) and 2(a) show sorted data compared to original data in figures 1(b) and 2(b). Any linear trend is easier to visualize in the sorted data as opposed its associated original data. We attempted this same plotting method with non-linear effects as well and ac hived a similar ease of trend visualization, as sorting preserves intercept and slope values. Future research may specifically look at data sorting as a viable means of plotting data for model selection in growth curve analysis.

Other measures of distributional qualities besides the mean may also increase the power of our proposed method to detect significant slope variation across individuals. We conducted a relatively small simulation study testing the efficacy of using median estimates above mean estimates and obtained similar results to using means. Other metrics may prove to be more useful however, and should be tested in order to further refine our proposed data permutation method.

Another possible future direction would be to continue to create permutation tests targeting specific parameters of interest. According to Wolpert and Macready (1997), no single method for optimization of a problem is the best possible method for solving all problems. According to this No Free Lunch Theorem, the better a single optimizer gets at solving a specific problem, the worse it gets at solving all other problems. This suggests two things. Firstly, for every global method for optimizing a problem (e.g., DIC based model comparison), there exists a more targeted method that will yield a more optimal solution to a problem. Secondly, every problem may have its own "best" solution. That is, every problem that is attempted to be optimized, may have its own best, and targeted, way to be optimized. While this second point implies that perhaps researchers should find targeted methodology for every possible effect in which they are interested, this would quickly spiral into many tests and would most likely create more confusion for individuals wishing to test specific effects.

Although targeted, our proposed method is easy to implement and solves the boundary problem for testing gamma/inverse gamma distributed random effects. This ease of implementation will allow more researchers to test for significant individual differences in intraindividual change. Additionally, our method offers one of the first steps for a paradigm shift of model comparison in a Bayesian framework. One where data is modified to destroy qualities of interest, as opposed to models being formed with/without qualities of interest. Indeed there may in fact be a hybrid form of these two methods that may prove more viable than either method in isolation. We hope our proposed permutation method spurs other researchers to consider data modifications for testing individual effects, leading to relatively uncomplicated methods that other researchers may use for testing whatever effects in which he/she is interested.

References

Baldwin, S. A., & Hoffmann, J. P. (2002). The Dynamics of Self-Esteem: A Growth-Curve Analysis. *Journal of Youth and Adolescence*, *31*(2), 101–113. doi: https://doi.org/10.1023/A:1014065825598

Baltes, P. B., & Nesselroade, J. R. (1979). *Longitudinal research in the study of behavior and development*. Academic Press New York, NY.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, *4*(December), 328. doi: https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. doi: https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*, 1–27. doi: https://doi.org/arXiv:1506.04967

Bayarri, M. J., & Berger, J. O. (2004). The interplay of bayesian and frequentist analysis. *Statistical Science*, 58–80. doi: https://doi.org/10.1214/088342304000000116

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Brooks, S. P., Giudici, P., & Philippe, A. (2003). Nonparametric convergence assessment for MCMC model selection. *Journal of Computational and Graphical Statistics*, *12*(1), 1–22. doi: https://doi.org/10.1198/1061860031347

Cai, B., & Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics*, *62*(2), 446–457. doi: https://doi.org/10.1111/j.1541-0420.2005.00499.x

Carlin, B. P., & Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(3), 473–484. doi: https://doi.org/10.2307/2346151

Chen, Z., & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, *59*(4), 762–769.

Collingridge, D. S. (2013). A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, *7*(1), 81–97. doi: https://doi.org/10.1177/1558689812454457

Cooper, H., & Findley, M. (1982). Expected Effect Sizes. *Personality and Social Psychology Bulletin*, *8*(1), 168–173. doi: https://doi.org/10.1177/014616728281026

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1), 1–26. doi: https://doi.org/10.1214/aos/1176344552

Fan, Y., & Li, R. (2012). Variable selection in linear mixed effects models. *Annals of Statistics*, *40*(4), 2043–2068. doi: https://doi.org/10.1214/12-AOS1028

Fearn, T. (1975). A Bayesian Approach to Growth Curves. *Biometrika*, *62*(1), 89. doi: https://doi.org/10.2307/2334490

Fillmore, K. M., Hartika, E., Johnstone, B. M., Leino, E. V., Motoyoshi, M., & Temple, M. T. (1991). A meta-analysis of life course variation in drinking. *British Journal of Addiction*, *86*(10), 1221–1268. doi: https://doi.org/10.1111/j.1360-0443.1991.tb01702.x

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. doi: https://doi.org/10.1214/06-BA117A

Gilks, W. R., Wang, C. C., Yvonnet, B., & Coursaget, P. (1993). Random-Effects Models for Longitudinal Data Using Gibbs Sampling. *Biometrics*, *49*(2), 441. doi: https://doi.org/10.2307/2532557

Golland, P., & Fischl, B. (2003). Permutation tests for classification: towards statistical significance in image-based studies. In *Biennial international conference on information processing in medical imaging* (pp. 330–341).

Gomez-Baya, D., Mendoza, R., Paino, S., Sanchez, A., & Romero, N. (2016). Latent growth curve analysis of gender differences in response styles and depressive symptoms during mid-adolescence. *Cognitive Therapy and Research*, 1–15. doi: https://doi.org/10.1007/s10608-016-9822-9

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications New York, NY.

Grimm, K. J., & Widaman, K. F. (2010). Residual structures in latent growth curve modeling. *Structural Equation Modeling*, *17*(3), 424–442. doi: https://doi.org/10.1080/10705511.2010.489006

Gutman, L. M., Sameroff, A. J., & Cole, R. (2003). Academic growth curve trajectories from 1st grade to 12th grade: Effects of multiple social risk factors and preschool child factors. *Developmental Psychology*, *39*(4), 777–790. doi: https://doi.org/10.1037/0012-1649.39.4.777

Hadfield, J. D. (2010). MCMCglmm: MCMC Methods for Multi-Response GLMMs in R. *Journal of Statistical Software*, *33*(2), 1–22. doi: https://doi.org/10.1002/ana.22635

Hartika, E., Johnstone, B., Leino, E. V., Motoyoshi, M., Temple, M. T., & Fill-

more, K. M. (1991). A meta-analysis of depressive symptomatology and alcohol consumption over time. *British Journal of Addiction*, *86*(10), 1283–1298. doi: https://doi.org/10.1111/j.1360-0443.1991.tb01704.x

Hertzog, C., & Nesselroade, J. R. (2003). Assessing Psychological Change in Adulthood: An Overview of Methodological Issues. *Psychology and Aging*, *18*(4), 639–657. doi: https://doi.org/10.1037/0882-7974.18.4.639

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Bayes Factors Have Frequency Properties—This Should Not Be Ignored: A Rejoinder to Morey, Wagenmakers, and Rouder. *Multivariate Behavioral Research*, *51*(1), 20–22. doi: https://doi.org/10.1080/00273171.2015.1071705

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Jounal of the american statistical association*, *90*(430), 773–795. doi: https://doi.org/10.1080/01621459.1995.10476572

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. doi: https://doi.org/10.1177/1745691611406925

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E. J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331–347. doi: https://doi.org/10.1016/j.jmp.2011.06.001

Lunn, D., Jackson, C., Best, N., Spiegelhalter, D., & Thomas, A. (2012). *The bugs book: A practical introduction to bayesian analysis*. Chapman and Hall/CRC.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*(25), 3049–3067. doi: https://doi.org/10.1002/sim.3680

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*(2013), 305–315. doi: https://doi.org/10.1016/j.jml.2017.01.001

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In *Handbook of psychology*. John Wiley & Sons, Inc. doi: https://doi.org/10.1002/0471264385.wei0218

Morey, R. D., Wagenmakers, E. J., & Rouder, J. N. (2016). Calibrated Bayes Factors Should Not Be Used: A Reply to Hoijtink, van Kooten, and Hulsker. *Multivariate Behavioral Research*, *51*(1), 11–19. doi: https://doi.org/10.1080/00273171.2015.1052710

Moulder, R. G., Boker, S. M., Ramseyer, F., & Tschacher, W. (2018). Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses. *Psychological Methods*. doi: https://doi.org/10.1037/met0000172

Oravecz, Z., & Muth, C. (2018). Fitting growth curve models in the Bayesian framework. *Psychonomic Bulletin and Review*, *25*(1), 235–255. doi: https://doi.org/10.3758/s13423-017-1281-0

Peng, H., & Lu, Y. (2012). Model selection in linear mixed ef-

fect models. *Journal of Multivariate Analysis*, *109*, 109–129. doi: https://doi.org/10.1016/j.jmva.2012.02.005

Pesarin, F., & Salmaso, L. (2010). The permutation testing approach: a review. *Statistica*, *70*(4), 481–509. doi: https://doi.org/10.6092/issn.1973-2201/3599

Piironen, J., & Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. doi: https://doi.org/10.1007/s11222-016-9649-y

R Core Team. (2013). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, *132*(1), 1. doi: https://doi.org/10.1037/0033-2909.132.1.1

Rouder, J. N., & Morey, R. D. (2012). Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *47*(6), 877–903. doi: https://doi.org/10.1080/00273171.2012.734737

Saville, B. R., & Herring, A. H. (2009). Testing Random Effects in the Linear Mixed Model Using Approximate Bayes Factors. *Biometrics*, *65*(2), 369–376. doi: https://doi.org/10.1111/j.1541-0420.2008.01107.x

Shedler, J., & Block, J. (1990). Adolescent drug use and psychological health: A longitudinal inquiry. *American psychologist*, *45*(5), 612. doi: https://doi.org/10.1037//0003-066X.45.5.612

Silver, R. C., Holman, E. A., McIntosh, D. N., Poulin, M., & Gil-Rivas, V. (2002). Nationwide longitudinal study of psychological responses to september 11. *Jama*, *288*(10), 1235–1244. doi: https://doi.org/10.1001/jama.288.10.1235

Spiegelhalter, D. J., Best, N. G., & Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters , and the comparison of arbitrarily complex models. *Technical Report, MRC Biostatistics Unit, Cambridge, UK*.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian Measures of Model Complexity anf Fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *64*(4), 583–639. doi: https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., et al. (2014). The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B*, *64*, 485–493. doi: https://doi.org/10.1111/rssb.12062

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on bayes factor design analysis using an informed prior. *Behavior research methods*, *51*(3), 1042–1058. doi: https://doi.org/10.3758/s13428-018-01189-8

Stram, D. O., & Lee, J. W. (1994, dec). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, *50*(4), 1171. doi: https://doi.org/10.2307/2533455

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2OpenBUGS: A Package for Running OpenBUGS from R. *Journal of Statistical Software*, *12*(3), 1–16. doi: https://doi.org/10.18637/jss.v012.i03

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., & Farmer, J. D. (1992). Testing for nonlinarity in time series: the method of surrogate data. *Physica D*, *58*(1-4), 77–94. doi: https://doi.org/10.1016/0167-2789(92)90102-S

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, *92*(2), 351–370. doi: https://doi.org/10.1093/biomet/92.2.351

Ward, E. J. (2008). A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. *Ecological Modelling*, *211*(1-2), 1–10. doi: https://doi.org/10.1016/j.ecolmodel.2007.10.030

Wilson, E. J., & Sherrell, D. L. (1993). Source Effects in Communication and Persuasion Research:. *Journal of the Academy of Marketing Science*, *21*(2), 101. doi: https://doi.org/10.1177/009207039302100202

Wolpert, D., & Macready, W. (1997, apr). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. Retrieved from `http://ieeexplore.ieee.org/document/585893/` doi: https://doi.org/10.1109/4235.585893

Zhang, Z. (2016). Modeling error distributions of growth curve models through Bayesian methods. *Behavior Research Methods*, *48*(2), 427–444. doi: https://doi.org/10.3758/s13428-015-0589-9

Zhang, Z., Hamagami, F., Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*(4), 374–383. doi: https://doi.org/10.1177/0165025407077764

Journal of Behavioral Data Science, 2025, *5* (1), 23–50.
DOI: https://doi.org/10.35566/jbds/jinliu

Extending Latent Basis Growth Model to Explore Joint Development in the Framework of Individual Measurement Occasions

Jin Liu[1 [0000−0001−5922−6643]]

Data Sciences Institute, Takeda Pharmaceuticals, Boston, MA 02142, USA
Veronica.Liu0206@gmail.com

Abstract. Longitudinal processes often exhibit nonlinear change patterns. Latent basis growth models (LBGMs) provide a versatile solution without requiring specific functional forms. Building on the LBGM specification for unequally-spaced waves and individual measurement occasions proposed by Liu and Perera (2024), we extend LBGMs to multivariate longitudinal outcomes. The extended models enable the analysis of nonlinear parallel longitudinal processes with unequally-spaced study waves in the framework of individual measurement occasions. We present the proposed models by simulation studies and real-world data analyses. Simulation studies demonstrate that the proposed model can provide unbiased and accurate estimates with target coverage probabilities for the parameters of interest. Real-world analyses of reading and mathematics scores demonstrate its effectiveness in analyzing joint developmental processes that vary in temporal patterns. Computational code is included.

Keywords: Latent Basis Growth Model · Parallel Nonlinear Longitudinal Processes · Individual Measurement Occasions · Simulation Studies

1 Introduction

In longitudinal studies, researchers often gather measurements on multiple outcomes to decipher how each evolves over time. While the focus has traditionally been on univariate outcomes, the inter-correlated nature of processes in domains such as development (Liu & Perera, 2022; Peralta, Kohli, Lock, & Davison, 2022; Shin, Davison, Long, Chan, & Heistad, 2013), behavioral sciences (Duncan & Duncan, 1994, 1996), and biomedicine (Dumenci et al., 2019) demands a multifaceted analysis. Recent research reflects a growing interest in exploring how these interconnected outcomes influence one another over time. Developmental studies, for example, often track achievement scores across multiple subjects (Liu & Perera, 2022, 2023; Peralta et al., 2022; Shin et al., 2013), facilitating an in-depth analysis of correlated growth in multiple domains. Similarly, clinical trials might collect multiple endpoints (Dumenci et al., 2019) to provide a

holistic evaluation of treatment effects. This complexity underscores the need for advanced modeling techniques that accurately capture the correlation between multiple longitudinal processes.

Another scenario highlighting the complexity of longitudinal studies involves the reconciliation of data from diverse sources. For instance, observational studies may utilize both child and parent reports to assess a child's health-related quality of life (Rajmil, López, López-Aguilà, & Alonso, 2013). In clinical trials, a single endpoint is often measured using different equipments, adding complexity to data interpretation. Additionally, analyzing repeated outcomes from individuals nested within pairs or small groups (Lyons et al., 2017; McNulty, Wenner, & Fisher, 2016) presents unique statistical challenges. These situations underscore the need for a model capable of describing the joint longitudinal processes, with the aim of elucidating the associations between varied data sources and outcomes. The objective of our study is to develop such models within the Structural Equation Modeling (SEM) framework, as SEM provides a flexible and comprehensive approach for capturing complex relationships and dependencies between variables.

Research in developmental psychology has provided insights into the joint development of cognitive abilities, such as the studies by Robitaille, Muniz, Piccinin, Johansson, and Hofer (2012), revealed complex nonlinear intercept and slope associations in the progression of visuospatial ability and processing speed, using multivariate growth models (MGMs) with linear growth curves. However, a model with linear function often falls short in capturing the full complexity of real-world longitudinal processes, which frequently exhibit nonlinearity and thus necessitate more sophisticated analytical approaches. To better model such complexity, Blozis (2004) developed MGMs with nonlinear parametric functions, such as polynomial and exponential forms, to capture nonlinear parallel growth, which were implemented using LISREL in Blozis, Harring, and Mels (2008). Despite these advancements, parametric models with predetermined nonlinear functional forms may not adequately represent actual change patterns that do not conform to the prespecified functional forms. Furthermore, while these MGMs can estimate correlations between growth factors, such as intercept-intercept and linear/quadratic slope-slope relationships for quadratic functions, they often fail to provide insights into the relationship between two nonlinear longitudinal processes directly.

As the field has evolved, there has been a notable shift toward semi-parametric methods to allow for more flexible analysis in multivariate nonlinear longitudinal processes. Liu and Perera (2022) exemplified this shift with their linear-linear piecewise function within the SEM framework. A longitudinal model with a linear-linear piecewise function divides the growth trajectory into two linear segments, each with its own slope, joined at a specific point (i.e., the knot). By breaking down the growth curve into two distinct phases, the model allows for the assessment of slope-slope correlation at each stage. More importantly, by estimating the knots and their variances, the model also examines knot-knot correlations. This approach provides a unique perspective on the developmental

process by helping to understand how the correlation changes in each stage and when the transition from one stage to the next occurs.

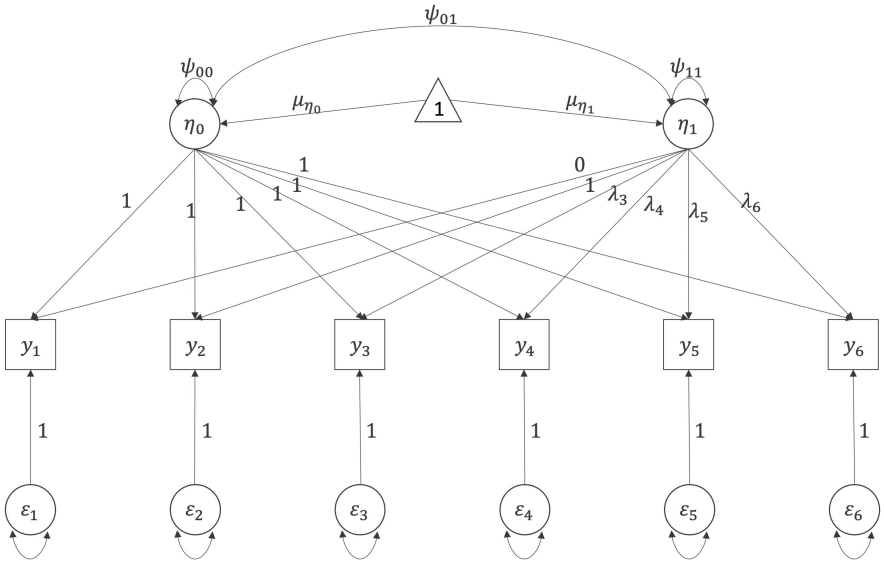
However, while semi-parametric functional forms like those introduced by Liu and Perera (2022) significantly enhance modeling flexibility, they inherently impose constraints by limiting the change patterns to only two distinct phases. Such two-piece functional forms may not adequately capture more complex developmental patterns that exhibit multiple phases over time, particularly in exploratory research stages where the underlying change patterns are not well-defined. Herein lies the advantage of latent basis growth models (LBGMs), which provide greater flexibility by allowing for the determination of the optimal curve shape without the constraints of prior assumptions (McArdle & Epstein, 1987; Meredith & Tisak, 1990). Our work builds on this flexibility to facilitate explorations of multiple longitudinal processes, addressing the need for adaptable analytical tools capable of handling the complexities of real-world challenges.

1.1 Traditional Specification of Latent Basis Growth Model

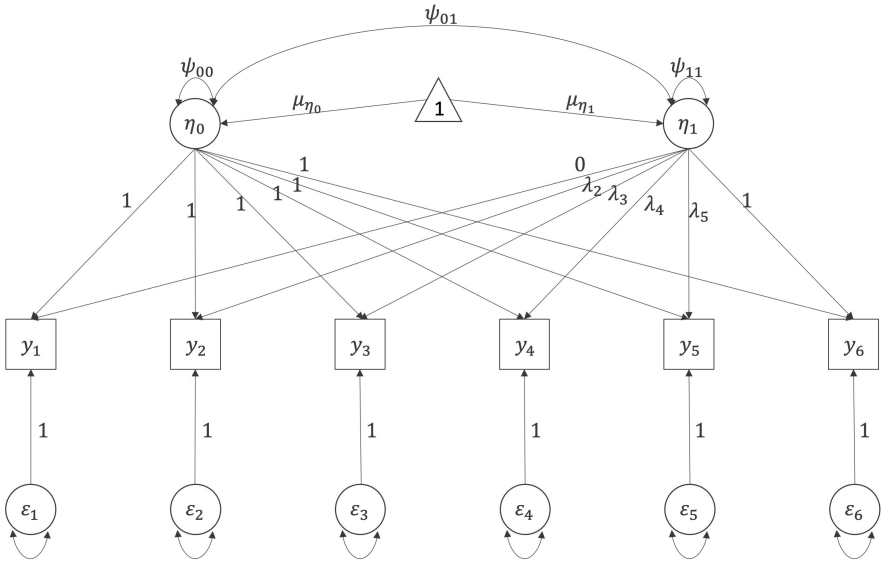
Grimm, Ram, and Estabrook (2016, Chapter 11) demonstrated that LBGMs can be constructed using both the Latent Growth Curve Modeling (LGCM) framework, a subset of the SEM framework, and the mixed-effects modeling framework. While LBGMs were not explicitly discussed, existing literature suggests that, for a majority of longitudinal models, these two frameworks are mathematically equivalent in evaluating between-individual differences in within-individual changes (Bauer, 2003; Curran, 2003). This study focuses on the SEM framework due to its greater modeling flexibility and widespread recognition within the social science research community.

Similar to other latent growth curve models, a LBGM can be expressed as $\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i$, where $\boldsymbol{y}_i$ represents the vector of repeated measurements for individual i, $\boldsymbol{\eta}_i$ is the vector of latent growth factors for individual i, $\boldsymbol{\Lambda}$ is the matrix of factor loadings, and $\boldsymbol{\epsilon}_i$ is the residual vector of individual i. Simply put, this equation captures how an individual's change patterns are represented by latent growth factors and measurement occasions. LBGMs typically consist of two growth factors: an intercept and a shape factor.

The factor loading matrix $\boldsymbol{\Lambda}$ is partially constrained for model identification. Specifically, in a setting with J measurements, factor loadings for the intercept are fixed at 1, while two factor loadings for the shape factor are also fixed, and the remaining $J-2$ are estimated. Figures 1a and 1b illustrate two common specifications of LBGM with six repeated measurements. In Figure 1a, the shape factor is scaled as the change during the initial time interval. In Figure 1b, the shape factor is scaled as the total change over the study duration. These methods allow for the flexible estimation of $\boldsymbol{\Lambda}$, thus freeing LBGM from being restricted to a specific functional form. This flexibility in specification allows LBGMs to adapt to different research questions and datasets, making them a powerful tool for longitudinal data analysis.



(a) Specification 1



(b) Specification 2

Figure 1: Path Diagram of Traditional Latent Basis Growth Models

1.2 Novel Specification of Latent Basis Growth Model

Although the LBGM described in Section 1.1 is a flexible statistical tool to explore trajectories, it does not specify whether nonlinearity exists in the growth patterns (Grimm, Steele, Ram, & Nesselroade, 2013) nor does it detail the nature of such nonlinearity (Wood, Steinley, & Jackson, 2015); it still has limitations. According to Grimm et al. (2016, Chapter 11), discrete time points are required when specifying an LBGM, and therefore, it cannot be fit in the framework of individual measurement occasions. One approximate method for such continuous measurement time is the time-bins approach, also known as the time-windows method. The time-bins approach involves dividing the assessment period into several intervals (time-bins), where each individual can have up to one response per bin. If a subject does not contribute data to a specific time window, it is treated as a missing record (Sterba, 2014).

However, several studies highlight the limitations of this approach. For example, Blozis and Cho (2008) demonstrated that using the time-bins approach may lead to inadmissible estimation, such as overestimating within-individual changes and underestimating between-individual differences, though these effects can be negligible if individual differences are not substantial. Moreover, Coulombe, Selig, and Delaney (2015) concluded that neglecting time differences often leads to undesirable outcomes, such as biased parameter estimates. Their evaluation of bias, efficiency, and Type I error rate under various conditions—different combinations of sample size, degree of heterogeneity, distribution of time, rate of change, and number of repeated measurements—showed that ignoring time differences can significantly affect the results.

Two parallel but distinct methods for accounting for individual measurement occasions have been developed by Sterba (2014) and Liu and Perera (2024). Sterba (2014) introduces an innovative approach by incorporating two growth factors—the intercept and shape factor. This model defines the loadings of the shape factor as a function of the specific timing of each individual's measurements, accounting for deviations from a linear progression.

In contrast, the framework by Liu and Perera (2024) specifies the latent basis growth model by incorporating linear piecewise functional forms, which effectively capture the dynamics across J measurements segmented into $J-1$ intervals. This model is designed to estimate interval-specific slopes and allows for an extension to derive both interval-specific changes and changes from the baseline. In particular, as illustrated in Figure 2a, the interval-specific change is quantified using the area under the curve (AUC) for the corresponding time interval, effectively representing the integral of the growth rate over that period. For example, consider the change from $t = 1$ to $t = 2$ calculated as: $0.8 \times (2-1) = 0.8$. This calculation is depicted in Figure 2b, where the change in growth is shown to increase by 0.8, from 21 to 21.8. Using AUC to represent interval-specific change relaxes the traditional constraints of LBGMs and allows for unequally-spaced study waves. For example, in Figure 2a, even if no measurement is taken at $t = 5$, the change from $t = 4$ to $t = 6$ can still be calculated

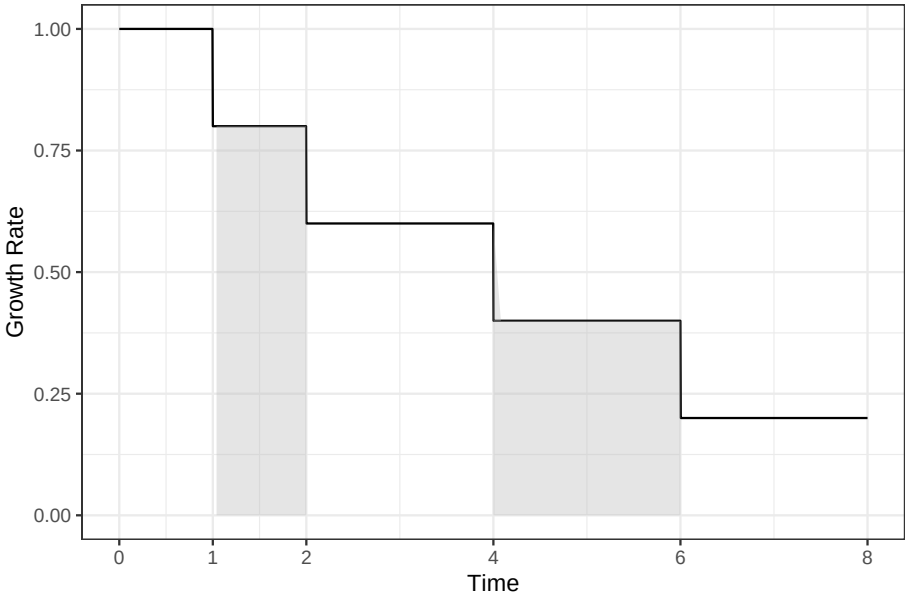
as $0.4 \times (6-4) = 0.8$. Similarly, the change from the baseline at any specific time is quantified using the AUC from the baseline to that particular time point.

The path diagram of the LBGM with six measurement occasions, as proposed by Liu and Perera (2024), is illustrated in Figure 3a. This model features two growth factors: the initial status, denoted as η_0, and the slope during the first interval, denoted as η_1. As depicted in Figure 3a, η_1 along with the relative rate γ_{j-1}, defines the interval-specific slopes (dy_{ij}). These slopes are then utilized, along with the length of each interval, to derive interval-specific changes. Each interval is enclosed in a diamond shape in the diagram, indicating that these intervals are allowed to vary among individuals. Such flexibility addresses the challenge of individual measurement occasions (which further lead to individual intervals), thus providing a more accurate representation of their growth trajectories. These time intervals are, therefore, considered 'definition variables', allowing the model to account for individual differences (Mehta & Neale, 2005; Mehta & West, 2000; Sterba, 2014).

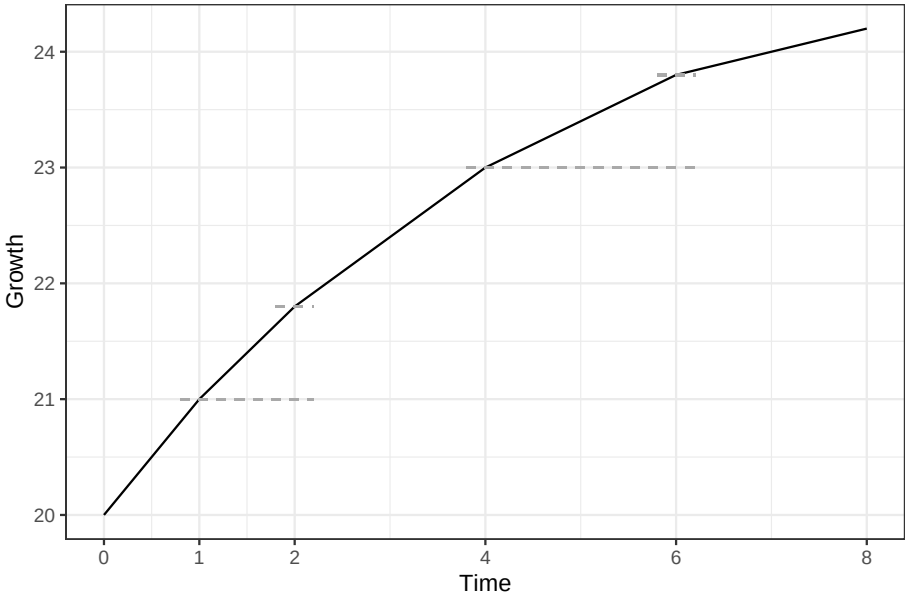
In addition to allowing for unequally-spaced study waves and individual measurement occasions in LBGM, this framework provides flexibility in scaling the growth rate factor, η_1. Instead of constraining η_1 to the first time interval, it can be adapted to represent growth rate during any selected time frame, such as the last time interval, as demonstrated in Figure 3b. Here, γ_{j-1} still serve as the relative growth rate in relation to η_1 for each $(j-1)^{th}$ time interval. Note that the models with different scalings of η_1 are mathematically equivalent. With such novel specifications, the shape factor's loading at each measurement occasion t_j is calculated by dividing the change-from-baseline (the difference between the current value and the initial value at t_1) at t_j by η_1. The setup of such factor loadings will be further explained in Section 2.1.

1.3 Parallel Latent Basis Growth Model

In the study of joint longitudinal processes, researchers frequently utilize MGMs, also known as parallel process and correlated growth models, which are thoroughly discussed in Grimm et al. (2016, Chapter 8) and McArdle (1988). MGMs are generally utilized to estimate three main types of associations based on the interactions they analyze: (1) within-process growth factors, (2) between-process growth factors, and (3) between-process residuals. Existing research, including studies by Robitaille et al. (2012), who investigated the co-evolution of processing speed and visuospatial ability using linear growth curves, and Blozis (2004); Blozis et al. (2008), who incorporated parametric nonlinear functional forms like polynomial and exponential curves, has significantly contributed to the understanding of these relationships. More recently, Peralta et al. (2022) and Liu and Perera (2022) have advanced this area by developing MGMs with linear-linear functional forms with unknown random knots, in the Bayesian mixed-effects and frequentist structural equation frameworks, respectively. Although effective for theory-driven research, these models sometimes lack the flexibility needed during the exploratory phases of research, especially in the absence of a guiding domain-specific theory for functional form selection.

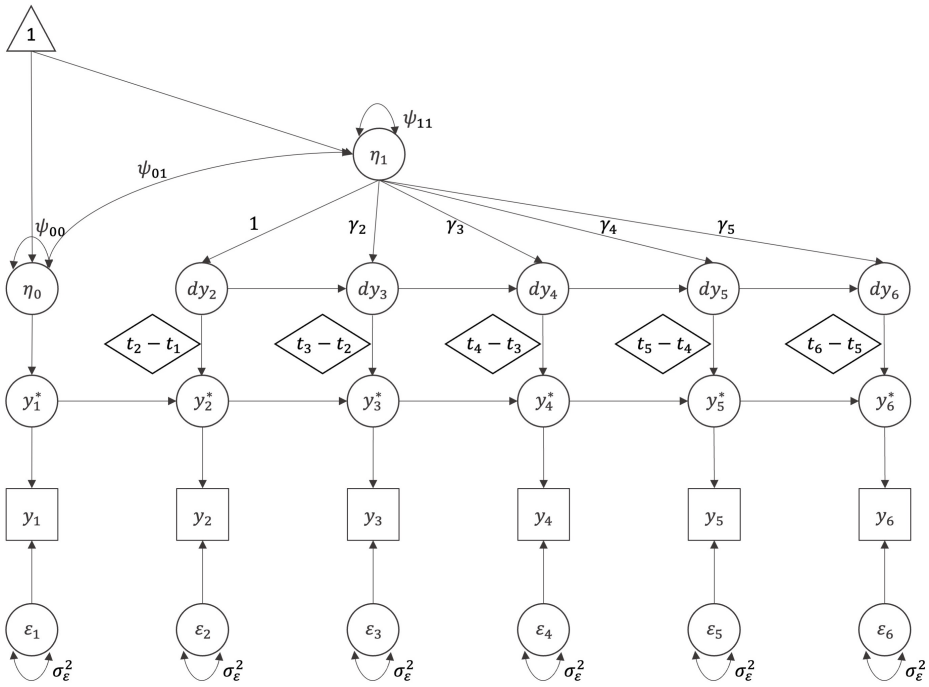


(a) Piecewise Linear Growth Rate

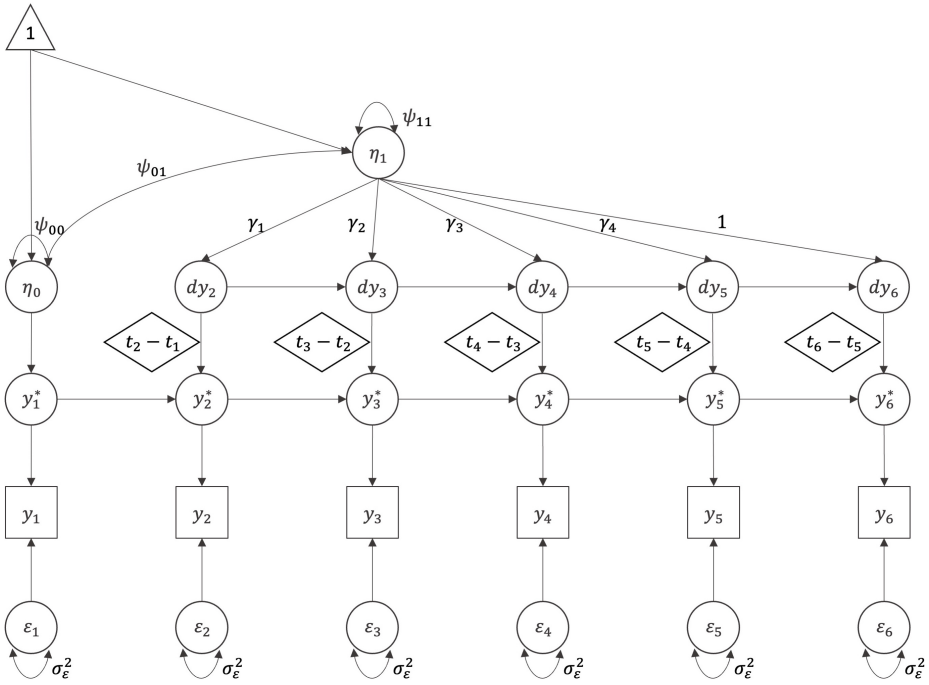


(b) Piecewise Linear Growth Curve

Figure 2: Piecewise Linear Growth Curve and Growth Rate (Values of the Intercept and Slope of Each Time Interval: $\eta_0 = 20$; $\gamma_1 = 1.0$; $\gamma_2 = 0.8$; $\gamma_3 = 0.6$; $\gamma_4 = 0.4$; $\gamma_5 = 0.2$)



(a) Specification 1



(b) Specification 2

Figure 3: Path Diagram of Novel Latent Basis Growth Models

This article aims to advance the field of joint longitudinal process modeling by extending the LBGM with the novel specification detailed in Section 1.2 to the MGM framework. The proposed model addresses existing gaps by demonstrating how to implement a parallel LBGM tailored to unequally-spaced study waves and individual measurement occasions. The structure of this article is organized as follows: We begin with a description of a LBGM for a univariate longitudinal process, incorporating our novel specification in the methods section. This model is then extended to a parallel growth curve framework, where we detail the model specification and estimation procedures. Subsequently, we evaluate the model's performance using a Monte Carlo simulation study, focusing on metrics such as relative bias, empirical standard error (SE), relative root-mean-square error (RMSE), and the coverage probability (CP) of a 95% confidence interval. We also illustrate the practical application of our model by analyzing a real-world dataset of longitudinal reading and mathematics scores from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011). In the application section, we explore how to derive and interpret insights from the model output. Finally, we conclude with discussions on practical and methodological considerations and directions for future research.

2 Method

2.1 Latent Basis Growth Model in the Framework of Individual Measurement Occasions

This section introduces the novel LBGM specification developed by Liu and Perera (2024) for univariate nonlinear developmental trajectories, applicable to analyzing univariate longitudinal outcomes such as reading or mathematics development. For individual i, the model can be specified as

$$y_{ij} = y^*_{ij} + \epsilon^{[y]}_{ij}, \tag{1}$$

$$y^*_{ij} = \begin{cases} \eta^{[y]}_{0i}, & \text{if } j = 1 \\ y^*_{i(j-1)} + dy_{ij} \times (t_{ij} - t_{i(j-1)}), & \text{if } j = 2, \ldots, J \end{cases}, \tag{2}$$

$$dy_{ij} = \eta^{[y]}_{1i} \times \gamma^{[y]}_{j-1} \qquad (j = 2, \ldots, J). \tag{3}$$

Equations 1 and 2 together specify a LBGM, where y_{ij}, y^*_{ij}, and $\epsilon^{[y]}_{ij}$ are the observed measurement, latent true score, and residual for the i^{th} individual at time j, respectively. At the baseline measurement (i.e., $j = 1$), the true score corresponds to the initial status growth factor ($\eta^{[y]}_{0i}$). For subsequent measurements (i.e., $j \geq 2$), the true score at time j is calculated as a linear combination of the score at the preceding time point $j-1$ and the true change from time $j-1$ to j. This true change is further defined as the product of the time interval $(t_{ij} - t_{i(j-1)})$ and the interval-specific slope (dy_{ij}). Equation 3 further represents the interval-specific slope, dy_{ij}, with a shape factor $\eta^{[y]}_{1i}$ and $\gamma^{[y]}_{j-1}$, where γ_{j-1}

$(j = 2, \ldots, J)$ can be interpreted as the relative growth rate in relation to $\eta_{1i}^{[y]}$ during the $(j-1)^{th}$ time interval. In Liu and Perera (2024), the term $\eta_{1i}^{[y]}$ is scaled to represent the growth rate during the first time interval (i.e., $\gamma_{2-1} = 1$), as illustrated in Figure 3a. As discussed in Section 1.2, this term can also be scaled to correspond with the growth rate during any other time interval, such as the last one (i.e., $\gamma_{J-1} = 1$), as depicted in Figure 3b.

The model specified in Equations 1-3 can also be written in a matrix form:

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}_i^{[y]} \times \boldsymbol{\eta}_i^{[y]} + \boldsymbol{\epsilon}_i^{[y]}, , \tag{4}$$

$$\boldsymbol{\eta}_i^{[y]} = \boldsymbol{\mu}_{\boldsymbol{\eta}}^{[y]} + \boldsymbol{\zeta}_i^{[y]}, \tag{5}$$

where $\boldsymbol{y}_i$ is a $J \times 1$ vector representing the i^{th} individual's repeated measurements (with J denoting the number of such measurements). The vector $\boldsymbol{\eta}_i^{[y]}$ is a 2×1 vector of growth factors, where the first element (η_{0i}) signifies the initial status and the second element (η_{1i}) indicates the growth rate within a specified time interval. The $J \times 2$ matrix $\boldsymbol{\Lambda}_i^{[y]}$ consists of associated factor loadings. Finally, $\boldsymbol{\epsilon}_i^{[y]}$ is a $J \times 1$ vector of the i^{th} individual's residuals. Equation (5) expresses $\boldsymbol{\eta}_i^{[y]}$ as deviations $(\boldsymbol{\zeta}_i^{[y]})$ from the mean values of the growth factors $(\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[y]})$.

While the scaling of η_{1i} affects its interpretation, the general form of the factor loading matrix, $\boldsymbol{\Lambda}_i^{[y]}$, remains consistent. The general form is given as:

$$\boldsymbol{\Lambda}_i^{[y]} = \begin{pmatrix} 1 & 0 \\ 1 & \gamma_{2-1}^{[y]} \times (t_{i2} - t_{i1}) \\ 1 & \sum_{j=2}^{3} \gamma_{j-1}^{[y]} \times (t_{ij} - t_{i(j-1)}) \\ \cdots & \cdots \\ 1 & \sum_{j=2}^{J} \gamma_{j-1}^{[y]} \times (t_{ij} - t_{i(j-1)}) \end{pmatrix}, \tag{6}$$

where the j^{th} element in the second column represents the cumulative value of the relative rate (i.e., $\gamma_{\cdot}^{[y]}$) over time up to time j, so the product of this element and $\eta_{1i}^{[y]}$ represents the change from the initial status of the i^{th} individual (Liu & Perera, 2024). In addition, the subscript i in $\boldsymbol{\Lambda}_i^{[y]}$ emphasizes that the model accommodates individual measurement occasions.

2.2 Model Specification of Parallel Latent Basis Growth Model

In this section, we extend the univariate Latent Basis Growth Model (LBGM) to its parallel version. This parallel version enables the joint analysis of multiple repeated outcomes, such as the joint development of reading and mathematics ability. The necessity for this extension arises from the various compelling reasons that have been discussed in Section 1. We describe the parallel LBGM in the context of individual measurement occasions, extending the univariate model given in Equation (4). Assume that we have bivariate growth trajectories for repeated outcomes, the parallel LBGM can then be formally defined as follows:

$$\begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{z}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_i^{[y]} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}_i^{[z]} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\eta}_i^{[y]} \\ \boldsymbol{\eta}_i^{[z]} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_i^{[y]} \\ \boldsymbol{\epsilon}_i^{[z]} \end{pmatrix}, \tag{7}$$

where $\boldsymbol{z}_i$ is also a $J \times 1$ vector of the repeated measurements for individual i, $\boldsymbol{\eta}_i^{[z]}$, $\boldsymbol{\Lambda}_i^{[z]}$ and $\boldsymbol{\epsilon}_i^{[z]}$ are its growth factors (a 2×1 vector), the corresponding factor loadings (a $J \times 2$ matrix), and the residuals of person i (a $J \times 1$ vector), respectively. Similar to $\boldsymbol{\Lambda}_i^{[y]}$, $\boldsymbol{\Lambda}_i^{[z]}$ has a general expression but with one fixed relative growth rate γ_{j-1}, corresponding to the growth rate of $(j-1)^{th}$ time interval that $\eta_{1i}^{[z]}$ represents. We then write the outcome-specific growth factors $\boldsymbol{\eta}_i^{[u]}$ $(u = y,\ z)$ as deviations from the corresponding outcome-specific growth factor means.

$$\begin{pmatrix} \boldsymbol{\eta}_i^{[y]} \\ \boldsymbol{\eta}_i^{[z]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\eta}}^{[y]} \\ \boldsymbol{\mu}_{\boldsymbol{\eta}}^{[z]} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\zeta}_i^{[y]} \\ \boldsymbol{\zeta}_i^{[z]} \end{pmatrix}, \tag{8}$$

where $\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[u]}$ is a 2×1 vector of outcome-specific growth factor means, and $\boldsymbol{\zeta}_i^{[u]}$ is a 2×1 vector of deviations of the i^{th} individual from the means. To simplify model, we assume that $\left(\boldsymbol{\zeta}_i^{[y]}\ \boldsymbol{\zeta}_i^{[z]}\right)^T$ follows a multivariate normal distribution

$$\begin{pmatrix} \boldsymbol{\zeta}_i^{[y]} \\ \boldsymbol{\zeta}_i^{[z]} \end{pmatrix} \sim \text{MVN}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[y]} & \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[yz]} \\ & \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[z]} \end{pmatrix}\right),$$

where both $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[u]}$ and $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[yz]}$ are 2×2 matrices: $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[u]}$ is the variance-covariance matrix of the outcome-specific growth factors while $\boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[yz]}$ is the covariances between the growth factors of $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$. To simplify the model, we also assume that the individual outcome-specific residual variances are identical and independent normal distributions over time, while the residual covariances are homogeneous over time, that is,

$$\begin{pmatrix} \boldsymbol{\epsilon}_i^{[y]} \\ \boldsymbol{\epsilon}_i^{[z]} \end{pmatrix} \sim \text{MVN}\left(\mathbf{0}, \begin{pmatrix} \theta_{\epsilon}^{[y]}\boldsymbol{I} & \theta_{\epsilon}^{[yz]}\boldsymbol{I} \\ & \theta_{\epsilon}^{[z]}\boldsymbol{I} \end{pmatrix}\right),$$

where $\boldsymbol{I}$ is a $J \times J$ identity matrix.

2.3 Model Estimation

We then write the expected mean vector and variance-covariance matrix of the bivariate repeated outcome $\boldsymbol{y}_i$ and $\boldsymbol{z}_i$ in the parallel LBGM specified in Equations (7) and (8) as

$$\boldsymbol{\mu}_i = \begin{pmatrix} \boldsymbol{\mu}_i^{[y]} \\ \boldsymbol{\mu}_i^{[z]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Lambda}_i^{[y]} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_i^{[z]} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\eta}}^{[y]} \\ \boldsymbol{\mu}_{\boldsymbol{\eta}}^{[z]} \end{pmatrix} \tag{9}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_i &= \begin{pmatrix} \boldsymbol{\Sigma}_i^{[y]} & \boldsymbol{\Sigma}_i^{[yz]} \\ & \boldsymbol{\Sigma}_i^{[z]} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Lambda}_i^{[y]} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_i^{[z]} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[y]} & \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[yz]} \\ & \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[z]} \end{pmatrix} \times \begin{pmatrix} \boldsymbol{\Lambda}_i^{[y]} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_i^{[z]} \end{pmatrix}^T \\ &+ \begin{pmatrix} \theta_\epsilon^{[y]} \boldsymbol{I} & \theta_\epsilon^{[yz]} \boldsymbol{I} \\ & \theta_\epsilon^{[z]} \boldsymbol{I} \end{pmatrix}.\end{aligned} \tag{10}$$

The parameters in the parallel LBGM specified in Equations (7) and (8) include the mean vector and variance-covariance matrix of the growth factors, the outcome-specific relative growth rate, the variance-covariance matrix of the residuals. Accordingly, we define

$$\begin{aligned}\boldsymbol{\Theta} =& \{\boldsymbol{\mu}_{\boldsymbol{\eta}}^{[u]}, \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[u]}, \boldsymbol{\Psi}_{\boldsymbol{\eta}}^{[yz]}, \boldsymbol{\gamma}^{[u]}, \theta_\epsilon^{[u]}, \theta_\epsilon^{[yz]}\} \\ =& \{\mu_{\eta_0}^{[u]}, \mu_{\eta_1}^{[u]}, \psi_{00}^{[u]}, \psi_{01}^{[u]}, \psi_{11}^{[u]}, \psi_{00}^{[yz]}, \psi_{01}^{[yz]}, \psi_{10}^{[yz]}, \psi_{11}^{[yz]}, \gamma_{j-1}^{[u]}, \\ & \theta_\epsilon^{[u]}, \theta_\epsilon^{[yz]}\}, \ u = y, \ z \\ j =& \begin{cases} 3, \ldots, J & \text{Model specification in Figure 3a} \\ 2, \ldots, J-1 & \text{Model specification in Figure 3b} \end{cases}\end{aligned} \tag{11}$$

to list the parameters that we need to estimated in the proposed model.

We estimate $\boldsymbol{\Theta}$ using full information maximum likelihood (FIML) to account for the individual measurement occasions and potential heterogeneity of individual contributions to the likelihood function. In this present study, the proposed model is built using the R package *OpenMx* with CSOLNP optimizer (Boker et al., 2020; Hunter, 2018; Neale et al., 2016; Pritikin, Hunter, & Boker, 2015). We provide *OpenMx* code of the proposed parallel LBGM and a demonstration in the online appendix (https://github.com/Veronica0206/LCSM_projects). We also provide *Mplus* 8 code of the proposed model for researchers who are interested in using *Mplus*.

3 Model Evaluation

We aim to assess the effectiveness of the proposed parallel LBGM by employing Monte Carlo simulation studies. Specifically, we examine the model's performance using several metrics: the relative bias, the empirical standard error (SE), the relative root-mean-square error (RMSE), and the empirical coverage probability for a nominal 95% confidence interval for each parameter. These metrics are commonly used in simulation studies to evaluate the performance of statistical methodologies or models. The definitions and estimates for these metrics are presented in Table 1.

Following practices in simulation studies as suggested by Morris, White, and Crowther (2019), we empirically determined the number of replications

Table 1: Performance Metrics: Definitions and Estimates

Criteria	**Definition**	**Estimate**
Relative Bias	$E_{\hat{\theta}}(\hat{\theta}-\theta)/\theta$	$\sum_{s=1}^{S}(\hat{\theta}_s-\theta)/S\theta$
Empirical SE	$\sqrt{Var(\hat{\theta})}$	$\sqrt{\sum_{s=1}^{S}(\hat{\theta}_s-\bar{\theta})^2/(S-1)}$
Relative RMSE	$\sqrt{E_{\hat{\theta}}(\hat{\theta}-\theta)^2}/\theta$	$\sqrt{\sum_{s=1}^{S}(\hat{\theta}_s-\theta)^2/S}/\theta$
Coverage Prob.	$Pr(\hat{\theta}_{\text{lower}} \leq \theta \leq \hat{\theta}_{\text{upper}})$	$\sum_{s=1}^{S} I(\hat{\theta}_{\text{lower},s} \leq \theta \leq \hat{\theta}_{\text{upper},s})/S$

[a] θ: the population value of the parameter of interest
[b] $\hat{\theta}$: the estimate of θ
[c] S: the number of replications and set as $1,000$ in our simulation study
[d] $s = 1, \ldots, S$: indexes the replications of the simulation
[e] $\hat{\theta}_s$: the estimate of θ from the s^{th} replication
[f] $\bar{\theta}$: the mean of $\hat{\theta}_s$'s across replications
[g] $I()$: an indicator function
[h] Coverage Prob.: coverage probability

to be $S = 1,000$. The pilot simulation study was conducted to ensure that the chosen number of replications would provide reliable performance metrics. Among the four performance metrics, the (relative) bias is of utmost importance. The pilot simulation revealed that the standard errors of bias, calculated as Monte Carlo SE(Bias) $= \sqrt{\frac{\text{Var}(\hat{\theta})}{S}}$, were less than 0.15 across all parameters, except for $\psi_{00}^{[u]}$ and $\psi_{00}^{[yz]}$. To maintain the Monte Carlo standard error of bias below 0.05, at least 900 replications are needed. Thus, we decided to proceed with $S = 1,000$ replications to account for variability and ensure a more robust evaluation.

3.1 Design of Simulation Study

To thoroughly evaluate the proposed parallel LBGM, we designed a comprehensive set of simulation studies, the conditions of which are outlined in Table 2. A key factor in the effectiveness of a model designed for longitudinal data is the number of repeated measures. We hypothesize that the proposed model's performance will improve with an increasing number of repeated measurements. To test this hypothesis, we considered two levels for the number of repeated measures: six and ten. For conditions with ten repeated measures, we investigated whether equally-placed study waves or unequally-placed waves affect model performance, assuming that the study duration remains constant across conditions. This consideration reflects real-world longitudinal study practices, where measurement waves are typically not equally spaced, often occurring more frequently at the beginning. We aimed to determine if such setups impact model performance. In scenarios with six repeated measures, we examined the model's performance under the more challenging condition of a shorter study duration with the hypothesis that a shorter duration poses greater challenges for the

model due to less available data to accurately capture the underlying growth trajectory. Measurement occasions are individuated by a 'medium'-width time window, $(-0.25, +0.25)$ around each wave (Coulombe et al., 2015).

Another key variable of interest is the correlation between the two trajectories, as the proposed model is designed for analyzing joint longitudinal processes. Three correlation levels for the between-construct growth factors are considered: ± 0.3 and 0. We are interested in how model over-specification affects performance in zero-correlation conditions, and whether the sign of the correlation (± 0.3) has any impact on model performance. Additionally, we explore the influence of varying trajectory shapes, quantified by the relative growth rate in each time interval. As specified in Table 2, the change patterns considered include both increasing and decreasing growth rates. Moreover, we evaluate the model's performance across different sample sizes ($n = 200$ and $n = 500$) and levels of outcome-specific residual variances ($\theta_{\epsilon}^{[u]} = 1$ or $\theta_{\epsilon}^{[u]} = 2$) to gauge the effects of sample size and measurement precision. In the simulation design, factors considered less critical to the proposed model's performance, such as the distribution of growth factors and the correlation of between-construct residuals, were held constant.

3.2 Data Generation and Simulation Step

To evaluate the performance of the proposed parallel LBGMs, we conducted a simulation study according to the design presented in Table 2. Each condition was replicated 1,000 times to ensure a robust assessment. The steps for the simulation are outlined as follows:

1. **Growth Factor Generation:** Utilizing the *MASS* R package (Venables & Ripley, 2002), generate the growth factors for both longitudinal processes based on the pre-defined mean vector and variance-covariance matrix as specified in Table 2. The *MASS* package is used for its reliability in generating multivariate Gaussian samples.
2. **Time Structure:** Generate the time structure with J waves t_j as defined in Table 2. Add a uniform disturbance following $U(t_j - \Delta, t_j + \Delta)$ around each wave to obtain individual measurement occasions t_{ij}.
3. **Factor Loadings Calculation:** Compute the factor loadings for each individual of each longitudinal process as Equation 6, using the relative growth rates and individual measurement intervals.
4. **Measurement Value Computation:** Calculate the values of bivariate repeated measurements, incorporating growth factors, factor loadings, and the pre-defined residual variance-covariance structure.
5. **LBGM Implementation:** Execute the proposed LBGM models on the generated dataset, estimating the model parameters and constructing 95% Wald confidence intervals.
6. **Replication:** Repeat steps 1-5 until 1,000 convergent solutions are obtained, as this number of replications provides a stable estimate of performance metrics such as bias and coverage probability.

Table 2: Simulation Design for Parallel Latent Basis Growth Model in the Framework of Individual Measurement Occasions

Fixed Conditions	
Variables	**Conditions**
Distribution of the Intercept	$\eta_{0i}^{[y]} \sim N(50, 5^2)$; $\eta_{0i}^{[z]} \sim N(30, 5^2)$ (i.e., $\mu_{\eta_0}^{[y]} = 50$, $\psi_{00}^{[y]} = 25$; $\mu_{\eta_0}^{[z]} = 30$, $\psi_{00}^{[z]} = 25$)
Distribution of the Shape Factor	$\eta_{1i}^{[y]} \sim N(4, 1^2)$; $\eta_{1i}^{[z]} \sim N(5, 1^2)$ (i.e., $\mu_{\eta_1}^{[y]} = 4$, $\psi_{11}^{[y]} = 1$; $\mu_{\eta_1}^{[z]} = 5$, $\psi_{11}^{[z]} = 1$)
Correlations of Within-construct GFs	$\rho^{[u]} = 0.3$ $(u = y, z)$
Correlation between Residuals	$\rho_\epsilon = 0.3$
Manipulated Conditions (Full Factorial)	
Variables	**Conditions**
Sample Size	$n = 200, 500$
Time (t_j)	6 equally-spaced: $t_j = 0, 1.00, 2.00, 3.00, 4.00, 5.00$
	10 equally-spaced: $t_j = 0, 1.00, 2.00, 3.00, 4.00, 5.00, 6.00, 7.00, 8.00, 9.00$
	10 unequally-spaced: $t_j = 0, 0.75, 1.50, 2.55, 3.00, 3.75, 4.50, 6.00, 7.50, 9.00$
Individual t_{ij}	$t_{ij} \sim U(t_j - \Delta, t_j + \Delta)$ $(\Delta = 0.25)$
Relative Growth Rate in Each Time Interval[a]	6 waves: $r^{[u]} = 1.0, 0.8, 0.6, 0.4, 0.2$ $(u = y, z)$
	6 waves: $r^{[u]} = 0.2, 0.4, 0.6, 0.8, 1.0$ $(u = y, z)$
	10 waves: $r^{[u]} = 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2$ $(u = y, z)$
	10 waves: $r^{[u]} = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0$ $(u = y, z)$
Correlation of Between-construct GFs	$\rho = 0, \pm 0.3$
Residual Variance	$\theta_\epsilon^{[u]} = 1, 2$ $(u = y, z)$

[a] Growth rate is the relative growth rate, which is defined as the absolute growth rate over the value of shape factor.

4 Result

4.1 Model Convergence

Before assessing the four performance measures of the proposed parallel LBGM, we first examined its convergence rate[1]. The model exhibited excellent convergence, as evidenced by a 100% rate across all simulation conditions listed in Table 2.

4.2 Performance Measures

This section summarizes the simulation results for four key performance metrics: relative bias, empirical SE, relative RMSE, and empirical coverage probability for a nominal 95% confidence interval. We calculated these metrics for each parameter across 1,000 repetitions under each condition, and summarized the median and range values for all conditions given the scale of parameters and simulation setups. The proposed model generally yielded unbiased and accurate point estimates with target coverage probabilities. Further details for each performance metric are provided in the Online Supplementary Document.

The proposed model produced unbiased and accurate point estimates. Specifically, the magnitudes of the relative biases for outcome-specific growth factor means, variances, and relative growth rates were below 0.004, 0.013, and 0.012, respectively[2]. The magnitudes of the relative RMSEs for outcome-specific growth factor means, variances, and relative growth rates were below 0.05, 0.15, and 0.23, respectively[3]. Moreover, the model demonstrated excellent empirical coverage probabilities, with median values approximating 0.95. Given these consistently strong performance metrics, further investigations into the effect of different simulation conditions were deemed unnecessary.

5 Application

In this section, we demonstrate how to employ the proposed parallel LBGM to analyze real-world data. This application section includes two examples. In the first example, we illustrate the recommended steps to construct the proposed model in practice. In the second example, we demonstrate how to apply the proposed model to analyze joint longitudinal processes with a more complicated

[1] In this study, we define convergence rate as the achievement of an *OpenMx* status code of 0, indicating successful optimization, in up to 10 runs with varied initial values (Neale et al., 2016).

[2] Previous simulations have often regarded relative bias in regression coefficients as acceptable if it was below 10%, which is commonly considered a guideline when assessing relative bias (Leite, 2017; Poon & Wang, 2010).

[3] Regarding relative RMSE, while there is no universally accepted benchmark for simulation studies, model accuracy is generally considered excellent when the score is below 10%, good when it ranges from 10% to 20%, fair when it falls between 20% and 30%, and poor when it exceeds 30% (Jadon, Patil, & Jadon, 2022).

data structure, where two repeated outcomes have different time frames. We randomly selected 400 students from the Early Childhood Longitudinal Study Kindergarten Cohort of 2010-2011 (ECLS-K: 2011), all of whom had complete records of repeated reading and mathematics scores based on Item Response Theory (IRT), as well as their age in months at each wave[4].

ECLS-K: 2011 is a national longitudinal study of US children enrolled in around 900 kindergarten programs beginning in the 2010-2011 school year. In ECLS-K: 2011, children's reading and mathematics abilities were assessed in nine waves: fall and spring of kindergarten (2010-2011), first (2011-2012) and second (2012-2013) grade, respectively, as well as spring of the 3^{rd} (2014), 4^{th} (2015), and 5^{th} (2016) grade. Only about 30% of students were assessed in the fall semesters of 2011 and 2012 (Lê, Norman, Tourangeau, Brick, & Mulligan, 2011). In the first example, we used all nine waves of reading and mathematics IRT scores to demonstrate how to apply the proposed model. In the second example, we utilized all nine waves of reading IRT scores but only the mathematics scores obtained in the spring semesters to mimic one possible complex time structure in practice. Note that the initial status and the number of measurement occasions of the two abilities are different in the second example. Additionally, we employed children's age in months rather than their grade-in-school to have individual measurement occasions. The subsample included 41.50% White, 7.25% Black, 37.00% Latinx, 8.25% Asian, and 6.00% of other ethnicity.

5.1 Analyze Joint Longitudinal Records with The Same Time Structure

Following Blozis et al. (2008) and Liu and Perera (2022), we first constructed a latent growth curve model to examine each longitudinal process in isolation before analyzing joint development. Specifically, we employed a LBGM to explore the univariate development of either reading or mathematics from Grade K to 5. Figure 4 illustrates the model-implied curves superimposed on the smooth lines for each ability. For each ability, the estimates from the LBGM produced model-implied trajectories that closely align with the smooth lines representing the observed individual data.

We then applied the proposed parallel LBGM to analyze the joint development of reading and mathematics abilities. Figure 5 illustrates the model-implied curves superimposed on the smooth lines for each ability obtained from the parallel model. From the figure, it is evident that the model-implied curves of the parallel models did not differ from those of the univariate growth models shown in Figure 4. Table 3 presents the parameter estimates of interest for joint development.

Note that we defined $\eta_1^{[u]}$ as the growth rate in the final time interval for each ability's longitudinal process (i.e., the model specification in Figure 3b).

[4] The total sample size of ECLS-K: 2011 is $n = 18174$. The number of rows after removing records with missing values (i.e., entries with any of NaN/-9/-8/-7/-1) is $n = 2290$.

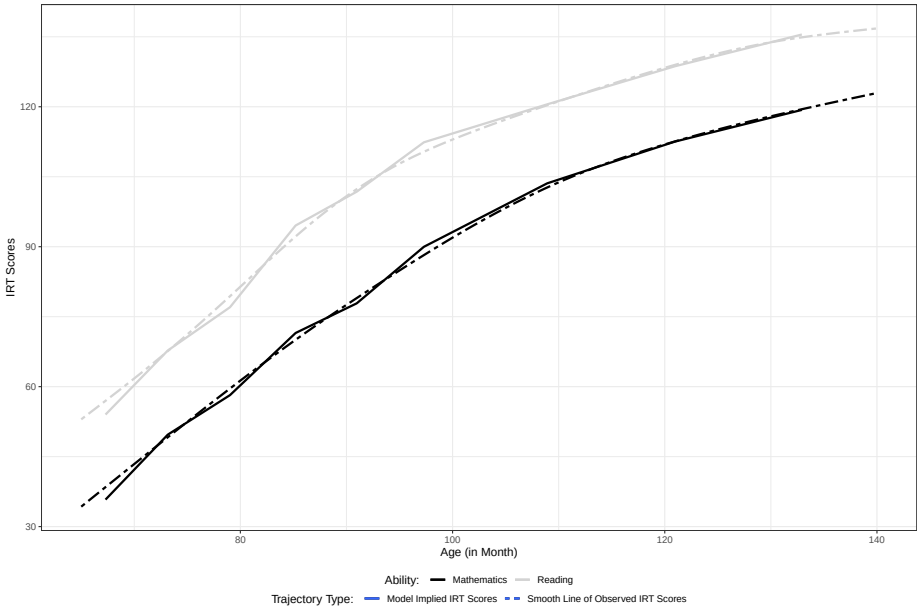


Figure 4: Model Implied Trajectory and Smooth Line of Univariate Development

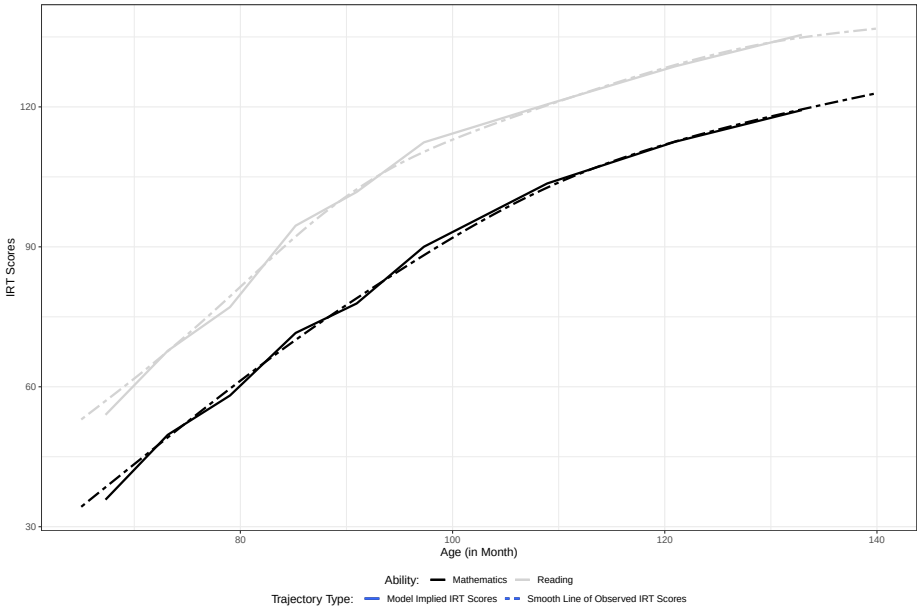


Figure 5: Model Implied Trajectory and Smooth Line of Bivariate Development with The Same Time Structures

Table 3: Estimates of Parallel Latent Basis Growth Model for Reading and Mathematics Ability with the Same Time Structures

	Reading IRT Scores		**Math IRT Scores**		**Covariance**	
Mean	Estimate (SE)	P value	Estimate (SE)	P value	Estimate (SE)	P value
Initial Status[a]	53.984 (0.724)	$< 0.0001^{*}$ [b]	35.797 (0.659)	$< 0.0001^{*}$	—[c]	—
Rate[d] of Interval 1	2.341 (0.080)	$< 0.0001^{*}$	2.369 (0.071)	$< 0.0001^{*}$	—	—
Rate of Interval 2	1.596 (0.077)	$< 0.0001^{*}$	1.437 (0.068)	$< 0.0001^{*}$	—	—
Rate of Interval 3	2.823 (0.077)	$< 0.0001^{*}$	2.169 (0.066)	$< 0.0001^{*}$	—	—
Rate of Interval 4	1.256 (0.080)	$< 0.0001^{*}$	1.098 (0.070)	$< 0.0001^{*}$	—	—
Rate of Interval 5	1.679 (0.074)	$< 0.0001^{*}$	1.920 (0.066)	$< 0.0001^{*}$	—	—
Rate of Interval 6	0.701 (0.041)	$< 0.0001^{*}$	1.167 (0.036)	$< 0.0001^{*}$	—	—
Rate of Interval 7	0.676 (0.039)	$< 0.0001^{*}$	0.742 (0.034)	$< 0.0001^{*}$	—	—
Rate of Interval 8	0.566 (0.040)	$< 0.0001^{*}$	0.568 (0.034)	$< 0.0001^{*}$	—	—
Variance	Estimate (SE)	P value	Estimate (SE)	P value	Estimate (SE)	P value
Initial Status	164.926 (13.085)	$< 0.0001^{*}$	139.878 (10.876)	$< 0.0001^{*}$	126.794 (10.650)	$< 0.0001^{*}$
Rate of Interval 1	0.111 (0.013)	$< 0.0001^{*}$	0.106 (0.011)	$< 0.0001^{*}$	0.064 (0.009)	$< 0.0001^{*}$
Rate of Interval 2	0.051 (0.007)	$< 0.0001^{*}$	0.039 (0.005)	$< 0.0001^{*}$	0.026 (0.004)	$< 0.0001^{*}$
Rate of Interval 3	0.161 (0.018)	$< 0.0001^{*}$	0.089 (0.010)	$< 0.0001^{*}$	0.070 (0.009)	$< 0.0001^{*}$
Rate of Interval 4	0.032 (0.005)	$< 0.0001^{*}$	0.023 (0.003)	$< 0.0001^{*}$	0.016 (0.002)	$< 0.0001^{*}$
Rate of Interval 5	0.057 (0.007)	$< 0.0001^{*}$	0.070 (0.008)	$< 0.0001^{*}$	0.037 (0.005)	$< 0.0001^{*}$
Rate of Interval 6	0.010 (0.001)	$< 0.0001^{*}$	0.026 (0.003)	$< 0.0001^{*}$	0.009 (0.001)	$< 0.0001^{*}$
Rate of Interval 7	0.009 (0.001)	$< 0.0001^{*}$	0.010 (0.001)	$< 0.0001^{*}$	0.006 (0.001)	$< 0.0001^{*}$
Rate of Interval 8	0.006 (0.001)	$< 0.0001^{*}$	0.006 (0.001)	$< 0.0001^{*}$	0.004 (0.001)	0.0001^{*}

[a] The initial Status was defined as 60 months old in this case.

[b] * indicates statistical significance at 0.05 level.

[c] — indicates that the metric was not available in the model.

[d] The mean, variance, and covariance of rate in each interval were the corresponding value of absolute growth rate, which can be obtained by *R* function *mxAlgebra()* from estimated shape factor and relative growth rate.

Consequently, in Table 3, the parameters related to 'initial status' and 'rate of Interval 8' were directly estimated from the proposed model, while others were obtained using the function *mxAlgebra()*[5] in the *R* package *OpenMx*. It is important to note that the estimated initial status and interval-specific slopes remain unaffected if $\eta_1^{[u]}$ is defined as the growth rate in the first time interval for each ability's longitudinal process (i.e., the model specification in Figure 3a). This difference in specification only affects the interpretation of $\eta_1^{[u]}$. Specifically, when $\eta_1^{[u]}$ is scaled to be the slope in the first time interval, the correlation of the two $\eta_1^{[u]}$ for the two outcomes indicates how the growth rates are related in the first interval. In contrast, when $\eta_1^{[u]}$ is scaled to be the slope in the last time interval, the correlation reflects how the growth rates are related during the final interval.

From Figure 5 and Table 3, we observed that the development of both reading and mathematics abilities generally slowed down after Grade 3, which aligns with earlier studies (Liu & Perera, 2022; Peralta et al., 2022). Additionally, there was a positive association between the development of reading and mathematics abilities, indicated by statistically significant intercept-intercept and slope-slope covariances in each time interval. After standardizing the covariances, the intercept-intercept correlation and each interval-specific slope-slope correlation were found to be 0.83 and 0.58, respectively. This suggests that, on average, a child who performed better in reading tests at Grade K also tended to perform better in mathematics examinations, and vice versa. Moreover, on average, children who showed more rapid gains in reading ability also tended to exhibit faster improvement in mathematics, and vice versa.

5.2 Analyze Joint Longitudinal Records with Different Time Structures

In this section, we use the proposed parallel LBGM to investigate the joint development trajectories of reading and mathematics abilities. We retained all nine measurement occasions for reading ability but included only the spring semester measurements for mathematics ability (i.e., Waves 2, 4, 6, 7, 8, and 9). In this configuration, both the initial statuses and the number of measurement occasions differ between the two abilities. Figure 6 illustrates the model-implied curves superimposed on the smooth lines representing each ability in this model. The figure reveals that the model-implied trajectories vary only minimally from those presented in Figure 5 due to fewer measurement occasions for mathematics ability, but they still sufficiently capture the smooth lines of the observed individual data.

Table 4 presents the estimated parameters of interest for the joint model with differing time structures. Note that there are 8 time intervals for the development of reading ability (corresponding to 9 measurement occasions) but only

[5] By using *mxAlgebra()*, we specify algebraic expressions for new parameters, enabling *OpenMx* to estimate their point values along with standard errors.

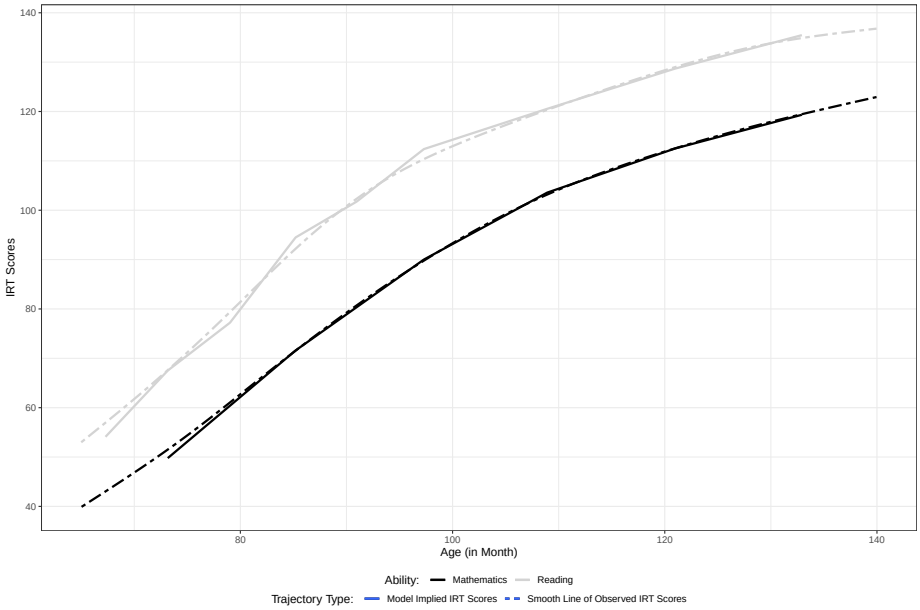


Figure 6: Model Implied Trajectory and Smooth Line of Bivariate Development with Different Time Structures

5 time intervals for the development of mathematics ability because three measurements from the fall semesters were excluded. During the first time interval for mathematics, which corresponds to Intervals 2 and 3 for reading ability (as detailed in Table 3), the estimated growth rate was 1.811. This value represents an average of the growth rates 1.437 and 2.169 from Interval 2 and Interval 3, respectively, in Table 3. These findings suggest that our proposed model effectively captures the underlying patterns of growth trajectories, even with fewer measurements.

Table 4: Estimates of Parallel Latent Basis Growth Model for Reading and Mathematics Ability with Different Time Structures[a]

	Reading IRT Scores		**Math IRT Scores**		**Covariance**	
Mean	Estimate (SE)	P value	Estimate (SE)	P value	Estimate (SE)	P value
Initial Status[b]	54.100 (0.724)	$< 0.0001^{*}$[c]	49.782 (0.694)	$< 0.0001^{*}$	—[d]	—
Rate[e]of Interval[f]1	2.288 (0.080)	$< 0.0001^{*}$	—	—	—	—
Rate of Interval 2	1.653 (0.077)	$< 0.0001^{*}$	1.811 (0.037)	$< 0.0001^{*}$	—	—
Rate of Interval 3	2.793 (0.077)	$< 0.0001^{*}$	1.811 (0.037)	$< 0.0001^{*}$	—	—
Rate of Interval 4	1.263 (0.080)	$< 0.0001^{*}$	1.523 (0.036)	$< 0.0001^{*}$	—	—
Rate of Interval 5	1.679 (0.074)	$< 0.0001^{*}$	1.523 (0.036)	$< 0.0001^{*}$	—	—
Rate of Interval 6	0.702 (0.041)	$< 0.0001^{*}$	1.167 (0.037)	$< 0.0001^{*}$	—	—
Rate of Interval 7	0.676 (0.039)	$< 0.0001^{*}$	0.742 (0.035)	$< 0.0001^{*}$	—	—
Rate of Interval 8	0.567 (0.039)	$< 0.0001^{*}$	0.572 (0.035)	$< 0.0001^{*}$	—	—
Variance	Estimate (SE)	P value	Estimate (SE)	P value	Estimate (SE)	P value
Initial Status	164.755 (13.095)	$< 0.0001^{*}$	156.78 (12.804)	$< 0.0001^{*}$	130.957 (11.353)	$< 0.0001^{*}$
Rate of Interval 1	0.105 (0.012)	$< 0.0001^{*}$	—	—	—	—
Rate of Interval 2	0.055 (0.007)	$< 0.0001^{*}$	0.060 (0.007)	$< 0.0001^{*}$	0.032 (0.005)	$< 0.0001^{*}$
Rate of Interval 3	0.157 (0.018)	$< 0.0001^{*}$	0.060 (0.007)	$< 0.0001^{*}$	0.055 (0.008)	$< 0.0001^{*}$
Rate of Interval 4	0.032 (0.005)	$< 0.0001^{*}$	0.042 (0.005)	$< 0.0001^{*}$	0.021 (0.003)	$< 0.0001^{*}$
Rate of Interval 5	0.057 (0.007)	$< 0.0001^{*}$	0.042 (0.005)	$< 0.0001^{*}$	0.028 (0.004)	$< 0.0001^{*}$
Rate of Interval 6	0.010 (0.001)	$< 0.0001^{*}$	0.025 (0.003)	$< 0.0001^{*}$	0.009 (0.001)	$< 0.0001^{*}$
Rate of Interval 7	0.009 (0.001)	$< 0.0001^{*}$	0.010 (0.001)	$< 0.0001^{*}$	0.005 (0.001)	$< 0.0001^{*}$
Rate of Interval 8	0.006 (0.001)	$< 0.0001^{*}$	0.006 (0.001)	$< 0.0001^{*}$	0.004 (0.001)	0.0001^{*}

[a] In this joint model, we included the measurements of reading ability at all nine waves, but we only included the measures of mathematics ability at Wave 2, 4, 6, 7, 8, and 9.

[b] The initial Status of reading ability was defined as the measurement at 60 months old, while that of mathematics ability was the measurement half a year later.

[c] * indicates statistical significance at 0.05 level.

[d] — indicates that the metric was not available in the model.

[e] The mean, variance, and covariance of rate in each interval were the corresponding value of absolute growth rate, which can be obtained by *R* function *mxAlgebra()* from estimated shape factor and relative growth rate.

[f] Each 'interval' was defined as the interval between any two consecutive measurement occasions of reading ability. The estimates of mathematics ability during the first interval are not applicable because the first measure of mathematics ability was Wave 2. The estimated means and variances of mathematics ability in Interval 2 (Interval 4) and Interval 3 (Interval 5) were the same because we took out its measurement at Wave 3 (Wave 5).

6 Discussion

This article extends the latent basis growth model with the novel specification proposed by Liu and Perera (2024) to explore joint nonlinear longitudinal processes in the framework of individual measurement occasions. This framework is particularly advantageous when investigating parallel development because it helps avoid inadmissible estimation and allows for different time structures across outcomes. Additionally, the proposed model allows scaling the second growth factor as the growth rate during any time interval. In the present study, we specify the second growth factor as the growth rate during either the first or last time interval and estimate the relative rates for each of the other intervals for each repeated outcome. We demonstrate that the proposed parallel LBGM can provide unbiased and accurate point estimates with target coverage probabilities through simulation studies. Additionally, we apply the proposed model to analyze the joint development of reading and mathematics abilities, using the same or different time structures. Our analysis relies on a subsample of $n = 400$ from ECLS-K: 2011.

6.1 Practical Considerations

In this section, we provide recommendations for empirical researchers based on both the simulation study and real-world data analyses. First, although we scale the shape factor η_1 as the growth rate in the first or last time interval of the study duration, it can be specified as the growth rate in any time interval. Note that the interpretation of γ_{j-1} remains as the relative growth rate to η_1 during the $(j-1)^{th}$ time interval. From the proposed parallel LBGM, we obtain the estimates of the mean and variance of shape factor and the fixed effects of relative growth rates for each construct. Using the *mxAlgebra()* function from the *OpenMx* R package, we derive both fixed and random effects for the absolute growth rate of each time interval, as detailed in the Application section.

In addition, the proposed model is capable of estimating the covariance of between-construct intercepts and that of between-construct shape factors directly. We can derive the covariance of between-construct growth rates for each interval by using the function *mxAlgebra()*. Note that the correlation of the between-construct growth rates is constant because we only estimate fixed effects of relative growth rates.

Third, as the latent basis growth model serves primarily as an exploratory tool, allowing trajectory characteristics to emerge from the data rather than being specified *a priori*, researchers may also be interested in exploring other aspects, such as the change-from-baseline values at each measurement wave for each repeated outcome. We can also derive these features with the function *mxAlgebra()*. In the online appendix (https://github.com/Veronica0206/LCSM_projects), we also provide code to demonstrate how to derive the values of change-from-baseline.

6.2 Methodological Considerations and Future Directions

There are several directions to consider for future studies. First, similar to the standard implementation of latent basis growth models, the proposed model requires a strict proportionality assumption (McNeish, 2020; Wu & Lang, 2016). Wu and Lang (2016) showed that this assumption might potentially result in biased estimates by simulation studies. McNeish (2020) demonstrated that this assumption could be relaxed by specifying random factor loadings of the shape factor. In the same way, we can also relax the proportionality assumption for the proposed parallel LBGM. Note that the extended model, where both the shape factor and relative growth rates are random coefficients, cannot be specified in a frequentist SEM software because these random coefficients enter the model in a multiplicative fashion (i.e., a nonlinear fashion). Similar to McNeish (2020), the extended model can be constructed in Bayesian software such as *jags* or *stan*.

Second, it is not our intention to show that the proposed parallel LBGM is better than any other parallel growth models with parametric or semi-parametric functional forms. The proposed model is a versatile tool for exploratory analyses; it should perform well to detect the trends of trajectories or whether a spike exists over the study duration. However, the insights directly related to research questions might be limited. Accordingly, subsequent analyses may need to be based on the estimates generated by the proposed model. For instance, if we obtain evidence suggesting that developmental processes can generally be divided into two stages, we may employ the parallel bilinear spline growth model (Liu & Perera, 2022) to further estimate the individual transition time to the stage with a slower growth rate. Alternatively, we can constrain the relative growth rates of multiple time intervals to be the same to have a more parsimonious model. Therefore, statistical methods for comparing the full model to a more parsimonious one need to be proposed and tested.

Third, as in any latent growth curve model, baseline covariates can be added to predict the intercept or the growth rate. Additionally, a time-varying covariate can also be added to estimate its effect on the measurements while simultaneously modeling parallel change patterns in these measurements.

6.3 Concluding Remarks

In this article, we propose a novel expression of latent basis growth models to allow for individual measurement occasions and further extend the model to analyze joint longitudinal processes. The results of both the simulation studies and real-world data analyses underscore the model's valuable capabilities for exploring parallel nonlinear change patterns. As discussed above, the proposed method offers avenues for both practical extensions and further methodological examination.

References

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135–167. doi: https://doi.org/10.3102/10769986028002135

Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures. *Psychological Methods*, *9*(3), 334-353. doi: https://doi.org/10.1037/1082-989x.9.3.334

Blozis, S. A., & Cho, Y. (2008). Coding and centering of time in latent curve models in the presence of interindividual time heterogeneity. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(3), 413-433. doi: https://doi.org/10.1080/10705510802154299

Blozis, S. A., Harring, J. R., & Mels, G. (2008). Using lisrel to fit nonlinear latent curve models. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(2), 346-369. doi: https://doi.org/10.1080/10705510801922639

Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R., ... Kirkpatrick, R. M. (2020). Openmx 2.17.2 user guide [Computer software manual]. Retrieved from `https://vipbg.vcu.edu/vipbg/OpenMx2/docs/OpenMx/2.17.2/OpenMxUserGuide.pdf`

Coulombe, P., Selig, J. P., & Delaney, H. D. (2015). Ignoring individual differences in times of assessment in growth curve modeling. *International Journal of Behavioral Development*, *40*(1), 76-86. doi: https://doi.org/10.1177/0165025415577684

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529–569. doi: https://doi.org/10.1207/s15327906mbr3804_5

Dumenci, L., Perera, R. A., Keefe, F. J., Ang, D. C., J., S., Jensen, M. P., & Riddle, D. L. (2019). Model-based pain and function outcome trajectory types for patients undergoing knee arthroplasty: A secondary analysis from a randomized clinical trial. *Osteoarthritis and cartilage*, *27*(6), 878-884. doi: https://doi.org/10.1016/j.joca.2019.01.004

Duncan, S. C., & Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate behavioral research*, *29*(4), 313–338. doi: https://doi.org/10.1207/s15327906mbr2904_1

Duncan, S. C., & Duncan, T. E. (1996). A multivariate latent growth curve analysis of adolescent substance use. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(4), 323–347. doi: https://doi.org/10.1080/10705519609540050

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Press.

Grimm, K. J., Steele, J. S., Ram, N., & Nesselroade, J. R. (2013). Exploratory latent growth models in the structural equation modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 568-591. doi: https://doi.org/10.1080/10705511.2013.824775

Hunter, M. D. (2018). State space modeling in an open source, modular, structural equation modeling environment. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(2), 307-324. doi: https://doi.org/10.1080/10705511.2017.1369354

Jadon, A., Patil, A., & Jadon, S. (2022). *A comprehensive survey of regression based loss functions for time series forecasting.* doi: https://doi.org/10.1007/978-981-97-3245-6_9

Lê, T., Norman, G., Tourangeau, K., Brick, J. M., & Mulligan, G. (2011). Early childhood longitudinal study: Kindergarten class of 2010-2011 - sample design issues. *JSM Proceedings*, 1629-1639.

Leite, W. (2017). *Practical propensity score methods using r.* Thousand Oaks, CA: Sage Publications. doi: https://doi.org/10.4135/9781071802854

Liu, J., & Perera, R. A. (2022). Estimating knots and their association in parallel bilinear spline growth curve models in the framework of individual measurement occasions. *Psychological Methods*, *27*(5), 703–729. doi: https://doi.org/10.1037/met0000309

Liu, J., & Perera, R. A. (2023). Extending growth mixture model to assess heterogeneity in joint development with piecewise linear trajectories in the framework of individual measurement occasions. *Psychological Methods*, *28*(5), 1029–1051. doi: https://doi.org/10.1037/met0000500

Liu, J., & Perera, R. A. (2024). Estimating rate of change for nonlinear trajectories in the framework of individual measurement occasions: A new perspective on growth curves. *Behavior Research Methods*, *56*, 1349–1375. doi: https://doi.org/10.3758/s13428-023-02097-2

Lyons, M. J., Panizzon, M. S., Liu, W., McKenzie, R., Bluestone, N. J., Grant, M. D., ... Xian, H. (2017). A longitudinal twin study of general cognitive ability over four decades. *Developmental psychology*, *53*(6), 1170–1177. doi: https://doi.org/10.1037/dev0000303

McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. Nesselroade & R. Cattell (Eds.), *Handbook of multivariate experimental psychology* (p. 561-614). Boston, MA: Springer. doi: https://doi.org/10.1007/978-1-4613-0893-5_17

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*(1), 110–133. doi: https://doi.org/10.2307/1130295

McNeish, D. (2020). Relaxing the proportionality assumption in latent basis models for nonlinear growth. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 817-824. doi: https://doi.org/10.1080/10705511.2019.1696201

McNulty, J. K., Wenner, C. A., & Fisher, T. D. (2016). Longitudinal associations among relationship satisfaction, sexual satisfaction, and frequency of sex in early marriage. *Archives of sexual behavior*, *45*(1), 85–97. doi: https://doi.org/10.1007/s10508-014-0444-6

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259-284. doi:

https://doi.org/10.1037/1082-989x.10.3.259

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*(1), 23-43. doi: https://doi.org/10.1037/1082-989x.5.1.23

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*, 107–122. doi: https://doi.org/10.1007/bf02294746

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074-2102. doi: https://doi.org/10.1002/sim.8086

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535-549. doi: https://doi.org/10.1007/s11336-014-9435-8

Peralta, Y., Kohli, N., Lock, E. F., & Davison, M. L. (2022). Bayesian modeling of associations in bivariate piecewise linear mixed-effects models. *Psychological Methods (Advance online publication)*, *27*(1), 44–64. doi: https://doi.org/10.1037/met0000358

Poon, W.-Y., & Wang, H.-B. (2010). Analysis of a two-level structural equation model with missing data. *Sociological Methods & Research*, *39*(1), 25-55. doi: https://doi.org/10.1177/0049124110371312

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for Item Factor Analysis. *Educational and Psychological Measurement*, *75*(3), 458-474. doi: https://doi.org/10.1177/0013164414554615

Rajmil, L., López, A. R., López-Aguilà, S., & Alonso, J. (2013). Parent–child agreement on health-related quality of life (hrqol): a longitudinal study. *Health Qual Life Outcomes*, *11*(101). doi: https://doi.org/10.1186/1477-7525-11-101

Robitaille, A., Muniz, G., Piccinin, A. M., Johansson, B., & Hofer, S. M. (2012). Multivariate longitudinal modeling of cognitive aging: Associations among change and variation in processing speed and visuospatial ability. *GeroPsych*, *25*(1), 15-24. doi: https://doi.org/10.1024/1662-9647/a000051

Shin, T., Davison, M. L., Long, J. D., Chan, C., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences*, *23*(4), 92-100. doi: https://doi.org/10.1016/j.lindif.2012.10.002

Sterba, S. K. (2014). Fitting nonlinear latent growth curve models with individually varying time points. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 630-647. doi: https://doi.org/10.1080/10705511.2014.919828

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. doi: https://doi.org/10.1007/978-0-387-21706-2

Wood, P. K., Steinley, D., & Jackson, K. M. (2015). Right-sizing statistical models for longitudinal data. *Psychological Methods*, *20*(4), 470–488. doi: https://doi.org/10.1037/met0000037

Wu, W., & Lang, K. M. (2016). Proportionality assumption in latent basis curve models: A cautionary note. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 140-154. doi: https://doi.org/10.1080/10705511.2014.938578

Journal of Behavioral Data Science, 2025, *5* (1), 51–66.
DOI: https://doi.org/10.35566/jbds/larzelere

An Innovation to Test Treatment X Pretest Interactions within Difference-in-Differences

Robert E. Larzelere* and Hua Lin

Department of Human Development and Family Science, Oklahoma State University
robert.larzelere@okstate.edu and hua.lin@okstate.edu

Abstract. We introduce a way to test Treatment X Pretest interactions within difference-in-differences (DID). Mathematically adding a Treatment X Pretest interaction to DID transforms the treatment estimate to an ANCOVA-type estimate, which differs from DID's estimate and is often biased against at-risk cases. Dual-centered ANCOVA duplicates DID's treatment estimate and can test whether that estimate varies by pretest scores. To illustrate, we test a Treatment X Pretest interaction for the effects of therapy for depression using the Fragile Families and Child Wellbeing longitudinal dataset. After centering posttest and pretest outcome data on pretest group means, DID and ANCOVA estimates are both equivalent to the original DID treatment estimate. ANCOVA of these dual-centered data can then test a Treatment X Pretest interaction.

Keywords: Difference-in-differences · Treatment X Pretest interaction · Longitudinal analyses · Causal validity · ANCOVA

1 Introduction

Longitudinal analyses that control for pre-existing differences with ANCOVA-type controls are biased against corrective actions (Larzelere, Lin, Payton, & Washburn, 2018) unless the covariates predict treatment condition perfectly (as in regression discontinuity designs). By definition, corrective actions are selected to reduce the poor prognosis of a presenting problem. Subsequent outcomes therefore constitute an unknown combination of the original poor prognosis of the problem and the extent to which the corrective action modified that prognosis. Controlling statistically for pre-existing differences via regression or matching reduces that selection bias, but rarely eliminates it. For example, a recent

* Correspondence concerning this article should be addressed to Robert E. Larzelere, Professor, Department of Human Development & Family, Science, 233 NRD Bldg., Oklahoma State University, Stillwater, OK 74078, United States. Email: robert.larzelere@okstate.edu Phone: (405) 744-2053. Fax: (405) 744-6344.

meta-analysis of efforts to improve low-performing schools found that analyses that used matching or regression methods predicted significantly worse effects than randomized studies on high-stakes math exams and marginally worse on language arts exams (Schueler, Asher, Larned, Mehrotra, & Pollard, 2021). That may be why 47% of the studies qualifying for that meta-analysis used difference-in-differences instead of regression-type controls to adjust for pre-existing differences.

There are two basic ways to analyze change in two-occasion longitudinal analyses: ANCOVA predicting residualized change scores, $Y_1|Y_0$ (Y_1 controlling for baseline Y_0), and difference-in-differences predicting simple difference scores, $Y_1 - Y_0$. ANCOVA has more statistical power (van Breukelen, 2013) but produces biased treatment estimates in non-randomized studies from invariant between-person differences (Berry & Willoughby, 2016; Hamaker, Kuiper, & Grasman, 2015) and measurement error (Huitema, 2011). Difference-in-differences overcomes these two biases, but is biased by any variations from its parallel-trends assumption. Some have recommended running both types of change-score analyses, either to bracket the true causal effect given some assumptions (Angrist & Pischke, 2009; Ding & Li, 2019) or to test robustness across alternative analyses (Duncan, Engel, Claessens, & Dowsett, 2014). A limitation of difference-in-differences has been its inability to test Treatment X Pretest interactions. For example, the meta-analysis of efforts to improve low-performing schools tested many moderators, but not whether the success of these efforts varied by the schools' previous performance on the outcomes (e.g., high-stakes testing). This article introduces a method to test whether treatment effects vary by pretest levels using difference-in-differences without inadvertently changing the treatment estimate to ANCOVA's estimate.

This article focuses on two-occasion data for two reasons. Many longitudinal studies have only two occasions (Usami, Todo, & Murayama, 2019), and these two change-score analyses are basic building blocks for more complex longitudinal analyses (Lin & Larzelere, 2024).

Treatment estimates become identical for the two change-score analyses after pretest means are equalized across treatment groups, but these robust estimates are not necessarily less biased. Different methods of equating pretest group means yield different treatment estimates (Lin & Larzelere, 2020). Pretest matching produces robust results that are equivalent to the original ANCOVA (Reichardt, 2019), which is unbiased only if the assumptions of the original ANCOVA are met (e.g., no measurement error in the covariates, equality of true pretest group means with each other: van Breukelen, 2013). Centering both posttest and pretest scores on pretest group means preserves everyone's difference score, rendering the treatment estimates robust and equivalent to the original difference-in-differences, which is unbiased under the assumption of parallel slopes under the null hypothesis. The two pairs of robust results therefore differ from each other as much as the original discrepancy between the two change-score analyses. But the dual-centered data can be analyzed with ANCOVA to

test a Treatment X Pretest interaction in a model duplicating the treatment effect from difference-in-differences (Lin & Larzelere, 2020).

1.1 Basics

Assume $X_{ij} = 1$ for treatment ($j = 2$), and $X_{ij} = 0$ for control ($j = 1$). Occasions are $t = 0$ (pretest) and $t = 1$ (posttest), with outcome variable Y_{ijt} for individual i within group j at occasion t. The equation for ANCOVA is:

$$Y_{ij1} = b_0 + b_1 X_{ij} + b_2 Y_{ij0} + e_{ij}. \tag{1}$$

The equation for difference-in-differences is:

$$Y_{ij1} - Y_{ij0} = \gamma_0 + \gamma_1 X_{ij} + \varepsilon_{ij}. \tag{2}$$

By adding Y_{ij0} to both sides of Equation (2), it can be shown that its treatment effect γ_1 is identical to the treatment effect b_1 in Equation (1) when b_2=1 in Equation (1). This is possible only when all $e_{ij} = 0$ or the variance of Y_{ij} is increasing over time. For the purposes of this article, we assume that some $e_{ij} > 0$ and that the variance of Y_{ij} is stable over time. Then the two treatment effect sizes equal each other ($b_1 = \gamma_1$) only if the pretest group means are equal to each other.

Dual-centered ANCOVA centers pretest and posttest scores on the pretest group means:

$$Y_{ij1} - \widehat{\mu}_{j0} = \omega_0 + \omega_1 X_{ij} + \omega_2 (Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}, \tag{3}$$

where the group-mean-centered pretest scores are the residuals τ_{ij} in the following equation:

$$Y_{ij0} = \widehat{\mu}_{j0} + \tau_{ij}. \tag{4}$$

Lin and Larzelere (2020) showed that, under the assumption of no pretest group mean differences, the treatment estimate in Equation (3) is identical to the treatment effect in difference-in-differences Equation (2), i.e., $\omega_1 = \gamma_1$. The $(Y_{ij0} - \widehat{\mu}_{j0})$ term is a generated regressor, however, which biases the standard error for the treatment effect ω_1 downward (Brorsen, Lin, & Larzelere, 2025; Pagan, 1984). The correct standard error can be obtained from Equation (2) or by analyzing Equations (3) and (4) together via two-stage least squares (Brorsen et al., 2025). Next we consider adding Treatment X Pretest interactions to the above analyses.

1.2 Treatment X Pretest Interactions

Standard ANCOVA. When there is a significant Treatment X Pretest interaction, treatment effects vary in magnitude and significance at different pretest scores (Huitema, 2011). Because significant interactions apply to both component predictors, the auto-regressive slope b_2 will then also vary significantly

across groups. A significant Treatment X Pretest interaction violates the ANCOVA assumption of homogeneity of the regression slope across groups. We follow Huitema (2011) and Lin (2020) in interpreting a significant Treatment X Pretest interaction.

Consider standard ANCOVA with a significant Treatment X Pretest interaction:

$$Y_{ij1} = b_0 + b_1 X_{ij} + b_2 Y_{ij0} + b_3 X_{ij} Y_{ij0} + e_{ij}. \tag{5}$$

Equation (5) can be re-arranged to indicate how the effect of Treatment X_{ij} varies by the pretest score (Lin, 2020):

$$Y_{ij1} = (b_0 + b_2 Y_{ij0}) + (b_1 + b_3 Y_{ij0}) X_{ij} + e_{ij}. \tag{6}$$

Reciprocally, the effect of the pretest Y_{ij0} on the posttest Y_{ij1} also varies by treatment condition (heterogeneity of regression slopes):

$$Y_{ij1} = (b_0 + b_1 X_{ij}) + (b_2 + b_3 X_{ij}) Y_{ij0} + e_{ij}. \tag{7}$$

One way to interpret significant Treatment X Pretest interactions is the Johnson and Neyman (1936) technique, which calculates regions of significant treatment effects at all pretest values. Alternatively, the picked-points analysis (Huitema, 2011; Lin, 2020) shows the estimated treatment effects at picked pretest values.

Equation (6) indicates that the estimated conditional effect of treatment X_{ij} on the posttest at any pretest score is:

$$\widehat{b}^*_{Tx} = b_1 + b_3 Y_{ij0}. \tag{8}$$

Reciprocally the conditional effect of the pretest on the posttest for either treatment condition according to Equation (7) is:

$$\widehat{b}^*_{lag1} = b_2 + b_3 X_{ij}. \tag{9}$$

Difference-in-Differences. To our knowledge, there is no generally accepted method of testing a Treatment X Pretest interaction within difference-in-differences without changing the main effect of treatment to ANCOVA's estimate. The reason is that tests of Treatment X Pretest interactions require both main effects to be included in the regression equation:

$$Y_{ij1} - Y_{ij0} = \gamma_0 + \gamma_1 X_{ij} + \gamma_2 Y_{ij0} + \gamma_3 X_{ij} Y_{ij0} + \varepsilon_{ij}. \tag{10}$$

But adding the pretest to both sides of Equation (10) yields the following:

$$Y_{ij1} = \gamma_0 + \gamma_1 X_{ij} + (1 + \gamma_2) Y_{ij0} + \gamma_3 X_{ij} Y_{ij0} + \varepsilon_{ij}. \tag{11}$$

Equation (11) is the same as Equation (5) for standard ANCOVA with a Treatment X Pretest interaction, with $b_2 = 1 + \gamma_2$. Therefore the treatment

effect γ_1 in Equations (10) and (11) is equivalent to ANCOVA's treatment effect b_1 in Equation (5). Omitting the auto-regressive term $\gamma_2 Y_{ij0}$ from Equation (10) is equivalent to fixing γ_2 to 0, which is usually nonsensical, since the pretest Y_{ij0} is one of the two components of the difference score being predicted. Fixing the slope coefficient γ_2 to 0 in Equation (10) is also equivalent to fixing the coefficient $(1+\gamma_2)$ to 1 in Equation (11), which makes the equations for ANCOVA and difference-in-difference identical. This is possible, however, only when the variance of the outcome scores is increasing over time or unless pretest scores predict posttest scores perfectly.

To add a Treatment X Pretest interaction to dual-centered ANCOVA in Equation (3), we apply the same steps as in Equations (5) through (9) for standard ANCOVA. In both cases, a significant interaction changes the unconditional marginal effects in Equations (1) and (2) to conditional effects that vary with pretest scores.

Adding a Treatment X Pretest interaction to Equation (3) for dual-centered ANCOVA yields:

$$Y_{ij1} - \widehat{\mu}_{j0} = \omega_0 + \omega_1 X_{ij} + \omega_2 (Y_{ij0} - \widehat{\mu}_{j0}) + \omega_3 X_{ij} (Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}. \tag{12}$$

Because dual-centered ANCOVA predicts the same treatment effect as difference-in-differences, the Treatment X Centered Pretest interaction can be interpreted in the same way as a Treatment X Pretest interaction in standard ANCOVA. Analyzing Equation (12) by itself yields the correct standard error for ω_3, according to our simulation (Lin, 2023). Equation (12) can be re-arranged to indicate how the treatment effect varies by the group-mean-centered pretest score (Lin, 2020).

$$Y_{ij1} - \widehat{\mu}_{j0} = (\omega_0 + \omega_2 [Y_{ij0} - \widehat{\mu}_{j0}]) + (\omega_1 + \omega_3 [Y_{ij0} - \widehat{\mu}_{j0}]) X_{ij} + \nu_{ij}. \tag{13}$$

Reciprocally, the effect of the group-mean-centered pretest score also varies by treatment condition:

$$Y_{ij1} - \widehat{\mu}_{j0} = (\omega_0 + \omega_1 X_{ij}) + (\omega_2 + \omega_3 X_{ij}) (Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}. \tag{14}$$

Equation (13) indicates that the effect of treatment on the pretest-group-mean-centered posttest at any group-mean-centered pretest score is:

$$\widehat{\omega}^*_{Tx} = \omega_1 + \omega_3 (Y_{ij0} - \widehat{\mu}_{j0}). \tag{15}$$

Reciprocally the effect of the group-mean-centered pretest on the centered posttest at either level of treatment according to Equation (14) is

$$\widehat{\omega}^*_{lag1} = \omega_2 + \omega_3 X_{ij}. \tag{16}$$

1.3 Illustrative Example

The following example estimates the effect of psychotherapy to treat depression in mothers from the Fragile Family & Child Wellbeing (FFCW) dataset. We selected this corrective action because its effectiveness has been documented in meta-analyses of hundreds of randomized trials (Cuijpers et al., 2023). Although these effect sizes shrink over time (Miguel et al., 2021) and in typical field implementations (Ormel, Hollon, Kessler, Cuijpers, & Monroe, 2022), there is no reason to think that such therapies are harmful on average.

We expect therapy to look more effective with difference-in-differences than with ANCOVA, as is typical for longitudinal analyses of corrective actions (Larzelere et al., 2018). We then illustrate how to use dual-centered ANCOVA to test the Treatment X Pretest interaction corresponding to the treatment estimate from the difference-in-differences model.

2 Methods

2.1 Participants.

The FFCW dataset started with baseline data on at-risk couples whose children were born from 1998 to 2000 in 20 large cities of the United States (Reichman, Teitler, Garfinkel, & McLanahan, 2001). It includes a wide range of data on household characteristics, physical and mental health, and parenting, first when the children were born (Time 1), and later when the children were approximately 1, 3, 5, and 9 years old (Times 2 to 5). The current example investigated the apparent effects of psychotherapy for maternal depression when their children were five years old, using data on maternal depression symptoms when their children were 5 and 9 years old (Time 4 and Time 5). At baseline (when the focal child was born), the 4566 mothers were 25.2 years old and had some college on average, and consisted of 21.0% White, 47.6% Black, 27.4% Hispanic, and 4.0% others. The sample size for this study consisted of the 3285 mothers with complete data on therapy for depression at Time 4 and on depression symptoms at Times 4 and 5. The data are available on the Open Science Framework Home website (`https://osf.io/532xt/?view_only=5857097b48034e7786a8933b4af22e3a`).

2.2 Measures

Depression treatment was based on maternal responses to questions about whether they had received any counseling or therapy in the past twelve months. "Yes" answers led to the question "Was this counseling or therapy for depression?" Mothers who reported receiving therapy for depression were contrasted with mothers who responded "No" to either of these questions.

Depression symptoms were assessed by maternal self-reports of relevant symptoms from the Composite International Diagnostic Interview--Short Form (CIDI-SF), Section A (Kessler, Andrews, Mroczek, Ustun, & Wittchen, 1998), a standardized survey instrument for assessing mental disorders. It uses two stem ques-

tions and four follow-up questions to identify possible eligibility for a Major Depressive Episode. Eligibility then led to eight symptom questions to determine depression severity. Sub-eligibility symptoms resulted in possible scores from 1 to 4. Four points were added to the number of the eight symptoms associated with a possible Major Depressive Episode. This produced a 13-point scale (0 to 12) for depression severity, with the majority of the scores being 0 (73.8% at Time 4; 73.9% at Time 5).

3 Results

Table 1 provides the means, standard deviations, and other descriptive statistics for therapy at Time 4 of the FFCW dataset and for depression symptoms at Times 4 and 5.

Table 1: Descriptive Statistics

	Treatment T4	N	Mean	SD	Minimum	Maximum
Depress T4	0	3078	1.54	3.36	0	12
	1	207	7.33	4.48	0	12
	Total	3285	1.9	3.72	0	12
Depress T5	0	3078	1.61	3.45	0	12
	1	207	5.1	4.94	0	12
	Total	3285	1.83	3.66	0	12

Note. T4 = Time 4 of the FFCW dataset. T5 = Time 5. Depress = Depression symptoms.

Prior to adding an interaction term, standard ANCOVA and difference-in-differences produced contradictory estimates of treatment effects, as is typical of longitudinal analyses of corrective actions (Larzelere et al., 2018). According to ANCOVA, therapy for depression led to more depression symptoms at Time 5 than predicted by initial symptoms at Time 4, b_1= 1.74, t (3284) = 6.59, $p <$ .001. In contrast, difference-in-differences indicated that depression symptoms decreased more following therapy than otherwise, γ_1= -2.31, t (3284) = -7.70, $p <$ .001. Because psychotherapy for depression has been shown to be effective in many randomized trials (Cuijpers et al., 2023), difference-in-differences may be less biased against corrective actions than ANCOVA. Most researchers, however, would also want to know whether these treatment effects vary by the level of presenting depression symptoms. We will illustrate the use of dual-centered ANCOVA to test a Treatment X Pretest interaction in difference-in-differences after summarizing a Treatment X Pretest interaction in standard ANCOVA for comparative purposes.

3.1 Standard ANCOVA

Analyzing the data with standard ANCOVA led to the following result from Equation (5):

$$Y_{ij1} = 1.13 + 2.54X_{ij} + .314Y_{ij0} - .119X_{ij}Y_{ij0} + e_{ij}. \tag{17}$$

indicating that therapy predicted worsening depression symptoms than controls, b_1= 2.54, $p < .001$, a harmful-looking effect that was reduced for those with worse initial symptoms, b_3= -.119, $p < .05$. Plugging coefficients into Equation (8) gives the magnitude of the estimated treatment effect for each pretest score:

$$\widehat{b}^*_{Tx} = 2.54 + (-.119)Y_{ij0}. \tag{18}$$

This signifies that the harmful-looking effect of therapy varied from 2.54 for those with pretest depression scores of 0 to a reduced harmful-looking treatment effect of only 1.11 for those with maximum pretest scores of 12. These effect sizes varied around the average treatment effect of 1.74 from standard ANCOVA before adding the interaction term.

Figure 1 uses picked-points analysis to show the conditional treatment effects predicted at the mean pretest scores for the treatment and control groups and at the maximum depression score (Lin, 2020). Figure 4 in Appendix A shows the 95% confidence intervals of these coefficients and the significance of these treatment effects at each pretest score. Next we illustrate similarities and differences in testing the same Treatment X Pretest interaction within difference-in-differences.

3.2 Difference-in-Differences via Dual-Centered ANCOVA

Using Equation (12), the results from dual-centered ANCOVA from the same data after centering all depression scores around their pretest group means, $Y_{ijt} - \widehat{\mu}_{j0}$, are

$$Y_{ij1} - \widehat{\mu}_{j0} = .10 + (-2.30)X_{ij} + .314(Y_{ij0} - \widehat{\mu}_{j0}) + (-.119)X_{ij}(Y_{ij0} - \widehat{\mu}_{j0}) + \nu_{ij}, \tag{19}$$

indicating that, for those with initial depression symptoms at their group mean ($Y_{ij0} - \widehat{\mu}_{j0} = 0$), depression symptoms decreased more for women in therapy than controls, ω_1= -2.30, $p < .001$, a beneficial-looking effect that was enhanced further for those with worse initial symptoms, ω_3= -.119, $p < .05$.

Using Equation (15), the estimated effect of therapy on the pretest-group-mean-centered posttest for any group-mean-centered pretest score was

$$\widehat{\omega}^*_{Tx} = -2.30 + (-.119)(Y_{ij0} - \widehat{\mu}_{j0}). \tag{20}$$

This signifies that therapy led to steeper decreases in depression symptoms than in controls, with that beneficial-looking effect varying from -2.12 for the minimum possible centered pretest score for the comparison group (0 – 1.5 = -1.5,

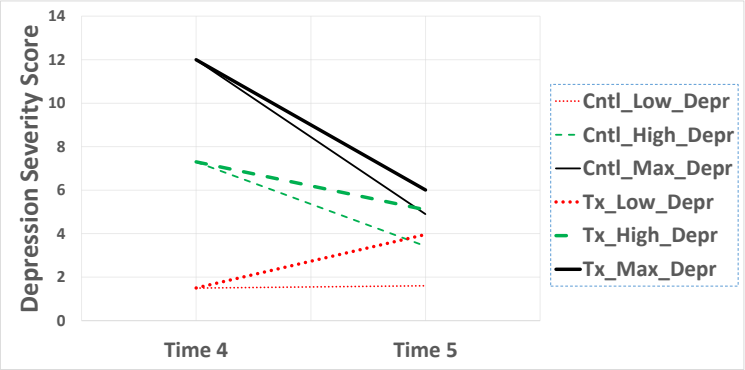


Figure 1: Predicted changes from Time 4 to Time 5 at different pretest scores according to standard ANCOVA (Low = mean pretest for controls, High = mean pretest for treatment, Max = maximum depression score) to illustrate the Treatment X Pretest interaction.

subtracting their mean pretest score) to a stronger beneficial-looking treatment effect of -2.86 for the maximum possible centered pretest score for the treatment group (12 – 7.3 = 4.7, subtracting their mean pretest score). These effect sizes varied around the average treatment effect of -2.31 from difference-in-differences before adding the interaction term.

This result and its confidence intervals are displayed in Figure 5 of Appendix A (Lin, 2020). Figure 2 uses picked-points analysis to illustrate how estimated treatment effects varied across the range of centered pretest scores that are possible in both treatment and comparison groups. Figure 3 illustrates the same treatment effects at the same picked pretest points after decentering all depression scores. This illustrates a potential problem with difference-in-differences in that its parallel-slopes assumption is less tenable at minimum and maximum scores. When centered pretest scores were at the minimum for the control group, they could not decrease further for that group, but could decrease further in the treatment group (a floor effect for the control group). In this case, however, this floor-effect bias is in the opposite direction of the Treatment X Pretest interaction and therefore does not invalidate it. (Therapy at a centered pretest of -1.5 [originally 5.8] decreased to a posttest mean of -2.49 [4.81 on original scale]. Controls at a centered pretest of -1.5 [originally 0.0] could not decrease, artificially increasing the extent to which therapy looked relatively effective at low depression levels. If controls could have decreased their centered depression pretest scores, the differential effectiveness would have been even smaller at low

depression levels, increasing the Treatment X Pretest interaction even more.) The ceiling effect bias was in the same direction as the Treatment X Pretest interaction, but was relatively minor as only 15 women in the therapy group had maximum posttest depression scores of 12 (7.2% of the therapy group, vs. 76.2% of controls with minimum posttest scores of 0).

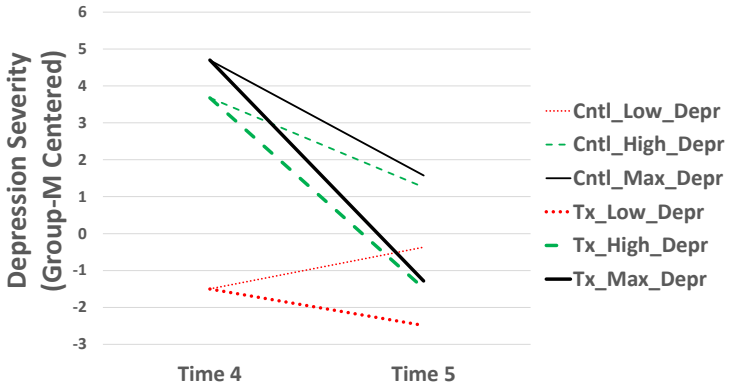


Figure 2: Predicted simple change scores from Time 4 to Time 5 for treatment vs. comparison groups at three levels of group-mean centered pretest scores, based on dual-centered ANCOVA (Low = minimum possible centered score for controls; High = one SD above the group mean pretest scores; Max = maximum possible centered pretest score for treatment).

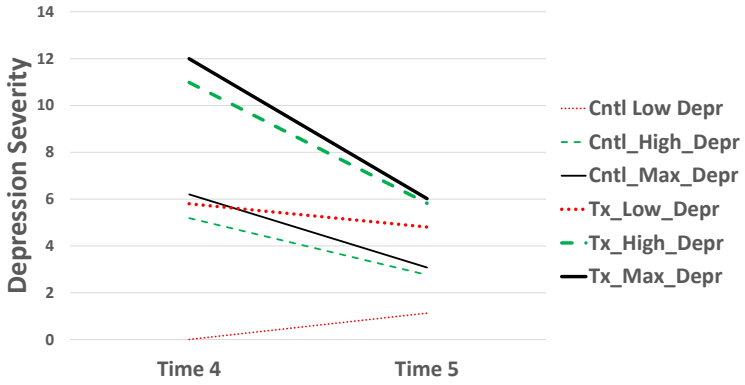


Figure 3: Predicted simple change scores from Time 4 to Time 5 for treatment vs. comparison groups at three levels of group-mean centered pretest scores according to dual-centered ANCOVA after decentering all scores (Low, High, & Max defined as in Figure 2).

4 Discussion

ANCOVA-type controls have been shown to be biased in longitudinal analyses (Berry & Willoughby, 2016; Hamaker et al., 2015; Hoffman, 2015), usually biased against corrective actions such as medical treatments and psychotherapy (Larzelere et al., 2018). This study demonstrates a novel way to overcome one disadvantage of the main alternative, difference-in-differences, which otherwise cannot test Treatment X Pretest interactions without changing the treatment effect to the estimate from ANCOVA. This innovation takes advantage of the fact that centering all longitudinal data around pretest group means makes the treatment effects of ANCOVA equal to estimates from difference-in-differences (Lin & Larzelere, 2020). This is called dual-centered ANCOVA in two-occasion analyses, which is used herein to test a Treatment X Pretest interaction corresponding to a difference-in-differences model.

We do not know of a better way to test Treatment X Pretest interactions in difference-in-differences. Without Treatment X Pretest interactions, difference-in-differences are limited to assuming that the estimated treatment effects are identical at every pretest score, an untenable assumption without sufficient evidence. When regression slopes are heterogeneous across treatment conditions, the effect of treatment also varies with the pretest score. For this situation, (Huitema, 2011, Chapter 11) showed how to calculate the conditional treatment effect at each level of pretest scores in standard ANCOVA. The lack of a parallel way to test Treatment X Pretest interactions in difference-in-differences appears

to be a limitation in such analyses, one that can be overcome after centering all data on the pretest group means.

Unless ANCOVA clearly produces a less-biased treatment estimate in longitudinal analyses, difference-in-differences should be used to test the robustness of the estimated treatment effect (Duncan et al., 2014), if not a less-biased estimate. The least-biased estimate is generally the one whose assumptions are best satisfied. From our experience, it is helpful to compare the plausibility of the no-treatment effect implied by their respective null hypotheses. A no-treatment effect in difference-in-differences assumes that the groups' average trends from pretest to posttest will be parallel to each other, with no shrinkage of the difference between group means. In contrast, the null hypothesis in ANCOVA assumes that any group difference on the pretest will spontaneously shrink from pretest to posttest according to regression toward the grand mean. This shrinkage is plausible in randomized trials when initial differences on the pretest group means are due only to random fluctuations (i.e., no true difference between the pretest group means). ANCOVA is also unbiased if the covariates fully determine treatment group assignment (van Breukelen, 2013). In many other applications, however, pretest group means reflect true differences as well as random fluctuations, and the covariates do not fully explain treatment assignment. The remaining bias is recognized as *residual confounding* by epidemiologists (Rothman, Greenland, & Lash, 2008), which often makes corrective actions such as therapy for depression look more harmful than they are (Larzelere et al., 2018). In contrast, difference-in-differences' treatment estimates are not biased by true differences that do not change from pretest to posttest nor by measurement error in the pretest, but it has its own biases in non-randomized studies (e.g., any variations from the parallel-slopes assumption not due to the treatment effect). Unless the original ANCOVA is less biased, difference-in-differences provides either a less biased treatment estimate or a test of that estimate's robustness (Duncan et al., 2014). Dual-centered ANCOVA can then be used to test a Treatment X Pretest interaction within difference-in-differences.

Acknowledgments

We gratefully acknowledge funding from research grant #R03 HD107307 from the National Institute of Child Health and Human Development and from the Oklahoma State University Foundation.

References

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton University Press. doi: https://doi.org/10.1515/9781400829828

Berry, D., & Willoughby, M. T. (2016). On the practical interpretability of cross-lagged panel models: Rethinking a developmental workhorse. *Child Development*, *88*(4), 1186–1206. doi: https://doi.org/10.1111/cdev.12660

Brorsen, B. W., Lin, H., & Larzelere, R. E. (2025). Critique of enhanced power claimed for Quasi-ANCOVA and Dual-Centered ANCOVA. *PLOS ONE*, *20*(1), e0317860. doi: https://doi.org/10.1371/journal.pone.0317860

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Papola, D., ... Karyotaki, E. (2023). Psychological treatment of depression: A systematic overview of a 'meta-analytic research domain'. *Journal of Affective Disorders*, *335*, 141–151. doi: https://doi.org/10.1016/j.jad.2023.05.011

Ding, P., & Li, F. (2019). A bracketing relationship between difference-in-differences and lagged-dependent-variable adjustment. *Political Analysis*, *27*(4), 605–615. doi: https://doi.org/10.1017/pan.2019.25

Duncan, G. J., Engel, M., Claessens, A., & Dowsett, C. J. (2014). Replication and robustness in developmental research. *Developmental Psychology*, *50*(11), 2417–2425. doi: https://doi.org/10.1037/a0037996

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. doi: https://doi.org/10.1037/a0038889

Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. Routledge.

Huitema, B. E. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. Wiley. doi: https://doi.org/10.1002/9781118067475

Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Kessler, R. C., Andrews, G., Mroczek, D., Ustun, B., & Wittchen, H. (1998). The world health organization composite international diagnostic interview short-form (cidi-sf). *International Journal of Methods in Psychiatric Research*, *7*(4), 171–185. doi: https://doi.org/10.1002/mpr.47

Larzelere, R. E., Lin, H., Payton, M. E., & Washburn, I. J. (2018). Longitudinal biases against corrective actions. *Archives of Scientific Psychology*, *6*(1), 243–250. doi: https://doi.org/10.1037/arc0000052

Lin, H. (2020). Probing two-way moderation effects: A review of software to easily plot Johnson-Neyman figures. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(3), 494–502. doi: https://doi.org/10.1080/10705511.2020.1732826

Lin, H. (2023). *Simulation test of standard errors of regression coefficients in DC-ANCOVA with Treatment X Pretest interaction* (Tech. Rep.). Department of Human Development and Family Science, Oklahoma State University.

Lin, H., & Larzelere, R. (2024). Lord's paradox illustrated in three-wave longitudinal analyses: Cross lagged panel models versus linear latent growth models. *Journal of Behavioral Data Science*, *4*(2), 51–63. doi: https://doi.org/10.35566/jbds/lin

Lin, H., & Larzelere, R. E. (2020). Dual-centered ANCOVA: Resolving contradictory results from Lord's paradox with implications for reducing bias

in longitudinal analyses. *Journal of Adolescence*, *85*(1), 135–147. doi: https://doi.org/10.1016/j.adolescence.2020.11.001

Miguel, C., Karyotaki, E., Ciharova, M., Cristea, I. A., Penninx, B. W., & Cuijpers, P. (2021). Psychotherapy for comorbid depression and somatic disorders: a systematic review and meta-analysis. *Psychological Medicine*, *53*(6), 2503–2513. doi: https://doi.org/10.1017/s0033291721004414

Ormel, J., Hollon, S. D., Kessler, R. C., Cuijpers, P., & Monroe, S. M. (2022). More treatment but no less depression: The treatment-prevalence paradox. *Clinical Psychology Review*, *91*, 102111. doi: https://doi.org/10.1016/j.cpr.2021.102111

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review*, *25*(1), 221–247. doi: https://doi.org/10.2307/2648877

Reichardt, C. S. (2019). *Quasi-experimentation: A guide to design and analysis*. Guilford.

Reichman, N. E., Teitler, J. O., Garfinkel, I., & McLanahan, S. S. (2001). Fragile families: Sample and design. *Children and Youth Services Review*, *23*(4–5), 303–326. doi: https://doi.org/10.1016/s0190-7409(01)00141-4

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3rd ed.). Wolters Kluwer.

Schueler, B. E., Asher, C. A., Larned, K. E., Mehrotra, S., & Pollard, C. (2021). Improving low-performing schools: A meta-analysis of impact evaluation studies. *American Educational Research Journal*, *59*(5), 975–1010. doi: https://doi.org/10.3102/00028312211060855

Usami, S., Todo, N., & Murayama, K. (2019). Modeling reciprocal effects in medical research: Critical discussion on the current practices and potential alternative models. *PLOS ONE*, *14*(9), e0209133. doi: https://doi.org/10.1371/journal.pone.0209133

van Breukelen, G. J. P. (2013). ANCOVA versus CHANGE from baseline in nonrandomized studies: The difference. *Multivariate Behavioral Research*, *48*(6), 895–922. doi: https://doi.org/10.1080/00273171.2013.831743

A Supporting Information

Whereas Figures 1, 2, and 3 in the main article use picked-points analysis to illustrate the significant Treatment X Pretest interaction at selected pretest scores, the following Supporting Figures illustrate the magnitude and significance of the estimated treatment effect for each possible pretest score. These plots are based on the Johnson-Newman technique, including 95% confidence intervals for the estimated treatment effect at each pretest score (Lin, 2020). The two figures illustrate the contradictory results from the two basic change-score analyses that are typical of longitudinal analyses of corrective actions (Larzelere et al., 2018). According to ANCOVA, therapy for maternal depression at Time 4 in the Fragile Families dataset appears to result in increased depression symptoms four years

later at Time 5, controlling for depression symptoms at Time 4 (Supporting Figure 4). In contrast, dual-centered ANCOVA duplicates difference-in-differences by indicating that therapy for maternal depression reduces depression scores more than for the comparison group (Supporting Figure 5). In both cases, therapy appears to be significantly more effective at high levels of initial depression than at low levels of initial depression (reducing the harmful-looking effect in standard ANCOVA, but increasing the beneficial-looking effect in dual-centered ANCOVA).

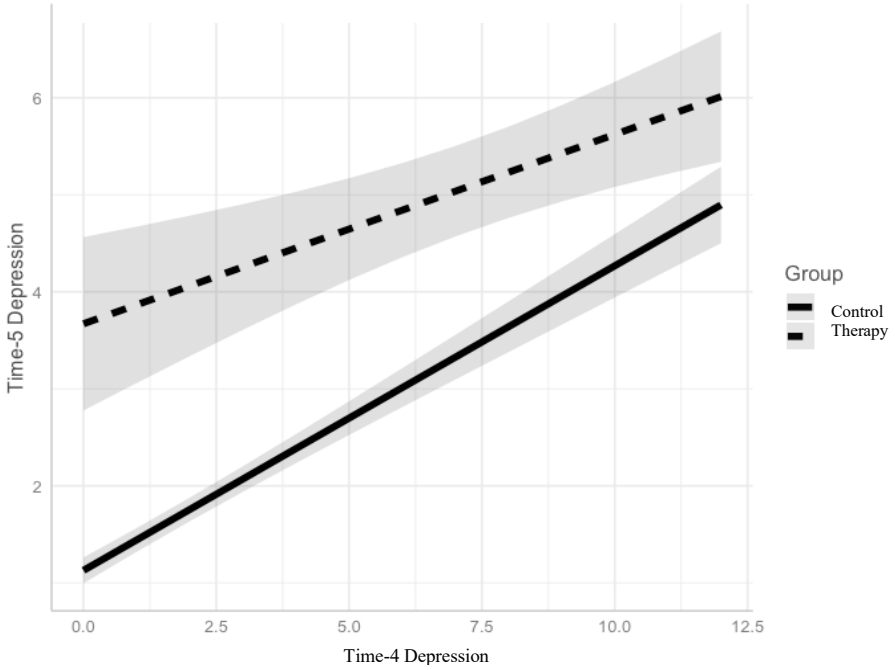


Figure 4: Predicted posttest depression scores for each pretest depression score for Therapy (dashed upper line) or Control (solid lower line) according to standard ANCOVA.

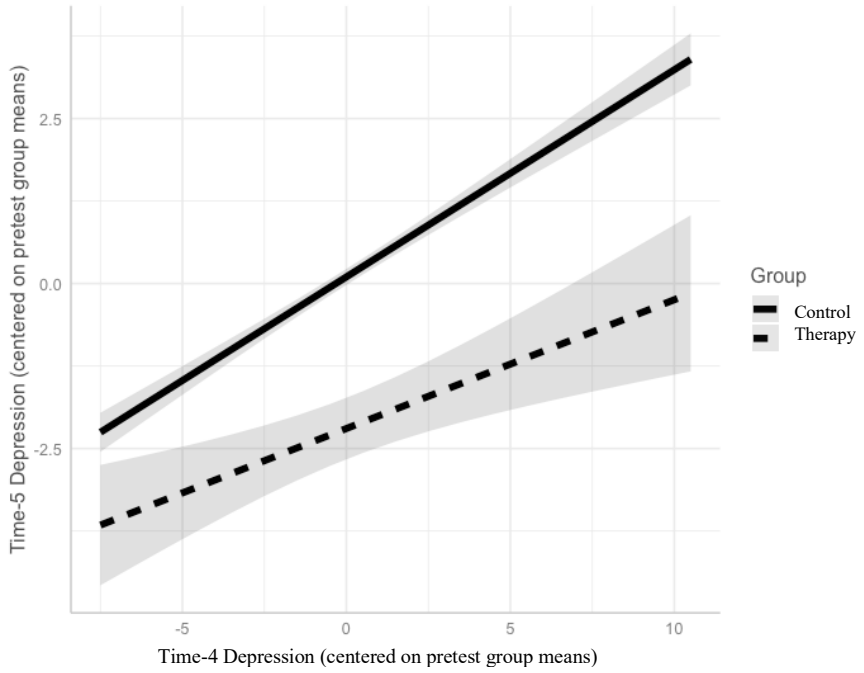


Figure 5: Predicted posttest depression score at Time-5 (centered on pretest group means at Time-4) for each pretest depression score (centered on pretest group means) according to dual-centered ANCOVA for Therapy (dashed lower line) and Control (solid upper line).

Journal of Behavioral Data Science, 2025, *5* (1), 67–102.
DOI: https://doi.org/10.35566/jbds/caoyc

Machine Learning Approaches for Mental Illness Detection on Social Media: A Systematic Review of Biases and Methodological Challenges

Yuchen Cao[1,⋆], Jianglai Dai[2,⋆], Zhongyan Wang[3], Yeyubei Zhang[4], Xiaorui Shen[1], Yunchong Liu[4], and Yexin Tian[5,⋆⋆]

[1] Khoury College of Computer Science, Northeastern University, USA
[2] Department of EECS, University of California, Berkeley, USA
[3] Center for Data Science, New York University, USA
[4] School of Engineering and Applied Science, University of Pennsylvania, USA
[5] Georgia Institute of Technology, College of Computing, USA

Abstract. The global increase in mental illness requires innovative detection methods for early intervention. Social media provides a valuable platform to identify mental illness through user-generated content. This systematic review examines machine learning (ML) models for detecting mental illness, with a particular focus on depression, using social media data. It highlights biases and methodological challenges encountered throughout the ML lifecycle. A search of PubMed, IEEE Xplore, and Google Scholar identified 47 relevant studies published after 2010. The Prediction model Risk Of Bias ASsessment Tool (PROBAST) was utilized to assess methodological quality and risk of bias. The review reveals significant biases affecting model reliability and generalizability. A predominant reliance on Twitter (63.8%) and English-language content (over 90%) limits diversity, with most studies focused on users from the United States and Europe. Non-probability sampling methods (approximately 80%) limit representativeness. Only 23% of studies explicitly addressed linguistic nuances like negations, crucial for accurate sentiment analysis. Inconsistent hyperparameter tuning was observed, with only 27.7% properly tuning models. About 17% did not adequately partition data into training, validation, and test sets, risking overfitting. While 74.5% used appropriate evaluation metrics for imbalanced data, others relied on accuracy without addressing class imbalance, potentially skewing results. Reporting transparency varied, often lacking critical methodological details. These findings highlight the need to diversify data sources, standardize preprocessing protocols, ensure consistent model development practices, address class imbalance, and enhance reporting transparency. By overcoming these challenges, future research can develop more robust and generalizable ML models for depression

⋆ These authors contributed equally to this work.
⋆⋆ Corresponding author

detection on social media, contributing to improved mental health outcomes globally.

Keywords: Mental illness · Social media · Bias evaluation with PROBAST · Systematic review · Machine learning and deep learning

1 Introduction

Mental health disorders, including depression, represent a critical global health challenge, impacting approximately 1 in 8 people worldwide—approximately 970 million individuals in 2019 (WHO, 2023). Depression, one of the most prevalent mental health conditions, affects over 280 million individuals globally, including around 23 million children and adolescents. The COVID-19 pandemic has further exacerbated mental health issues, with notable increases in depression and anxiety observed during this period (WHO, 2023). The prevalence of mental health conditions, especially depression, highlights an urgent need for innovative detection methods and interventions. Early identification can lead to more effective treatment outcomes, alleviating the burdens placed on individuals, their families, and healthcare systems (Kessler et al., 2017).

In today's digital age, social media platforms such as Twitter, Facebook, and Reddit play a central role in daily life for millions of people. Studies have shown that individuals often openly express their thoughts, emotions, and mental states on Twitter, making it a valuable platform for examining mental health trends and developing tools for detection and intervention (De Choudhury, Counts, & Horvitz, 2013). The extensive user-generated content on these platforms provides a unique opportunity for mental health research, enabling the real-time analysis of linguistic patterns and behavioral trends, and providing insights that may otherwise be inaccessible (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017).

Advancements in machine learning and deep learning have significantly enhanced the ability to process and analyze large-scale datasets. These technologies are particularly suited for handling the complex and nuanced data found on social media, as they identify patterns and make predictions based on textual and behavioral cues. This capability offers practical tools for mental health detection, allowing researchers to develop models that can potentially identify at-risk individuals based on their social media activity (Shatte, Hutchinson, & Teague, 2019). By leveraging algorithms capable of learning from such diverse and rich datasets, researchers are able to develop models that contribute to early intervention efforts in mental health care.

1.1 Overview of Historical Studies on Machine Learning Approaches for Mental Health Detection in Social Media

A growing body of research has explored the application of machine learning techniques to detect depression through social media platforms. Approaches range from traditional machine learning techniques such as logistic regression

and support vector machines to advanced deep learning models and ensemble methods—have been employed to classify user posts and predict mental health conditions based on linguistic features, patterns, and metadata (Calvo, Milne, Hussain, & Christensen, 2017; De Choudhury et al., 2013; Yazdavar et al., 2020). Platforms like Twitter, Facebook, and Reddit are frequently utilized due to their large user bases and the accessibility of publicly available text-based data. In contrast, TikTok, with its short-video format, provides a distinct medium that captures audiovisual cues such as tone, facial expressions, and gestures, providing researchers with additional dimensions for understanding mental health dynamics.

One of the most common approaches within this research involves sentiment analysis, which aims to determine the emotional tone of user-generated content. By assessing positive, negative, or neutral sentiment (Kumar, Khan, & Kalra, 2020), researchers attempt to correlate language patterns with indicators of depression. For instance, several studies have examined the use of first-person singular pronouns and negative emotion words as potential depression signals (Rude, Gortner, & Pennebaker, 2004).

Despite promising results, multiple challenges remain. First, many studies suffer from limited generalizability due to small or homogeneous samples that may not represent the broader population. Data bias is a significant concern, stemming from the overrepresentation of certain demographic groups or linguistic communities while underrepresenting others (Olteanu, Castillo, Diaz, & Kiciman, 2019). Moreover, the dispersion of research in advanced machine learning methods for mental health detection across the literature, combined with a lack of robust sampling methods and standardized protocols, impedes the reliability of findings. Additionally, insufficient handling of complex linguistic nuances, such as context-dependent meanings, further limits the effectiveness of these detection efforts (Calvo et al., 2017).

1.2 Research Gaps and Objectives of the Current Study

While individual studies have provided valuable insights into the application of machine learning for mental health detection, significant gaps persist in the literature. These include the broader implications of biases and limitations across studies and the lack of comprehensive reviews consolidating the effectiveness of machine learning models (Calvo et al., 2017). Additionally, existing research does not consistently address methodological challenges across different stages of machine learning applications, such as sampling, preprocessing, model development, and evaluation (Thieme, Belgrave, & Doherty, 2020). Therefore, a systematic review is essential to unify findings and evaluate the pervasiveness and impact of biases across studies.

To address these gaps, this study aims to conduct a systematic review that synthesizes and evaluates existing machine-learning models for detecting depression on social media. The specific objectives are:

1. Examine the effectiveness of machine learning and deep learning models by focusing on bias present in sampling, data preprocessing, model construction,

fine-tuning, evaluation, and comparison, as well as the challenges associated with model generalizability across different social media platforms.
2. Explore methodological challenges, including those unique to mental health detection—such as handling class imbalance where depressive posts are the minority and preprocessing for sentiment analysis involving negations. Additionally, more general machine learning challenges, like improving model evaluation techniques and addressing data biases related to language and platform-specific factors, also persist. It is important to recognize that most of these biases are unintentional, either from practical challenges or from a lack of standardized guidelines for applying machine learning to mental health detection. By addressing these biases, the review aims to provide insights and strategies to mitigate these unintended biases, advancing the development of more reliable and generalizable models.
3. Provide recommendations for future research to enhance the reliability and applicability of machine learning models in mental health detection. These insights aim to inform strategies that improve early intervention efforts and contribute to the development of more robust, generalizable, and ethically sound machine learning applications. In doing so, the review seeks to provide guidance that fills the gap left by current practice, where a lack of formal guidelines has sometimes led to the persistence of unintended biases.

By addressing these objectives, this review seeks to provide a comprehensive understanding of the current practices and limitations within the field. The findings aim to guide future research and development into more robust, generalizable, and ethical applications of machine-learning models for mental health detection using social media data. In the following sections, we will first examine the methodologies and models used across studies, followed by an analysis of common biases and limitations. We will conclude with a discussion on best practices and recommendations for advancing the field.

2 Methodology

2.1 Search Strategy

The search focused on publications on machine learning and deep learning models for detecting depression and other mental health conditions using social media data, primarily from platforms like Twitter, Facebook, and Reddit. To identify relevant studies, a systematic search was conducted across multiple academic databases including PubMed, ACM, and IEEE Xplore, with Google Scholar used for additional sources. The search included combinations of 'machine learning,' 'deep learning,' 'artificial intelligence,' 'social media,' 'Twitter,' 'Facebook,' 'Reddit,' 'depression,' 'sentiment analysis,' and 'mental health.' To broaden the scope of the search, additional terms such as 'anxiety,' 'mental disorders,' 'neural networks,' and 'supervised learning' were included. The search process was carried out from June to July 2024.

The search strategy was structured around three main categories: social media platforms (e.g., 'social media,' 'Twitter,' 'Facebook,' 'Reddit'), mental health topics (e.g., 'depression,' 'sentiment analysis'), and machine learning and data analysis techniques (e.g., 'machine learning,' 'deep learning,' 'artificial intelligence'). The comprehensive search query[1] formulated for this review is:

```
((social media OR 'Twitter' OR 'Facebook' OR 'Reddit')
   AND ('depression' OR 'sentiment analysis' OR '
   mental health' OR 'anxiety' OR 'mental disorders')
   AND ('machine learning' OR 'deep learning' OR '
   artificial intelligence' OR 'neural networks' OR '
   supervised learning'))
```

2.2 Inclusion and Exclusion Criteria

To be included in this review, studies needed to meet the following criteria:

- **Publication Date:** Studies published after 2010 were included to ensure contemporary research and methods were considered.
- **Language:** Only studies published in English were included.
- **Research Focus:** The study must use machine learning or deep learning models for detecting depression or other mental health conditions, with a particular focus on analyzing data from social media platforms like Twitter, Facebook, or Reddit.
- **Study Type:** The review included primary research articles, specifically those that involved data-driven analyses.

Studies were excluded based on the following criteria:

- **Publication Type:** Review articles, systematic reviews, conference abstracts, editorials, opinion pieces, and non-peer-reviewed literature were excluded.
- **Scope:** Studies not directly focused on mental health detection through social media or not employing machine learning models were excluded.
- **Methodology:** Studies that did not directly employ machine learning or deep learning and applied solely on quantitative analysis were excluded.

[1] The search query used the term 'Twitter' to align with the naming convention at the time of the review, which covered literature up to June/July 2024. Twitter was rebranded as 'X' after this period. The search algorithm was adjusted to include both 'Twitter' and 'X' where applicable to ensure coverage of relevant results under the new name. However, no additional papers published up to June/July 2024 were identified using the term 'X.' Notably, one manuscript, Jamali, Berger, and Spiteri (2023), included both terms.

2.3 Study Selection Process

The selection process was conducted in three stages to ensure a rigorous and unbiased review of relevant studies. The process, which followed the search process that concluded in July 2024, lasted until August 2024.

1. **Initial Identification:** Duplicates were removed, and an initial screening was conducted based on titles and abstracts to filter out irrelevant studies. All authors contributed to this step.
2. **Title and Abstract Screening:** An independent review was conducted by two authors, Y.T. and J.D., to assess the relevance of studies based on their titles and abstracts. Both authors have expertise in machine learning and mental health research, ensuring a thorough evaluation. Any discrepancies in their assessments were discussed and resolved to ensure a consistent screening process.
3. **Full-Text Screening:** A comprehensive review of the full texts of selected studies was conducted. Any disagreements were resolved through discussion to maintain an unbiased selection process. Additionally, relevant studies identified through references in full-text articles were included for consideration. All authors contributed to this step.

2.4 Data Extraction and Analysis

The data extraction process involved using a standardized form to systematically capture detailed information from each selected study. The form included fields to record author names, study titles, publication journals, and publication years. It also documented the study designs, settings, and sample sizes, alongside specific inclusion and exclusion criteria. In addition, the form provided details on the machine learning models employed, the social media platforms analyzed (such as Twitter, Facebook, and Weibo), and the primary and secondary outcomes measured. Additionally, performance metrics, including accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic (AUROC)[2], which were collected when applicable.

Special attention was given to identifying potential sources of bias, study limitations, and funding sources, ensuring a comprehensive overview of each study's context and reliability. Table 1 below outlines the key categories and details included in the data extraction form.

[2] Accuracy measures the proportion of correctly classified instances among all instances. Precision focuses on the correctness of positive predictions, while recall measures the ability to identify actual positive cases. Both F1-score and Area Under the Receiver Operating Characteristic Curve (AUROC) are composite metrics that combine aspects of precision and recall to evaluate the performance of models. A detailed explanation of these metrics is provided in Section 3.7

Table 1: Key Data Extraction Categories for Systematic Review.

Category	**Details**
Study Details	Title, Authors, Year of Publication, Journal or Source, DOI or URL
Research Objectives	Purpose of the Study, Research Questions or Hypotheses
Methodological Aspects	Study Design, Settings, Sample Sizes, Inclusion and Exclusion Criteria, Data Collection Methods, ML/DL Models Employed
Criteria Applied	Data included, e.g., publicly available tweets, specific language posts. Data excluded, e.g., private or insufficiently detailed posts
Performance Metrics	Metrics Used (e.g., Accuracy, Precision, Recall, F1-score, AUROC, etc.)
Bias Evaluation	Data Collection and Preprocessing, Model Development and Tuning, Model Evaluation and Reporting
Additional Information	Confounding Factors, Study Limitations, Ethical Considerations, Funding Sources

2.5 Analytical Methods Used to Synthesize Findings

The extracted data were synthesized using a narrative approach, systematically examining each aspect of the machine learning lifecycle—sampling, data preprocessing, model construction, tuning, evaluation, comparison, and reporting—across the selected studies. This synthesis involved reviewing how studies approached sampling and data preprocessing, examining their approaches to model construction and tuning, and assessing model evaluation and comparison based on quantitative metrics such as accuracy, precision, recall, F1 scores, and AUROCs. For each stage, we summarized the methodologies employed by the studies and identified potential biases with established tools. This comprehensive approach provided insights into the current state of research, highlighting areas for future investigation to enhance the accuracy, generalizability, and applicability of machine learning models in this field.

2.6 Systematic Review Registration

This systematic review has been registered in the International Prospective Register of Systematic Reviews (PROSPERO) database under the title *Systematic Review of Machine Learning and Deep Learning Algorithms for Detecting Depression and Mental Health Conditions on Social Media* (ID: 617763). The registration has been approved.

3 Results

3.1 Study Selection

The search process began by identifying a total of 328 studies from three databases: 192 from Google Scholar, 101 from PubMed, and 35 from IEEE Xplore. After removing 57 duplicate studies, 271 unique titles and abstracts were retained for screening. During the title and abstract review, 174 studies were excluded. These exclusions were due to issues related to methodology (53 studies), scope (77 studies), and publication type (44 studies). This left 97 full-text studies to be reviewed in detail.

Upon reviewing the full texts, another 50 publications were excluded. The reasons for exclusion included being outside the scope or irrelevant (32 studies), methodological concerns (6 studies), publication type (9 studies), and unavailability (3 studies). Ultimately, 47 studies were included in the final narrative synthesis.

Figure 1 outlines how the initial pool of studies was refined down to the most relevant research for inclusion.

3.2 Characteristics of Included Studies

In this systematic review, key details of all 47 included studies, as summarized in Table 1, are provided in an online supplementary document. The majority of studies focused on Twitter (32 studies), Reddit (8 studies), and Facebook (7 studies). Additionally, one study examined Blued, a platform for MSM communities, and another focused on Indian social networking sites (SNS). Notably, 8 studies (17.02%) analyzed data from multiple platforms. Upon further examination, the datasets used in these 47 studies were found to be independent. The most commonly used models included traditional machine learning approaches such as Support Vector Machines (SVM) (19 studies), tree-based models (e.g., Decision Trees in 6 studies, Random Forests in 13 studies, and eXtreme Gradient Boosting (XGBoost) in 3 studies), and Logistic Regression (6 studies). Some studies also utilized deep learning models, including Convolutional Neural Networks (CNNs) (9 studies), Long Short-Term Memory (LSTM) networks (5 studies), and Bidirectional Encoder Representations from Transformers (BERT) (9 studies) for depression detection.

3.3 Methodological Quality and Risk of Bias

The risk of bias in the studies included in this systematic review was assessed using the Prediction model Risk Of Bias Assessment Tool (PROBAST, Wolff et al., 2019). PROBAST is a structured tool designed to assess the risk of bias and applicability of prediction models. It evaluates four key domains: participants, predictors, outcomes, and analysis, ensuring methodological rigor in studies. This tool provides a systematic framework for identifying biases and limitations

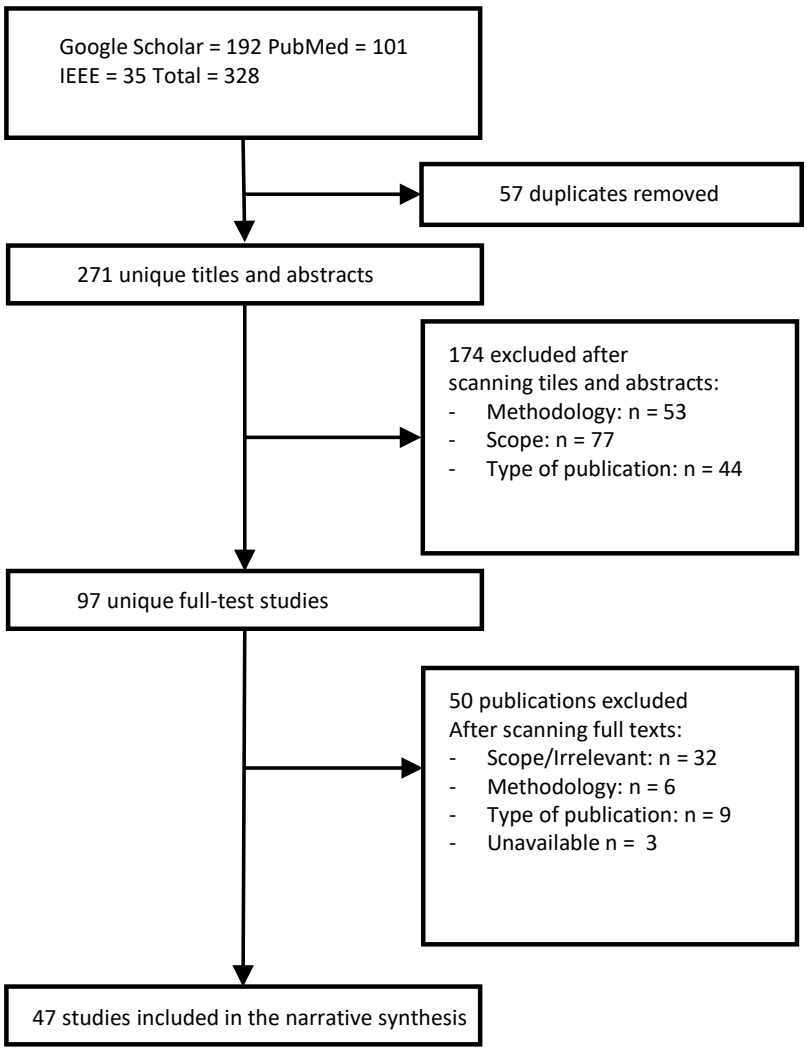


Figure 1: PRISMA Flow Diagram of Study Selection Process for Systematic Review on Machine Learning Models for Depression Detection Using Social Media Data

in prediction models, offering critical insights into their validity and applicability. PROBAST is particularly relevant in this systematic review as it allows for a comprehensive assessment of potential methodological biases throughout the machine learning lifecycle, including data collection, preprocessing, model development, and evaluation. By identifying biases in these areas, the tool supports a rigorous evaluation of the reliability and generalizability of machine learning

models used for mental health detection on social media. In addition to its role in assessing the risk of bias, PROBAST was used to evaluate transparency and completeness in the reporting of study methodologies and findings. The assessment covered 20 structured questions across the four domains, as detailed in Table 2. By incorporating PROBAST, this review identifies methodological weaknesses in the included studies, assesses their implications for the validity of findings, and evaluates the overall applicability of machine learning models used for mental health detection on social media. This ensures a thorough understanding of bias and enhances the reliability of the review's conclusions.

Table 2: Bias Evaluation Questions for Each Domain.

Domain	**Evaluation Questions**
Sample Selection and Representativeness	Q1. What is the sample used in this study, including the platform, sampling criteria, and sampling method? Q2. Does the sample represent the target population of social media users or posts?
Data Preprocessing	Q3. Did the study specify its approach to handling negative words when using traditional or machine learning methods for sentiment analysis?
Model Development	Q4. Did this study report hyperparameters? Q5. If reported, did this study tune (optimize) hyperparameters or use default settings? Q6. If tuned hyperparameters in this study, was this done on all models mentioned in the study?
Model Evaluation	Q7. Did the study divide the dataset into training, validation, and test sets, and were the reported metrics based only on training data? Q8. What evaluation metric was used in this study? Q9. Is the evaluation metric appropriate for this context (i.e., class-imbalanced settings)? Q10. If the study used accuracy as an evaluation metric, did it mention preprocessing steps to address class imbalance?

3.4 Sample Selection and Representativeness (Q1 & Q2)

The reviewed studies employed diverse sampling methods across various social media platforms, primarily focusing on Twitter (63.8%) with additional data from Reddit (23.5%), Facebook (8.5%), and other social media (2.1%). Most studies (around 80%) used non-probability sampling techniques, such as convenience sampling or keyword filtering, often utilizing APIs (e.g., Twitter API,

Reddit API) to filter posts by specific mental health-related keywords like 'depression' or '#MentalHealth,' or leveraging pre-existing datasets from repositories like Kaggle.

The diversity in sampling criteria, sample sizes, demographic details, language focus, and geographic regions across the studies introduces potential biases. Sample sizes and levels of representation varied significantly among the studies, from small-scale studies (e.g., Study #46, which analyzed 4,124 Facebook posts from 43 undergraduate students with pre-specified criteria from the U.S.) to large-scale analyses (e.g., Study #5, which analyzed 56,411,200 tweets from 70,000 users across seven major U.S. cities). Many studies lacked detailed demographic information. The majority of studies focused predominantly on English-language posts, which are commonly associated with specific regions such as the U.S., U.K., Japan, Spain, and Portugal (although geographic information was explicitly reported in only about one-third of the studies) limiting the generalizability of the findings. Only a few studies examined posts in other languages, like Study #15, which analyzed Arabic tweets. Even within these regions and language-specific studies, demographic distribution was not always fully balanced. For example, Study #1 reported a mean participant age of 30.5 years (ranging from 18 to 68) and had a slight overrepresentation of female participants at 66.4%.

The non-representative sampling approaches observed across studies suggest limited generalizability to broader social media user populations. The primary biases identified include:

- **Platform Bias:** The predominance of Twitter (63.8%) over other platforms means that findings may not represent behaviors on platforms like Facebook, Instagram, or Reddit. As suggested by Olteanu et al. (2019), utilizing multi-platform data can reduce platform-specific biases and provide a more comprehensive view of user behaviors. However, while multi-platform data broaden the scope and reduce single-platform bias, platform-specific user demographics and engagement patterns may affect generalizability, with some platforms carrying more weight due to larger user bases or data volume.
- **Selection Bias:** Some studies relied on keyword-based sampling, which may overlook users not explicitly mentioning mental health. Study #7, for instance, searched for tweets containing 'I was diagnosed with depression.' As suggested by Morstatter, Pfeffer, Liu, and Carley (2013), combining keyword-based and random sampling can capture a broader range of user behaviors and discussions. Additionally, the limitations of Twitter's API exacerbate platform-specific challenges. As highlighted by Morstatter et al. (2013), Twitter's API does not provide access to all user-generated content, raising concerns about whether sampled data is representative of the platform's overall activity. This issue may lead to incomplete or skewed representations of user behavior, particularly in studies relying solely on API data. Researchers must critically evaluate the validity of conclusions drawn from API-retrieved data and consider combining multiple sampling strategies to mitigate such biases.

- **Language Bias:** The overwhelming focus on English-language content (over 90%) excludes insights from non-English-speaking communities, limiting the generalizability of findings across diverse linguistic groups. For instance, Study #15 was one of the few that analyzed non-English tweets, indicating the rarity of multilingual studies in this field. To address this, Danet and Herring (2007) recommended leveraging multilingual analysis methods, such as machine translation, or employing multilingual research teams to capture a more diverse linguistic landscape.
- **Geographic Bias:** While explicit geographic information was reported in only about one-third of the studies, the predominance of English-language posts suggests an implicit bias toward regions where English is the primary language, such as the U.S., U.K., and other English-speaking countries. Among the studies that reported geographic information, this predominance was evident. For example, Study #5 analyzed tweets from seven major U.S. cities, and Study #19 focused on Twitter users in Spain and Portugal. Hargittai (2015) suggested broadening the geographic scope to better represent global populations and avoid region-specific findings.
- **Self-selection Bias:** Platforms like Mechanical Turk (MTurk) or Clickworker, used in some studies (e.g., Studies #45 and #1, respectively), may attract specific demographic or employment profiles (e.g., higher digital literacy, particular age ranges, or specific socioeconomic statuses), affecting generalizability. While Chandler and Shapiro (2016) assessed the use of MTurk as a crowdsourcing tool, highlighting limitations in participant diversity and representativeness, which may skew results and underscore the need for multiple recruitment sources and stratified sampling for better generalizability.

In summary, no study in the review provided a fully representative sample of all social media users or posts. Key limitations include platform-specific focus (mostly Twitter), heavy reliance on non-probability sampling techniques (e.g., approximately 80% of the studies utilized convenience sampling or keyword filtering), and geographic and linguistic constraints. Notably, over 90% of the studies themselves acknowledged these limitations, recognizing the challenges of achieving representativeness in social media research. These limitations are, to a large extent, unavoidable due to the nature of social media platforms and the constraints of current data collection methodologies. This underscores the need for ongoing efforts to develop more sophisticated sampling techniques and analytical methods to mitigate these biases.

Similarly, some studies explicitly stated that their findings were intended to represent only specific populations. For instance, Study #8 and Study #21 focused on users discussing mental health or particular demographic groups on specific platforms. These limitations significantly impact the generalizability of findings to the broader population of social media users. Future research should strive for more diverse and representative sampling across platforms, languages, and geographic regions to enhance the applicability of results in the field of mental health and social media research.

3.5 Data Preprocessing with Focus on Negative Words Handling (Q3)

Across all studies, several common preprocessing tasks were consistently performed. Tokenization was conducted in all studies to break text into individual words or tokens, and text normalization steps included converting text to lowercase, as well as removing punctuation, URLs, and special characters. Many studies also performed stop-word removal to eliminate common words that are generally not informative for modeling. Additionally, some studies applied stemming and lemmatization to reduce words to their base or root forms, thereby unifying different morphological variants. Feature extraction techniques such as Bag of Words (BoW, Harris, 1954)[3], Term Frequency-Inverse Document Frequency (TF-IDF, Salton & Buckley, 1988)[4], and various word embedding methods were widely used to represent textual data numerically for modeling purposes.

While these standard preprocessing steps were broadly applied, certain aspects of sentiment analysis in mental health detection require additional attention. One such aspect is the effective handling of negative words, which is crucial for accurately interpreting sentiment and emotional tone, especially within this context. Among the 47 reviewed studies, approaches to negative words varied significantly:

First, only a minority of studies (11 out of 47 studies, approximately 23%) explicitly addressed negative words or negations in their preprocessing steps. Methods included standardizing all negative words to a basic form, like 'not," during preprocessing, which simplifies the representation of negations and improves sentiment recognition (e.g., Studies #3 and #34). Some studies quantified negative words as features by calculating metrics such as the user-specific average number of negative words per post. This metric captures the frequency of negative expressions per user and is then used as input for machine learning models to identify depressive emotions (e.g., Study #21). Others (e.g., Study #25) assigned a weight of -1 to negative adverbs to account for their inversion effect on sentence sentiment, ensuring more accurate sentiment quantification. Moreover, several studies employed specific methods for managing negations within their sentiment analysis frameworks. For example, some studies used sentiment analysis tools like TextBlob to determine the polarity of words in context, identifying negative words as indicators of depressive symptoms (e.g., Study #31). Others incorporated linguistic inquiry and word count (LIWC) categories related to

[3] BoW represents text as a vector by creating a vocabulary of all unique words in a corpus and counting the frequency of each word in a document. While simple and effective, BoW disregards word order and context, treating documents as collections of independent words.

[4] TF-IDF evaluates the importance of a word in a document relative to a collection of documents. It combines term frequency (how often a word appears in a document) with inverse document frequency (reducing the weight of common words that appear across many documents). This technique highlights terms that are more informative for classification or clustering tasks.

negations and negative emotions, indirectly addressing negations through predefined lexicon categories (Studies #1, #40, #42, #46, and #47).

The importance of negation handling has also been recognized in studies currently under review. For instance, Study #6 specifically explored the role of negation preprocessing in sentiment analysis for depression detection. By comparing datasets with and without negation handling, the authors demonstrated that addressing negations can significantly improve the accuracy of both sentiment analysis and depression detection, underscoring the need to address them in preprocessing. This study highlights the critical need for comprehensive negation handling in preprocessing to enhance the reliability of machine learning models in mental health contexts.

Second, a subset of studies (9 out of 47 studies, approximately 19%) did not explicitly handle negative words but employed advanced language models capable of inherently managing negations due to their contextual understanding, such as transformer-based models like Bidirectional Encoder Representations from Transformers (BERT, Devlin, Chang, Lee, & Toutanova, 2018) and Mental Health BERT (MentalBERT, Ji et al., 2022) (e.g., Studies #8, #9, #15, #16, and #39). These transformer-based models can capture the context of negations by processing text bi-directionally without explicit preprocessing steps. Other studies used attention mechanisms[5] (Vaswani et al., 2017) with word embeddings, such as attention layers combined with Global Vectors for Word Representation (GloVe) embeddings (Pennington, Socher, & Manning, 2014), allowing models to inherently understand and assign appropriate weights to negations through contextual embeddings (e.g., Studies #7, #10, and #13). Additionally, Embeddings from Language Models (ELMo, Peters et al., 2018), which capture the entire context of a word within a sentence, was also noted as a method that could capture the effect of negative words without explicit handling (Study #45).

However, the majority (27 out of 47 studies, approximately 57%) neither explicitly addressed negative words in their preprocessing nor used models inherently capable of handling negations (i.e., Studies #2, #4, #5, #11, #12, #14, #17, #18, #19, #20, #22, #23, #24, #26, #27, #28, #29, #30, #32, #33, #35, #36, #37, #38, #41, #43, and #44). These studies primarily focused on standard preprocessing tasks (e.g., tokenization, lowercasing, stop-word removal, stemming, and lemmatization), feature extraction methods (e.g., TF-IDF, BoW), and basic word embeddings (e.g., Word to Vector [Word2Vec]), without any special consideration for negations.

The impact on model performance and potential bias varied depending on how negative words were handled. Studies that explicitly addressed negative word handling reported improvements in model accuracy and a more nuanced understanding of sentiment (Helmy, Nassar, & Ramadan, 2024). Proper handling of negations allowed these models to correctly interpret phrases where

[5] Attention mechanisms allow models to focus on specific parts of the input data by assigning different weights to different elements. This enables the model to capture and utilize relevant contextual information more effectively during processing.

negations invert the sentiment (e.g., 'not happy" versus 'happy"), leading to more reliable results. In contrast, studies that did not explicitly account for negative words risked misinterpreting negated expressions, introducing bias into their findings. This oversight can cause models to incorrectly assign positive sentiment to negated negative expressions or vice versa, thus skewing the analysis. Such biases can significantly affect the overall performance and generalizability of the models, particularly in sensitive applications like depression detection. While some studies used advanced models capable of inherently handling negations (e.g., Studies #7, #8, #9, #10, #13, #15, #16, #39, and #45), reliance solely on the model's ability without explicit preprocessing might not capture all nuances of negations. Explicitly addressing negations can further enhance model performance, even when using sophisticated language models (Khandelwal & Sawant, 2020). Therefore, integrating both advanced modeling techniques and careful preprocessing of negative words may provide the most effective approach.

In summary, the review highlights a significant gap in the explicit handling of negative words in data preprocessing among studies focused on sentiment analysis and related fields. Proper management of negations is crucial, as it can substantially impact both model accuracy and reliability. Without adequately handling negative words, models may introduce bias and reduce their effectiveness, particularly in applications such as mental analysis and depression detection, where understanding sentiment nuances is critical. Future studies should prioritize the inclusion of explicit negation handling techniques within their preprocessing pipelines to enhance model performance and ensure more accurate interpretations of textual data.

3.6 Model Development

Hyperparameter Tuning (Q3, Q4 & Q5) Hyperparameters are external configurations set before the training process of machine learning models. Unlike model parameters, which are learned from the data during training, hyperparameters govern the learning process itself, such as the learning rate, regularization strength, and the number of hidden layers. Proper hyperparameter tuning ensures optimal model performance by balancing underfitting and overfitting, thus improving the model's ability to generalize to unseen data. Hyperparameter tuning is a critical aspect of optimizing machine learning models, directly impacting their performance and reliability. Our evaluation of the 47 reviewed studies focused on whether the studies reported their hyperparameters, the extent to which these hyperparameters were optimized, and whether tuning was applied consistently across all models within each study.

In particular, 27 studies (approximately 60%) reported using hyperparameters, but not all of them performed proper tuning. Only a limited number of studies ensured consistent tuning across all models, with many opting for default settings or tuning only specific models, leaving significant performance potential unexplored (Yang & Shami, 2020). This practice suggests that while hyperparameters are acknowledged by researchers, there is still a notable gap in their

comprehensive and consistent optimization across studies. The breakdown of hyperparameter reporting and tuning practices is presented in Table 3.

Table 3: Hyperparameter Reporting and Tuning Practices in Reviewed Studies.

Hyperparameter Reporting	**Number (%) of Studies**	**Studies #**
Reported & Tuned for All Models	13 (27.7%)	#11, #12, #13, #16, #18, #21, #22, #25, #26, #28, #33, #45, #47
Reported but Partially Tuned	4 (8.5%)	#1, #8, #15, #23
Reported but Not Tuned	11 (23.4%)	#3, #4, #7, #9, #10, #31, #36, #39, #40, #41, #43
Not Reported or Tuned	19 (40.4%)	#2, #5, #6, #14, #17, #19, #20, #24, #27, #29, #30, #32, #34, #35, #37, #38, #42, #44, #46

The absence of consistent hyperparameter tuning can result in suboptimal model performance, reduced generalizability, or biased model comparisons. Key hyperparameters such as learning rate, regularization terms, or the number of hidden layers directly impact a model's training process and final accuracy (Mantovani, Rossi, Vanschoren, Bischl, & de Carvalho, 2015; Probst, Boulesteix, & Bischl, 2019). Without proper tuning, models may overfit, meaning they perform well on training data but poorly on unseen data, or underfit, failing to capture the complexity of the data altogether. For example, Study #2 did not report any tuning, which likely affected its model's ability to generalize to unseen data, leading to reduced model performance.

When only some models are tuned, comparisons across models become biased, as those with optimized hyperparameters gain an undue advantage. In Study #1, for instance, the Elastic Net model had its hyperparameters tuned, while other models, such as random forest, were left with default settings. This discrepancy can misleadingly suggest the superiority of the Elastic Net model due to tuning alone, rather than any inherent advantage in its architecture, leading to biased model comparisons.

A significant proportion of studies did not report hyperparameter tuning (approximately 40%) or failed to consistently tune them across all models (approximately 32%), which compromises the validity of their findings. For example, Studies #2 and #4 used default settings and missed opportunities to enhance performance, while Study #1 tuned hyperparameters for only one model, resulting in biased comparisons. Proper hyperparameter tuning is essential to avoid issues like overfitting or underfitting. Consistent tuning across all models ensures fair comparisons and enhances result validity.

Providing detailed descriptions of hyperparameter settings and optimization processes enhances transparency and reproducibility. Standardized tuning protocols, such as grid search, random search, or Bayesian optimization, should be employed to explore optimal configurations. Clearly documenting tuning strategies and any challenges encountered will provide valuable context for interpreting model performance results and strengthen the credibility of future machine learning studies. Future research should prioritize consistent tuning strategies and detailed reporting to enhance the credibility and reproducibility of their machine learning studies.

Data Partitioning (Q6) Proper data partitioning is fundamental to developing robust machine learning models that generalize well to unseen data. Typically, datasets are divided into three subsets: the training set, used to train the model and learn patterns; the validation set, used to fine-tune hyperparameters and avoid overfitting; and the test set, reserved for evaluating the model's final performance on unseen data. Of the 47 reviewed studies, 32 studies (approximately 68%) adhered to recommended machine learning protocols by appropriately dividing their datasets into training, validation, and test sets or by employing cross-validation techniques. The breakdown of data partitioning practices is summarized in Table 4.

Among the studies that explicitly partitioned their datasets, such as Studies #1, #6, and #7, performance metrics were reported based on the test sets, adhering to the best practices outlined by Goodfellow, Bengio, and Courville (2016). By evaluating their models on unseen data, they ensured that the models' performance accurately reflected their generalizability.

Table 4: Summary of Data Partitioning Practices Across Reviewed Studies.

Data Partitioning Practices	**Number (%) of Studies**	**Studies #**
Training / Validation /Test Split	32 (68.1%)	#1, #6, #7, #8, #10, #11, #13, #15, #16, #17, #18, #19, #21, #22, #23, #25, #26, #28, #29, #30, #32, #33, #34, #35, #36, #40, #41, #42, #43, #45, #46, #47
Cross-validation without Traditional Split	7 (14.9%)	#3, #4, #14, #24, #38, #39, #44
Inadequate or Unreported Partitioning	8 (17.0%)	#2, #5, #9, #12, #20, #27, #31, #37

Seven studies used cross-validation methods instead of a traditional train/-validation/test split. Techniques like k-fold cross-validation provide a robust assessment of a model's ability to generalize by iteratively training and testing on different subsets of the dataset (Hastie, Friedman, & Tibshirani, 2009). For instance, Study #39 utilized 5-fold cross-validation, where the dataset was divided into five subsets, with each subset used as a test set once while the remaining subsets formed the training set. The reported metrics—Positive Predictive Value (PPV), Sensitivity, and F1 Score—were averaged across the five test folds in the cross-validation process, ensuring that evaluation was based on separate test data rather than solely on the training data.

Conversely, as shown in Table 4, approximately 17% of studies (8 out of 47) did not report sufficient details on data partitioning or did not employ partitioning techniques. For example, Study #2 provided limited information about its dataset division and did not elaborate on how model performance was evaluated, while Study #5 applied pre-existing models without conducting new data partitioning or validation within their analysis, thereby limiting the validity of their performance assessments.

Inadequate data partitioning practices introduce significant risks of bias, particularly overfitting. Models that lack proper data division tend to memorize the training data, leading to overly optimistic performance metrics that do not accurately reflect real-world applicability (Bishop, 2006).

According to A. Ng (2018), proper validation and testing sets are crucial for assessing generalization and preventing overfitting. Without these, models may appear overly effective due to inflated performance metrics, misleading when applied beyond the training context. For example, studies that evaluated models solely on training data, such as Studies #2 and #5, likely overestimate their real-world performance.

In summary, while the majority of the reviewed studies adhered to best practices in data partitioning—thereby enhancing the credibility and generalizability of their findings—a significant minority did not. The lack of proper data partitioning in approximately 17% of studies introduces risks of bias, underscoring the need for more rigorous practices. For the development of robust models, future research should consistently apply proper data partitioning and report performance based on validation or test sets to provide accurate, unbiased evaluations. Transparent data partitioning and evaluation reporting, as emphasized by Bishop (2006) and Goodfellow et al. (2016), is fundamental to enhancing reproducibility and reliability in machine learning research. By incorporating these practices, researchers can enhance the reliability of their models, ensure that findings are both valid and applicable in real-world scenarios, and contribute to the advancement of the field.

3.7 Model Evaluation: Evaluation Metrics for Imbalanced Class Scenarios (Q8, Q9 & Q10)

In the domain of depression-related emotion detection, datasets often exhibit significant class imbalance, with non-depressed cases vastly outnumbering de-

pressed ones. This imbalance poses challenges for model evaluation, as traditional metrics like accuracy can be misleading. According to He and Garcia (2009), accuracy may not adequately reflect a model's performance in imbalanced scenarios because a model could achieve high accuracy by simply predicting the majority class. Therefore, metrics such as recall, precision, F1 score, and Area Under the Receiver Operating Characteristic Curve (AUROC or AUC) are preferred, as they provide a more balanced evaluation by accounting for both false positives and false negatives. He and Ma (2013) and Japkowicz and Stephen (2002) further emphasize the necessity of using these metrics, arguing that they are crucial for a comprehensive assessment of model performance in the presence of class imbalance.

In the context of depression detection, recall, measures the proportion of actual positive cases (individuals with depression) that the model correctly identifies (i.e., Recall $= \frac{\text{True Positive}}{\text{True Positive+False Negative}}$), is particularly important. A high recall indicates that the model is successfully identifying most individuals who are truly depressed (true positives), although this often comes at the cost of more false positives, where individuals without depression are incorrectly flagged as depressed. Failing to identify someone who is depressed (a false negative) could have serious consequences, as it may result in a missed opportunity to provide help or intervention. Therefore, prioritizing recall ensures that the model captures as many true positive cases as possible, even if it risks increasing false positives. In this context, minimizing false negatives is often a higher priority, given the potential implications for those who might otherwise go undiagnosed and unsupported (Bradford, Meyer, Khan, Giardina, & Singh, 2024).

Precision, on the other hand, measures the proportion of positive predictions that are correct (i.e., Precision $= \frac{\text{True Positive}}{\text{True Positive+False Positive}}$), highlighting the model's ability to avoid false positives. In depression detection, a low precision score indicates a high rate of false positives, where individuals who are not depressed are incorrectly labeled as depressed. This could lead to unnecessary concern or even stigmatization for those wrongly flagged. While high precision is desirable to avoid false alarms, an overly strict focus on precision could inadvertently lower recall, leading to more false negatives. Therefore, balancing precision and recall is essential to ensure that the model is not only identifying true cases of depression but also minimizing the number of false alarms. This balance is particularly critical in applications where both false negatives (missing a depressed individual) and false positives (incorrectly flagging someone as depressed) carry significant consequences (Bradford et al., 2024).

The F1 score, representing the harmonic mean of precision and recall, provides a balanced measure of both recall and precision. It is particularly useful in imbalanced datasets, where a balance between recall and precision is essential.

Finally, AUROC measures the model's ability to distinguish between positive and negative classes across different threshold settings, providing a comprehensive view of the model's discriminatory power. A higher AUROC indicates a better capability of distinguishing between depressed and non-depressed individuals, making it a robust metric for evaluating models in this domain. Among the

47 studies reviewed, approximately 35 (Studies #1, #3, #6, #7, #8, #13, #14, #15, #16, #17, #19, #21, #22, #23, #25, #26, #27, #28, #29, #30, #31, #32, #33, #34, #35, #36, #37, #39, #40, #41, #42, #43, #44, #45, #46) utilized these preferred metrics. For example, Study #6, "Depression Detection for Twitter Users Using Sentiment Analysis in English and Arabic Tweets," employed precision, recall, F1 score, and AUC to evaluate their models, acknowledging the importance of these metrics for imbalanced data. Similarly, Study #42, "Classification of Helpful Comments on Online Suicide Watch Forums," emphasized recall as a key metric in evaluating their model's effectiveness in identifying individuals at risk.

Other than the utilization of preferred metrics, an alternative way to address imbalanced data involves implementing data balancing techniques, including resampling and reweighting. For instance, Study #6 employed dynamic sampling methods, such as oversampling the minority class and undersampling the majority class, to balance the dataset. This approach ensured that the model had sufficient exposure to both classes before model construction and evaluation. Similarly, Study #41, "A Deep Learning Model for Detecting Mental Illness from User Content on Social Media," used Synthetic Minority Oversampling Technique (SMOTE) to enhance the representation of the minority class, leading to improved classification performance, particularly for underrepresented classes.

Notably, some studies (Studies #3, #6, #13, #15, #34, #40, #41, #42, #43) applied both data balancing techniques and preferred evaluation metrics together to comprehensively address the class imbalance. For example, 'Explainable Depression Detection with Multi-Aspect Features Using a Hybrid Deep Learning Model on Social Media' (Study #13) first implemented preprocessing steps to balance the dataset, enhancing the model's ability to learn from both classes equally. After addressing the class imbalance, the study then used the F1 score and related metrics to evaluate model performance, ensuring a more accurate and fair assessment. These examples indicate that researchers are increasingly aware of the class imbalance issue and are employing various approaches to address it effectively.

Conversely, some studies primarily relied on accuracy without addressing class imbalance issues. For example, Studies #2, #10, and #24 reported high accuracy but did not mention techniques to mitigate the effects of class imbalance.

In the context of depression detection, addressing class imbalance is essential for achieving reliable model evaluation. When instances of the non-depressed class significantly outnumber those of the depressed class, the resulting imbalance can skew model outcomes if not properly managed. Two primary strategies are commonly employed to mitigate this issue: the use of evaluation metrics that accommodate class imbalance and data preprocessing techniques, such as resampling and reweighting. Japkowicz and Stephen (2002) emphasize that metrics like recall, precision, and F1 score offer a more nuanced evaluation by accounting for both positive and negative classes, thus reducing potential bias. Additionally, data preprocessing methods like reweighting or resampling adjust the dataset

to provide a balanced exposure to both classes, enhancing model training on imbalanced data.

While some studies utilized both strategies, demonstrating a thorough approach to handling imbalance, others employed just one—either through preferred evaluation metrics or data balancing. Even when only one strategy is adopted, it can still reduce potential bias to some extent. However, solely relying on accuracy introduces a significant risk of bias, as it often leads the model to favor the majority class, thereby failing to identify depressed individuals accurately. Chawla, Japkowicz, and Kotcz (2004) highlight that this reliance on accuracy alone can lead to misleading conclusions in imbalanced datasets, as it does not accurately reflect the model's ability to detect minority class instances.

Out of the 47 studies analyzed, approximately 35 employed preferred metrics such as F1 score, precision, recall, or AUROC, recognizing their importance in evaluating models on imbalanced datasets. Seven studies explicitly mentioned preprocessing steps like resampling to mitigate class imbalance, even when using accuracy as an evaluation metric. However, several studies relied mainly on accuracy without addressing class imbalance, potentially introducing bias into their evaluations.

In conclusion, while a significant number of studies have adopted appropriate evaluation metrics and techniques to address class imbalance, there remains a need for broader implementation of these practices. Incorporating balanced metrics and addressing class imbalance is essential for reliable and valid model evaluations in depression detection research. As Fernandez et al. (2018) recommended, employing these strategies enhances the robustness of machine learning models in domains characterized by imbalanced datasets.

3.8 Reporting: Transparency and Completeness

Transparency and completeness in reporting are fundamental to the integrity and reproducibility of scientific research. In our examination of the 47 studies, we assessed the extent to which they transparently reported their methodologies, findings, and limitations. Notably, all studies (100%) included a limitation section, indicating an overall acknowledgment of the importance of addressing potential shortcomings. However, the depth and specificity of these disclosures varied significantly across the studies.

While every study mentioned limitations, not all of them fully recognized or disclosed all critical methodological issues that could impact their findings. For instance, as highlighted in our earlier assessments, approximately 23% of the studies (11 out of 47) did not properly partition their data or failed to report their data partitioning methods adequately (Studies #2, #5, #9, #12, #20, #27, #31, and #37). Despite this, only a few of these studies explicitly acknowledged the potential biases introduced by improper data partitioning in their sections of limitations. This suggests that while researchers are generally aware of the necessity to report limitations, there is a gap in fully understanding or disclosing specific methodological shortcomings, such as data partitioning, which is crucial for model generalizability and validity.

Similarly, in the context of hyperparameter tuning, approximately 43% of the studies did not report or properly tune hyperparameters across all models used (e.g., Studies #1, #2, #4, #5, #12, #14, #17, #19, #20, #24, #27, #29, #30, #32, #34, #35, #37, #38, #42, #44, and #46). Only a few acknowledged this limitation in their reports. This lack of comprehensive reporting on hyperparameter tuning can lead to biased model comparisons and affect the reproducibility of the studies.

Incomplete or non-transparent reporting can introduce significant bias and limit the reproducibility and applicability of research findings. When critical methodological details are omitted or underreported, it hinders the ability of other researchers to replicate studies or to understand the context in which the results are valid. For instance, failing to disclose improper data partitioning can lead to overestimation of model performance due to overfitting (Bishop, 2006). Models evaluated on training data or without appropriate validation may appear to perform well, but this performance may not generalize to new, unseen data. This oversight can mislead stakeholders about the efficacy of the models and affect subsequent research or practical applications that build upon these findings.

Similarly, not reporting on hyperparameter tuning practices can result in unfair comparisons between models and misinterpretations of their relative performances (Claesen & Moor, 2015; Zhang et al., 2025). Models with optimized hyperparameters may outperform others not because they are inherently better but because they were given an optimization advantage. Without transparency in reporting these practices, readers cannot assess the fairness of the comparisons or replicate the optimization process.

In conclusion, while all 47 studies recognized the importance of reporting limitations, there remains a notable disparity in the thoroughness and transparency of their reporting. For the field to advance, transparent and comprehensive reporting of methodologies and limitations is essential. Future research should strive for complete disclosure of data collection, preprocessing, model development, hyperparameter tuning, and evaluation metrics. This includes acknowledging specific methodological limitations, such as data partitioning practices and sampling biases, and discussing how these limitations may impact results and generalizability. Such transparency will allow others to interpret findings accurately, replicate studies, and build upon prior work effectively.

3.9 Summary of Findings and Implications for Future Research

This systematic review evaluated biases throughout the entire lifecycle of machine learning and deep learning models for depression detection on social media. In sampling, biases arose from a predominant reliance on Twitter, English-language data, and specific geographic regions, limiting the representativeness of findings. Data preprocessing commonly showed inadequate handling of negations, which can skew sentiment analysis results. Model development was often compromised by inconsistent hyperparameter tuning and improper data partitioning, reducing model reliability and generalizability. Lastly, in model eval-

uation, an overreliance on accuracy without addressing class imbalance risked favoring majority class predictions, potentially misleading results. These findings highlight the importance of enhancing methodologies to bolster the validity and applicability of future research.

To address these biases, future research should improve practices across all stages of the machine learning lifecycle. Expanding data sources across multiple platforms, languages, and regions will help mitigate platform and language biases and improve representativeness. Standardizing data preprocessing, especially with explicit negation handling, and employing resampling and reweighting techniques will enhance sentiment analysis accuracy and balance datasets. Consistent hyperparameter tuning protocols are essential to ensure fair model comparisons and optimal performance. Lastly, prioritizing evaluation metrics like precision, recall, F1 score, and AUROC in imbalanced datasets, particularly for depression detection, will yield more accurate and insightful assessments. By implementing these improvements, future studies can achieve greater model robustness and generalizability, contributing to more effective mental health detection tools.

4 Discussion

The escalating prevalence of mental health conditions, particularly depression, poses a significant global health challenge. Social media platforms have emerged as rich data sources where individuals express their thoughts and emotions, offering a unique opportunity to detect mental health issues through advanced computational methods. Machine learning and deep learning models hold promise for analyzing this vast, unstructured data to identify patterns indicative of depression. This systematic review aimed to evaluate the effectiveness of these models in detecting depression on social media, focusing on identifying and analyzing biases throughout the ML lifecycle.

4.1 Summary of Key Findings

Our review uncovered several key biases and methodological challenges that impact the reliability and generalizability of machine learning and deep learning models in this domain. Sampling biases emerged due to a predominant reliance on specific social media platforms, particularly Twitter, which was used in 63.8% of the studies. Additionally, most studies focused on English-language content and users from specific geographic regions, primarily the United States and Europe. These biases limit the representativeness of findings, as they do not capture the diversity of global social media users. In data preprocessing, many studies inadequately handled linguistic nuances, such as negations and sarcasm. Only about 23% of the studies explicitly addressed the handling of negative words or negations, which are crucial for accurate sentiment analysis in depression detection.

Model development issues were also prominent. Inconsistent hyperparameter tuning practices were observed, with only 27.7% of the studies properly tuning hyperparameters for all models. Moreover, approximately 17% of the studies did not adequately partition their data into training, validation, and test sets. These practices can lead to overfitting, reducing the models' ability to generalize to new data. Regarding model evaluation, many studies relied heavily on accuracy as the primary evaluation metric without addressing class imbalances inherent in depression detection datasets. While about 74.5% of the studies used metrics suitable for imbalanced data, such as precision, recall, F1 score, and AUROC, others did not, potentially skewing the evaluation of model performance. Finally, despite all studies including a limitations section, transparency varied significantly, with critical methodological details like data partitioning methods and hyperparameter settings often underreported. This inconsistency hinders reproducibility and the ability to fully assess the validity of the findings.

4.2 Strengths and Limitations of the Review

This systematic review stands out for its comprehensive scope, examining biases across the entire ML lifecycle, from sampling to reporting, in depression detection on social media. By not limiting the analysis to specific aspects, the review offers a holistic view of how biases can influence model validity. Another strength is the structured methodological approach, adhering to established guidelines with a well-defined search strategy and clear inclusion criteria. Focusing on studies published after 2010, it reflects the latest advancements in ML and DL applications for mental health.

The use of established bias assessment tools, particularly PROBAST, adds rigor by systematically evaluating bias across key methodological domains. Additionally, the review's detailed data extraction process facilitated a structured analysis, allowing for the identification of patterns and providing actionable recommendations, such as diversifying data sources and improving transparency.

However, the review also has limitations. Limited database coverage and the English-only restriction may exclude valuable insights from non-English research, potentially affecting the generalizability of the findings. The focus on recent studies (post-2010) might have overlooked earlier influential works, while heterogeneity in study designs hindered direct comparisons and precluded a quantitative meta-analysis. Moreover, publication bias could skew findings toward positive results, and excluding grey literature means emerging methodologies may not be fully captured. Lastly, while ethical considerations were acknowledged, a deeper exploration of issues like data privacy and informed consent is warranted.

These limitations suggest areas for improvement in future research, such as broadening database and language coverage, including grey literature, and conducting a meta-analysis where feasible. By addressing these areas, future studies can enhance the robustness of ML models for mental health detection and provide a more comprehensive, ethical, and globally relevant understanding of the field.

4.3 Implications for Future Research

To enhance the generalizability and applicability of machine learning and deep learning models in depression detection on social media, addressing identified biases is essential. First, diversifying data sources across multiple social media platforms and incorporating non-English languages and underrepresented regions will improve representativeness and generalizability. Improving sampling methods is crucial. Combining keyword-based sampling with random sampling techniques can help reduce selection bias and capture users who may not explicitly mention depression but exhibit relevant behaviors. In the data preprocessing step, researchers should standardize practices to explicitly handle linguistic nuances like negations and sarcasm, which are vital for accurate sentiment analysis. Additionally, applying resampling or reweighting techniques can help balance datasets, ensuring that both classes—particularly the minority depressive class—are adequately represented. Advanced natural language processing techniques that account for linguistic nuances, such as sarcasm and context-dependent meanings, should be employed.

Consistent and comprehensive hyperparameter tuning across all models is essential to ensure fair comparisons and optimize model performance. Proper data partitioning practices, including the use of validation and test sets, should be implemented to prevent overfitting and assess model generalizability. When evaluating models, researchers should prioritize metrics that account for class imbalance, such as precision, recall, F1 score, and AUROC. These metrics provide a more balanced assessment of model performance and are more informative in the context of detecting depression, where the minority class is of primary interest.

4.4 Concluding Remarks

This systematic review highlights significant methodological limitations in current research on detecting depression through social media analysis using machine learning and deep learning models. Addressing these limitations is critical to developing more accurate, reliable, and generalizable models that can effectively identify individuals at risk of depression. Future research should focus on diversifying data sources, improving sampling methods, enhancing data preprocessing and model development practices, and employing appropriate evaluation metrics to ensure balanced and meaningful assessments.

By advancing these methodological approaches, researchers can contribute to the advancement of mental health detection tools that are ethically sound and effective across diverse populations and platforms. Such advancements hold the potential to facilitate early intervention strategies, ultimately improving mental health outcomes on a global scale.

Acknowledgments

The authors thank Dr. Jin (Veronica) Liu (ORCID: 0000-0001-5922-6643) for her valuable comments and insightful feedback on the development of this manuscript.

Her input has contributed to improving the clarity and overall presentation of the work.

Availability of Data and Materials

The reviewed titles, authors, and publication years of the included studies have been provided in Table A.1. Detailed information on each reviewed paper is hosted on GitHub: `https://github.com/odile1999/Systematic-Review-Machine-Learning-on-Depression`.

Authors' Contributions

Project administration: Y.T. and Y.C.; Conceptualization: Y.T. and Y.C.; Methodology: Y.T., J.D., and Y.C.; Investigation: Y.T., Y.C., J.D., Z.W., Y.Z., X.S., and Y.L.; Formal Analysis: Y.T., Y.C., J.D., Z.W., Y.Z., X.S., and Y.L.; Writing - Original Draft: Y.T., Y.C., and J.D.; Writing - Review and Editing: Y.C., Z.W., Y.Z., X.S., and Y.L.

References

Agarwal, A. K., Mittal, J., Tran, A., Merchant, R., & Guntuku, S. C. (2023). Investigating social media to evaluate emergency medicine physicians' emotional well-being during covid-19. *JAMA Netw Open*, *6*(5), e321708. doi: https://doi.org/10.1001/jamanetworkopen.2023.12708

Angskun, J., Tipprasert, S., & Angskun, T. (2022). Big data analytics on social networks for real-time depression detection. *J Big Data*, *9*, 69. doi: https://doi.org/10.1186/s40537-022-00622-2

Baghdadi, N. A., Malki, A., Magdy Balaha, H., Abdul-Azeem, Y., Badawy, M., & Elhossieni, M. (2022). An optimized deep learning approach for suicide detection through arabic tweets. *PeerJ Comput Sci*, *8*, e1070. doi: https://doi.org/10.7717/peerj-cs.1070

Baird, A., Xia, Y., & Cheng, Y. (2022). Consumer perceptions of telehealth for mental health or substance abuse: A twitter-based topic modeling analysis. *JAMIA Open*, *5*(2), ooac028. doi: https://doi.org/10.1093/jamiaopen/ooac028

Beier, F., Pryss, R., & Aizawa, A. (2023). Sentiments about mental health on twitter—before and during the covid-19 pandemic. *Healthcare (Basel)*, *11*(21), 2893. doi: https://doi.org/10.3390/healthcare11212893

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. doi: https://doi.org/http://www.loc.gov/catdir/enhancements/fy0818/2006922522-d.html

Borba de Souza, V., Campos Nobre, J., & Becker, K. (2022). Dac stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts. *IEEE J Biomed Health Inform*, *26*(7), 3303-3311. doi: https://doi.org/10.1109/JBHI.2022.3151589

Bradford, A., Meyer, A. N. D., Khan, S., Giardina, T. D., & Singh, H. (2024). Diagnostic error in mental health: a review. *BMJ Quality & Safety*, *33*(10), 663–672. doi: https://doi.org/https://qualitysafety.bmj.com/content/33/10/663

Calvo, R. A., Milne, D. N., Hussain, M. S., & Christensen, H. (2017). Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, *23*(5), 649-685. doi: https://doi.org/10.1017/S1351324916000383

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annu Rev Clin Psychol*, *12*, 53-81. doi: https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chandra, R., & Krishna, A. (2021). Covid-19 sentiment analysis via deep learning during the rise of novel cases. *PLOS ONE*, *16*(8), e0255615. doi: https://doi.org/10.1371/journal.pone.0255615

Chawla, N., Japkowicz, N., & Kotcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations*, *6*(1), 1-6. doi: https://doi.org/10.1145/1007730.1007733

Chen, L., Gong, T., Kosinski, M., Stillwell, D., & Davidson, R. L. (2017). Building a profile of subjective well-being for social media users. *PLOS ONE*, *12*(11), e0187278. doi: https://doi.org/10.1371/journal.pone.0187278

Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Comput Biol Med*, *135*, 104499. doi: https://doi.org/10.1016/j.compbiomed.2021.104499

Claesen, M., & Moor, B. D. (2015). *Hyperparameter search in machine learning.* Retrieved from `https://arxiv.org/abs/1502.02127`

Danet, B., & Herring, S. C. (2007). *The multilingual internet: Language, culture, and communication online.* Oxford University Press.

Das Swain, V., Ye, J., Ramesh, S. K., Mondal, A., Abowd, G. D., & De Choudhury, M. (2024). Leveraging social media to predict covid-19-induced disruptions to mental well-being among university students: Modeling study. *JMIR Form Res*, *8*, e52316. doi: https://doi.org/10.2196/52316

Davis, B. D., McKnight, D. E., Teodorescu, D., Quan-Haase, A., Chunara, R., Fyshe, A., & Lizotte, D. J. (2020). Quantifying depression-related language on social media during the covid-19 pandemic. *Int J Popul Data Sci*, *5*(4), 1716. doi: https://doi.org/10.23889/ijpds.v5i4.1716

de Anta, L., Alvarez-Mon, M. A., Ortega, M. A., Salazar, C., Donat-Vargas, C., Santoma-Vilaclara, J., ... Alvarez-Mon, M. (2022). Areas of interest and social consideration of antidepressants on english tweets: A natural language processing classification study. *Journal of Personalized Medicine*, *12*(2), 20155. doi: https://doi.org/10.3390/jpm12020155

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement of depression in populations. In *Proceedings of the acm annual web science conference* (p. 47-56). New York, NY, USA. doi: https://doi.org/10.1145/2464464.2464480

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, *abs/1810.04805*. doi: https://doi.org/https://arxiv.org/abs/1810.04805

Doan, S., Ritchart, A., Perry, N., Chaparro, J. D., & Conway, M. (2017). How do you #relax when you're #stressed? a content analysis and infodemiology study of stress-related tweets. *JMIR Public Health Surveill*, *3*(2), e35. doi: https://doi.org/10.2196/publichealth.5939

Fernandez, A., Garcia, S., Galar, M., Prati, R., Krawczyk, B., & Herrera, F. (2018). *Learning from imbalanced data sets*. Springer. doi: https://doi.org/10.1007/978-3-319-98074-4

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Guntuku, S. C., Schneider, R., Pelullo, A., Young, J., Wong, V., Ungar, L., ... Merchant, R. (2019). Studying expressions of loneliness in individuals using twitter: an observational study. *BMJ Open*, *9*(1), e030355. doi: https://doi.org/10.1136/bmjopen-2019-030355

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49. doi: https://doi.org/10.1016/j.cobeha.2017.07.005

Hargittai, E. (2015). Is bigger always better? potential biases of big data derived from social network sites. *The Annals of the American Academy of Political and Social Science*, *659*, 63-76. doi: https://doi.org/http://www.jstor.org/stable/24541849

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. doi: https://doi.org/10.1080/00437956.1954.11659520

Hastie, T., Friedman, J. H., & Tibshirani, R. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: https://doi.org/10.1109/TKDE.2008.239

He, H., & Ma, Y. (2013). Assessment metrics for imbalanced learning. In *Imbalanced learning: Foundations, algorithms, and applications* (p. Chapter Reference (add specific pages if available)). Wiley-IEEE Press. (Book Chapter) doi: https://doi.org/10.1002/9781118646106.ch8

Helmy, A., Nassar, R., & Ramadan, N. (2024). Depression detection for twitter users using sentiment analysis in english and arabic tweets. *Artificial Intelligence in Medicine*, *147*, 102716. doi: https://doi.org/10.1016/j.artmed.2023.102716

Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R. M., Wang, H., & Ulhag, A. (2018). Depression detection from social network data using machine learning techniques. *Health Inf Sci Syst*, *6*(1), 8. doi: https://doi.org/10.1007/s13755-018-0046-0

Jamali, A. A., Berger, C., & Spiteri, R. J. (2023). Momentary depressive feeling detection using x (formerly twitter) data: Contextual language approach.

JMIR AI, *2*, e49531. doi: https://doi.org/10.2196/49531

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intell Data Anal*, *6*, 429-449. doi: https://doi.org/10.3233/IDA-2002-6504

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022, June). MentalBERT: Publicly available pretrained language models for mental healthcare. In N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 7184–7190). Marseille, France: European Language Resources Association. doi: https://doi.org/https://aclanthology.org/2022.lrec-1.778/

Kaur, R., Ahassan, S. U., Alankar, B., & Chang, V. (2021). A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets. *Inf Syst Front*, *23*(6), 1417-1429. doi: https://doi.org/10.1007/s10796-021-10135-7

Kavuluru, R., Williams, A. G., Ramos-Morales, M., Haye, L., Holaday, T., & Cerel, J. (2016). Classification of helpful comments on online suicide watch forums. In *Proceedings of the 7th acm conference on bioinformatics, computational biology, and health informatics (acm-bcb)* (pp. 32–40). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/2975167.2975170

Kelley, S. W., Monaghan, C. N., Burke, L., Whelan, R., & Gillan, C. M. (2022). Machine learning for anxiety language on twitter reveals weak and non-specific predictions. *NPJ Digit Med*, *5*, 35. doi: https://doi.org/10.1038/s41746-022-00576-y

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., ... Koenen, K. C. (2017). Trauma and ptsd in the who world mental health surveys. *European Journal of Psychotraumatology*, *8*(sup5), 1353383. doi: https://doi.org/10.1080/20008198.2017.1353383

Khandelwal, A., & Sawant, S. (2020, May). NegBERT: A transfer learning approach for negation detection and scope resolution. In N. Calzolari et al. (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 5739–5748). Marseille, France: European Language Resources Association. doi: https://doi.org/https://aclanthology.org/2020.lrec-1.704/

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Sci Rep*, *10*, 18446. doi: https://doi.org/10.1038/s41598-020-68764-y

Kumar, A., Khan, S. U., & Kalra, A. (2020). Covid-19 pandemic: a sentiment analysis. *European Heart Journal*, *41*(39), 3782–3783. doi: https://doi.org/10.1093/eurheartj/ehaa597

Levanti, D., Monastero, R. N., Zamani, M., Eichstaedt, J. C., Giorgi, S., Schwartz, H. A., & Meilxer, J. R. (2023). Depression and anxiety on twitter during the covid-19 stay-at-home period in 7 major u.s. cities. *AJPM Focus*, *2*, 100062. doi: https://doi.org/10.1016/j.focus.2022.100062

Li, Y., Cai, M., Qin, S., & Lu, X. (2020). Depressive emotion detection and

behavior analysis of men who have sex with men via social media. *Front Psychiatry*, *11*, 830. doi: https://doi.org/10.3389/fpsyt.2020.00830

Low, D. M., Munoz, F. L., Talkar, T., Torres, J., Cecchi, G., & Ghosh, S. S. (2020). Natural language processing reveals vulnerable mental health support groups and heightened anxiety on reddit during covid-19. *JMIR Ment Health*, *8*(6), e22635. doi: https://doi.org/10.2196/22635

Mantovani, R. G., Rossi, A. L. D., Vanschoren, J., Bischl, B., & de Carvalho, A. C. P. L. F. (2015). Effectiveness of random search in svm hyper-parameter tuning. In *Proceedings of the 2015 international joint conference on neural networks (ijcnn)* (pp. 1–8). IEEE. doi: https://doi.org/https://ieeexplore.ieee.org/document/7280664

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming api with twitter firehose. In *Proceedings of the 7th international conference on weblogs and social media, icwsm 2013* (pp. 400–408).

Ng, A. (2018). *Machine learning yearning.* doi: https://doi.org/https://info.deeplearning.ai/machine-learning-yearning-book

Ng, Q. X., Lim, Y. L., Ong, C., New, S., Fam, J., & Liew, T. M. (2024). Hype or hope? ketamine for the treatment of depression: results from the application of deep learning to twitter posts from 2010 to 2023. *Front Psychiatry*, *15*, 1369727. doi: https://doi.org/10.3389/fpsyt.2024.1369727

Obagbuwa, I. C., Danster, S., & Chibaya, O. C. (2023). Supervised machine learning models for depression sentiment analysis. *Front Artif Intell*, *6*, 1230649. doi: https://doi.org/10.3389/frai.2023.1230649

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Front Big Data*, *2*, 13. doi: https://doi.org/10.3389/fdata.2019.00013

Ophir, Y., Tikochinski, R., Asterhan, C. S. C., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. *Sci Rep*, *10*(1), 16685. doi: https://doi.org/10.1038/s41598-020-73917-0

Owen, D., Antypas, D., Hassoulas, A., Pardinas, A. F., Espinosa-Anke, L., & De Choudhury, M. (2023). Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation. *JMIR AI*, *2*, e41205. doi: https://doi.org/10.2196/41205

Patel, D., Sumner, S. A., Bowen, D., Zwald, M., Yard, E., Wang, J., ... Chen, Y. (2023). Predicting state-level suicide fatalities in the united states with real-time data and machine learning. *NPJ Ment Health Res*, *3*(1), 3. doi: https://doi.org/10.1038/s44184-023-00045-8

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543). Association for Computational Linguistics. doi: https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 2227–2237). Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/N18-1202

Prieto, V. M., Matos, S., Alvarez, M., Cacheda, F., & Oliveira, J. L. (2014). Twitter: A good place to detect health conditions. *PLoS One*, *9*(1), e86191. doi: https://doi.org/10.1371/journal.pone.0086191

Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, *20*(53), 1-32. doi: https://doi.org/http://jmlr.org/papers/v20/18-444.html

Roy, A., Nikolitch, K., McGinn, R., Jinah, S., Klement, W., & Kaminsky, Z. A. (2020). A machine learning approach predicts future risk of suicidal ideation from social media data. *NPJ Digit Med*, *3*, 78. doi: https://doi.org/10.1038/s41746-020-0287-6

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(6), 1121-1133. doi: https://doi.org/10.1080/02699930441000030

Saha, K., Yousuf, A., Boyd, R. L., Pennebaker, J. W., & De Choudhury, M. (2022). Social media discussions predict mental health consultations on college campuses. *Sci Rep*, *12*(1), 123. doi: https://doi.org/10.1038/s41598-021-03423-4

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. doi: https://doi.org/10.1016/0306-4573(88)90021-0

Shatte, A. B. R., Hutchinson, D.-M., Fuller-Tyszkiewicz, M., & Teague, S. J. (2020). Social media markers to identify fathers at risk of postpartum depression: A machine learning approach. *Cyberpsychol Behav Soc Netw*, *23*(9), 611-618. doi: https://doi.org/10.1089/cyber.2019.0746

Shatte, A. B. R., Hutchinson, D.-M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(8), 1426-1448. doi: https://doi.org/10.1017/S0033291719000151

Singh, A., & Singh, J. (2022). Synthesis of affective expressions and artificial intelligence to discover mental distress in online community. *Int J Ment Health Addict*, 1-26. doi: https://doi.org/10.1007/s11469-022-00966-z

Sun, B., Zhang, Y., He, J., Xiao, Y., & Xiao, R. (2019). An automatic diagnostic network using skew-robust adversarial discriminative domain adaptation to evaluate the severity of depression. *Comput Methods Programs Biomed*, *173*, 185-195. doi: https://doi.org/10.1016/j.cmpb.2019.01.006

Swapnarekha, H., Nayak, J., Behera, H. S., Dash, P. B., & Pelusi, D. (2023). An optimistic firefly algorithm-based deep learning approach for sentiment analysis of covid-19 tweets. *Math Biosci Eng*, *20*(2), 2582-2607. doi:

https://doi.org/10.3934/mbe.2023112

Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *27*(5), 1–53. doi: https://doi.org/10.1145/3398069

Thorstad, R., & Wolff, P. (2019). Predicting future mental illness from social media: A big data approach. *Behav Res Methods*, *51*(4), 1586-1600. doi: https://doi.org/10.3758/s13428-019-01235-z

Trifan, A., Semeraro, D., Drake, J., Bukowski, R., & Oliveira, J. L. (2020). Social media mining for postpartum depression prediction. *Stud Health Technol Inform*, *270*, 1391-1392. doi: https://doi.org/10.3233/SHTI200457

Ueda, M., Watanabe, K., & Sueki, H. (2023). Correction: Emotional distress during covid-19 by mental health conditions and economic vulnerability: Retrospective analysis of survey-linked twitter data with a semi-supervised machine learning algorithm. *J Med Internet Res*, *25*, e759. doi: https://doi.org/10.2196/47549

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems (neurips)* (pp. 5998–6008). Curran Associates Inc. doi: https://doi.org/https://arxiv.org/abs/1706.03762

WHO. (2023). *Mental disorders.* (Retrieved February 2, 2025) doi: https://doi.org/https://www.who.int/news-room/fact-sheets/detail/mental-disorders

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., ... Group, P. (2019). Probast: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, *170*(1), 51–58. doi: https://doi.org/10.7326/M18-1376

Wongkoblap, A. (2023). Automatic profiles collection from twitter users with depressive symptoms. *Stud Health Technol Inform*, *305*, 419-422. doi: https://doi.org/10.3233/SHTI230520

Wongkoblap, A., Vadillo, M. A., & Curcin, V. (2021). Deep learning with anaphora resolution for the detection of tweeters with depression: Algorithm development and validation study. *JMIR Ment Health*, *8*(6), e19624. doi: https://doi.org/10.2196/19824

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295-316. doi: https://doi.org/10.1016/j.neucom.2020.07.061

Yao, H., Rashidian, S., Dong, X., Duanmu, H., Rosenthal, R. N., & Wang, F. (2020). Detection of suicidality among opioid users on reddit: Machine learning-based approach. *J Med Internet Res*, *22*(21), e15293. doi: https://doi.org/10.2196/15293

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., ... Sheth, A. (2017). Semi-supervised

approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the ieee/acm international conference on advances in social networks analysis and mining* (p. 1191-1198). doi: https://doi.org/10.1145/3110025.3123028

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., ... Hitzler, P. (2020). Multimodal mental health analysis in social media. *PLoS One*, *15*(8). doi: https://doi.org/10.1371/journal.pone.0226248

Yin, Z., Fabbri, D., Rosenbloom, S. T., & Malin, B. (2015). A scalable framework to detect personal health mentions on twitter. *J Med Internet Res*, *17*(6), e138. doi: https://doi.org/10.2196/jmir.4305

Zhang, Y., Wang, Z., Ding, Z., Tian, Y., Dai, J., Shen, X., ... Cao, Y. (2025). *Tutorial on using machine learning and deep learning models for mental illness detection.* Retrieved from https://arxiv.org/abs/2502.04342

Zhou, T. H., Hu, G. L., & Wang, L. (2019). Psychological disorder identifying method based on emotion perception over social networks. *Int J Environ Res Public Health*, *16*(6). doi: https://doi.org/10.3390/ijerph16060953

Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, *25*(1), 281-304. doi: https://doi.org/10.1007/s11280-021-00992-2

Appendix: Reviewed Studies on Machine Learning Models for Depression Detection on Social Media

Table A1: Reviewed Studies on Machine Learning Models for Depression Detection on Social Media

Index	Title of the Paper	Reference
#1	Machine learning of language use on Twitter reveals weak and non-specific predictions	Kelley, Monaghan, Burke, Whelan, and Gillan (2022)
#2	Supervised machine learning models for depression sentiment analysis	Obagbuwa, Danster, and Chibaya (2023)
#3	A textual-based featuring approach for depression detection using machine learning classifiers and social media texts	Chiong, Budhi, Dhakal, and Chiong (2021)
#4	Emotional Distress During COVID-19 by Mental Health Conditions and Economic Vulnerability: Retrospective Analysis of Survey-Linked Twitter Data With a Semisupervised Machine Learning Algorithm	Ueda, Watanabe, and Sueki (2023)

Continued on next page

Index	Title of the Paper	Reference
#5	Depression and Anxiety on Twitter During the COVID-19 Stay-At-Home Period in 7 Major U.S. Cities	Levanti et al. (2023)
#6	Depression detection for Twitter users using sentiment analysis in English and Arabic tweets	Helmy et al. (2024)
#7	Deep Learning With Anaphora Resolution for the Detection of Tweeters With Depression: Algorithm Development and Validation Study	Wongkoblap, Vadillo, and Curcin (2021)
#8	Sentiments about Mental Health on Twitter-Before and during the COVID-19 Pandemic	Beier, Pryss, and Aizawa (2023)
#9	Hype or hope? Ketamine for the treatment of depression: results from the application of deep learning to Twitter posts from 2010 to 2023	Q. X. Ng et al. (2024)
#10	Quantifying depression-related language on social media during the COVID-19 pandemic	Davis et al. (2020)
#11	Predicting state-level suicide fatalities in the United States with realtime data and machine learning	Patel et al. (2023)
#12	Investigating Social Media to Evaluate Emergency Medicine Physicians' Emotional Well-being During COVID-19	Agarwal, Mittal, Tran, Merchant, and Guntuku (2023)
#13	Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media	Zogan, Razzak, Wang, Jameel, and Xu (2022)
#14	Big data analytics on social networks for real-time depression detection	Angskun, Tipprasert, and Angskun (2022)
#15	An optimized deep learning approach for suicide detection through Arabic tweets	Baghdadi et al. (2022)
#16	COVID-19 sentiment analysis via deep learning during the rise of novel cases	Chandra and Krishna (2021)
#17	A Scalable Framework to Detect Personal Health Mentions on Twitter	Yin, Fabbri, Rosenbloom, and Malin (2015)
#18	An automatic diagnostic network using skew-robust adversarial discriminative domain adaptation to evaluate the severity of depression	Sun, Zhang, He, Xiao, and Xiao (2019)
#19	Twitter: a good place to detect health conditions	Prieto, Matos, Alvarez, Cacheda, and Oliveira (2014)
#20	Consumer perceptions of telehealth for mental health or substance abuse: A Twitter-based topic modeling analysis	Baird, Xia, and Cheng (2022)

Continued on next page

Index	Title of the Paper	Reference
#21	Depressive Emotion Detection and Behavior Analysis of Men Who Have Sex With Men via Social Media	Li, Cai, Qin, and Lu (2020)
#22	Areas of Interest and Social Consideration of Antidepressants on English Tweets: A Natural Language Processing Classification Study	de Anta et al. (2022)
#23	An Optimistic Firefly Algorithm-Based Deep Learning Approach for Sentiment Analysis of COVID-19 Tweets	Swapnarekha, Nayak, Behera, Dash, and Pelusi (2023)
#24	How Do You #relax When You're #stressed? A Content Analysis and Infodemiology Study of Stress-Related Tweets	Doan, Ritchart, Perry, Chaparro, and Conway (2017)
#25	Psychological Disorder Identifying Method Based on Emotion Perception over Social Networks	Zhou, Hu, and Wang (2019)
#26	Momentary Depressive Feeling Detection Using X (Formerly Twitter) Data: Contextual Language Approach	Jamali et al. (2023)
#27	A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets	Kaur, Ahassan, Alankar, and Chang (2021)
#28	A machine learning approach predicts future risk to suicidal ideation from social media data	Roy et al. (2020)
#29	Studying expressions of loneliness in individuals using Twitter: an observational study	Guntuku et al. (2019)
#30	Automatic Profiles Collection from Twitter Users with Depressive Symptoms	Wongkoblap (2023)
#31	Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media	Yazdavar et al. (2017)
#32	Synthesis of Affective Expressions and Artificial Intelligence to Discover Mental Distress in Online Community	Singh and Singh (2022)
#33	DAC Stacking: A Deep Learning Ensemble to Classify Anxiety, Depression, and Their Comorbidity from Reddit Texts	Borba de Souza, Campos Nobre, and Becker (2022)
#34	A textual-based featuring approach for depression detection using machine learning classifiers and social media texts	Chiong et al. (2021)
#35	Detection of Suicidality Among Opioid Users on Reddit: Machine Learning-Based Approach	Yao et al. (2020)
#36	Social Media Markers to Identify Fathers at Risk of Postpartum Depression: A Machine Learning Approach	Shatte, Hutchinson, Fuller-Tyszkiewicz, and Teague (2020)

Continued on next page

Index	Title of the Paper	Reference
#37	Social Media Mining for Postpartum Depression Prediction	Trifan, Semeraro, Drake, Bukowski, and Oliveira (2020)
#38	Social Media Discussions Predict Mental Health Consultations on College Campuses	Saha, Yousuf, Boyd, Pennebaker, and De Choudhury (2022)
#39	Enabling Early Health Care Intervention by Detecting Depression in Users of Web-Based Forums using Language Models: Longitudinal Analysis and Evaluation	Owen et al. (2023)
#40	Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study	Low et al. (2020)
#41	A deep learning model for detecting mental illness from user content on social media	Kim, Lee, Park, and Han (2020)
#42	Classification of Helpful Comments on Online Suicide Watch Forums	Kavuluru et al. (2016)
#43	Predicting future mental illness from social media: A big-data approach	Thorstad and Wolff (2019)
#44	Depression detection from social network data using machine learning techniques	Islam et al. (2018)
#45	Deep neural networks detect suicide risk from textual Facebook posts	Ophir, Tikochinski, Asterhan, Sisso, and Reichart (2020)
#46	Leveraging Social Media to Predict COVID-19-Induced Disruptions to Mental Well-Being Among University Students: Modeling Study	Das Swain et al. (2024)
#47	Building a profile of subjective well-being for social media users	Chen, Gong, Kosinski, Stillwell, and Davidson (2017)

Journal of Behavioral Data Science, 2025, *5* (1), 103–147.
DOI: https://doi.org/10.35566/jbds/bain

A Tutorial on Supervised Machine Learning Variable Selection Methods in Classification for the Social and Health Sciences in R

Catherine M. Bain[1 [0000−0002−2767−6882]], Dingjing Shi[1,2 [0000−0002−5652−3818]], Yaser M. Banad[3 [0000−0001−7339−810X]], Lauren E. Ethridge[1,4 [0000−0003−0601−6911]], Jordan E. Norris[1 [0000−0002−4438−3416]], and Jordan E. Loeffelman[1 [0000−0002−0269−7708]]

[1] Department of Psychology, University of Oklahoma, Norman, OK, USA
cbain1@ou.edu
[2] School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA
[3] School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK, USA
[4] Department of Pediatrics, Section on Developmental and Behavioral Pediatrics, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

Abstract. With the increasing availability of large datasets in the behavioral and health sciences, the need for efficient and effective variable selection techniques has grown. While traditional methods like stepwise regression remain prevalent, numerous advanced techniques are available but underutilized in these fields. This tutorial aims to increase awareness and understanding of five variable selection methods available in the popular statistical software R: LASSO, Elastic Net, a penalized SVM classifier, random forest, and the genetic algorithm. Using a recent survey-based assessment dataset on misophonia diagnosis, we provide step-by-step guidance on variables selections and implementation of each method in the context of classification. We discuss the strengths, weaknesses, and performance of each technique, emphasizing the importance of selecting appropriate performance metrics. The associated code and data implemented in this tutorial are available on Open Science Framework and provide an interactive learning experience. We encourage social and health science researchers to adopt these advanced variable selection methods, leading to more robust, interpretable, and impactful models. This paper is written with the assumption that individuals have at least a basic understanding of R.

Keywords: Machine learning · Variable selection · Big data · R · Data classification

1 Introduction

In the behavioral and health sciences, selecting the right variables for a model is crucial for understanding human behavior's complexity. Researchers strive to uncover how personality traits influence treatment engagement, how symptoms manifest in disorders, and how to accurately classify individuals into meaningful groups for diagnosis or intervention. They not only want to understand how these aspects (i.e., variables) are related to each other and to overarching constructs but may also want to use the variables to classify individuals into groups (e.g., diagnosing clinical disorders, determining participant compliance, etc.). The accuracy of these classifications or predictions is greatly influenced by which variables a researcher uses to create the classifications. For example, if a researcher is interested in diagnosing someone with depression, the accuracy of the diagnosis would suffer if relying solely on the presence of a depressed mood. However, if they use a variety of variables like depressed mood, loss of interest in activities, hours slept, and change in appetite or weight, their classification would be more accurate.

Researchers must carefully construct their classification models to understand variable interrelationships while maximizing predictive accuracy. Variable selection techniques can help researchers to identify and select informative variables to build these models. The use of variable selection techniques can lead to more accurate predictions, reduce the computational cost of creating the model, and improve the parsimony of the model by eliminating redundant and irrelevant variables. For example, variable selection techniques have been used to build models pertaining to identifying exposure-outcome associations (Lenters, Vermeulen, & Portengen, 2018) as well as predicting mortality rates (Amene, Hanson, Zahn, Wild, & Döpfer, 2016; Bourdès et al., 2010), psychological strain in teachers (Wettstein et al., 2023), and nomophobia (Luo, Ren, Li, & Liu, 2021).

Behavioral researchers often turn to stepwise regression to perform variable selection. An APA PsychINFO database search for the term "stepwise regression" returned 222 peer-reviewed articles published in the last 3 years using stepwise regression for variable selection. Stepwise regression, however, has many severe limitations and statistical experts do not recommend it (Smith, 2018; Thompson, 1995; Whittingham, Stephens, Bradbury, & Freckleton, 2006). These limitations include the inability to distinguish signal (i.e., true predictor variables) from noise (Derksen & Keselman, 1992; Kok, Choi, Oh, & Choi, 2021; Whittingham et al., 2006; Wiegand, 2010), underestimation of p-values, and failure to replicate (Smith, 2018; Thompson, 1995). As such, many alternative variable selection algorithms have been proposed in the literature, but behavioral researchers have been slow to adopt these new methods in place of more traditional methods (Serang, Jacobucci, Brimhall, & Grimm, 2017; Shi, Shi, & Fairchild, 2023). One potential reason for this delay may be the disconnect between methodological and applied behavioral researchers, as much methodological research is often inaccessible for applied researchers at first (e.g., complex techniques, lack of published code, or no tutorials). An APA PsychINFO database search for the term "variable selection" returned 253 papers published

in quantitative methods journals in the last 20 years, indicating that methodological researchers are dedicated to developing better approaches to variable selection than stepwise regression. Of these publications, however, only one is a tutorial (Gunn, Hayati Rezvan, Fernández, & Comulada, 2023).

Given the clear gap in the popularity of variable selection methodological research and the lack of tutorials on how to apply them, the field would benefit greatly from additional tutorials on variable selection techniques with demonstrations of how to apply them to psychological datasets. The following groups would benefit, specifically, from this tutorial. First, behavioral and health science researchers who are working with big data or looking to further enhance their understanding of advanced variable selection techniques to build more robust and interpretable models. Second, graduate students and early career researchers who are new to machine learning and variable selection methods and seek practical guidance on applying these techniques in their own research. Third, those who may be teaching courses on data analysis, machine learning, or statistics who are looking for comprehensive examples to illustrate advanced techniques to their students. By following this tutorial, readers will gain practical knowledge on implementing five advanced variable selection methods in R, insights into the strengths and weaknesses of each method, helping researchers to choose the most appropriate technique for their specific research question, and access to the associated code on Open Science Framework, providing an interactive learning experience. We encourage social and health sciences researchers to adopt these advanced methods, leading to more robust, interpretable models.

Specifically, the goal of this paper is to provide a tutorial on five variable selection techniques freely available to researchers in R. We will introduce the Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net, a version of the genetic algorithm (GA), and implementations of Support Vector Machines (SVMs) and Random Forest that have been adapted to perform variable selection. The manuscript is organized as follows. The first section illustrates the importance of variable selection in machine learning and explains why each of the five methods was selected. Then, a motivating example pertaining to the diagnosis of misophonia is provided. The dataset was collected from a psychology research pool and represents an excellent example of a dataset available to many behavioral and health researchers (Norris, Kimball, Nemri, & Ethridge, 2022). Within this example, there are three major sections. The first discusses methods using a logistic regression model (i.e., LASSO, EN, and the GA), the second discusses SVM, and the third pertains to random forest. Each technique is introduced, the code necessary to implement each technique is provided, and each technique's associated strengths and weaknesses are discussed. This paper is written with the assumption that individuals have at least a basic understanding of R.

1.1 Variable Selection in Machine Learning

Objectives of Variable Selection Variable selection is a fundamental step in the process of building robust and efficient machine learning models, and its

importance cannot be overstated (Chowdhury & Turin, 2020; Guyon & Elisseeff, 2003). It serves as a critical mechanism for optimizing model performance and ensuring its reliability across various tasks and datasets. The goal of variable selection (also known as feature selection in machine learning literature) is to identify the most informative (i.e., best) subset of variables for a given task. The criteria for defining "best" vary depending on the researcher's objectives, as highlighted by Huang (2015). Highlights of Huang's discussion argue that there are two main objectives of variable selection: (1) to improve the accuracy of the model, and (2) to determine the relevance of the variables in the model so as to better guide researchers' hypothesis generation.

Types of Variable Selection In the field of machine learning, variable selection techniques are often classified into one of three categories, initially discussed in the seminal paper by Guyon and Elisseeff (2003): filter methods, wrapper methods, and embedded methods .

Filter methods (e.g., χ^2, Euclidean distance, or the i-test) are often used as a pre-processing step, but they can be used as a stand-alone variable selection method. These techniques choose variables (or features) before building any model to measure the construct of interest. For example, a filter could select items based on a particular feature relevance score, a variable's correlation with the constructs of interest, or the variable's amount of variance. Most often, significance testing is used as a filter method to determine variable selection (e.g., a variable would need to correlate significantly, as determined by a p-value, with the outcome variable). However, these significance tests occur in a univariate fashion (i.e., one variable is tested at a time), which ignores possible interaction effects or covariance among variables. No filter methods are presented in this tutorial, as past research indicates they provide inferior results and miss important information as the selection is separate from model estimation (Blum & Langley, 1997; Guyon & Elisseeff, 2003; Kohavi, 1996), but we include a brief overview to provide the reader with a full picture of the types of variable selection methods that exist.

Wrapper methods improve upon filter methods by accounting for a variable's ability to measure the construct of interest. Each wrapper method operates under a specific algorithmic ideology from machine learning (e.g., stepwise regression techniques operate as greedy algorithms, choosing the variable that will optimize the selected criteria at each step). Wrapper methods are flexible in that they are not constrained to any one type of model (e.g., regression, structural equation modeling, etc.) but rather can be "wrapped" around the researcher's chosen model. The wrapper method explained in this tutorial is the genetic algorithm, which we have wrapped around a logistic regression model for classification purposes. More details about the genetic algorithm will be provided in a later section of this paper.

Embedded methods are similar to wrapper methods in how well a set of variables predicts the given construct of interest. Embedded methods differ from wrapper methods in that they perform variable selection while simultaneously

estimating the prediction model (Guyon & Elisseeff, 2003). Although this often results in higher efficiency than wrapper methods, embedded methods are constrained to one type of model. The embedded methods discussed in this tutorial are LASSO and Elastic Net which use a logistic regression classification model (Engebretsen & Bohlin, 2019), Elastic SCAD SVM which uses an SVM classifier (Becker, Toedt, Lichter, & Benner, 2011), and Boruta which uses a random forest classifier (Kursa & Rudnicki, 2010).

Variable Selection Importance Variable selection is advantageous with any model (e.g., regression, structural equation modeling, etc.) because, as mentioned previously, it leads to more accurate predictions, reduces the computational cost of the model, and improves the parsimony of the model by eliminating redundant and irrelevant variables. However, there are additional advantages to variable selection when paired with machine learning models. First, variable selection helps manage dimensionality problems (i.e., when a dataset contains more predictors than observations). Over the years, technology such as the invention of online data collection platforms like Prolific or the creation of mobile health apps has allowed researchers to collect more complex data from increasingly larger samples. As datasets grow in both size and complexity, the number of variables may also increase, leading to computational inefficiencies and reduced model interpretability (Barceló, Monet, Pérez, & Subercaseaux, 2020). By carefully selecting relevant variables, we can effectively reduce the dimensionality of the data, thereby streamlining the computational process and facilitating easier interpretation of the model (Jia, Sun, Lian, & Hou, 2022)

Moreover, the variable selection process enables models to achieve higher accuracy and better generalization capabilities. For example, van Vuuren et al. (2021) found that LASSO created a model that was able to classify students as at risk for suicide with a higher accuracy than simple inclusion rules (i.e., predicting based on history of suicide alone). Pratik, Nayak, Prasath, and Swarnkar (2022) utilized Elastic Net to select variables that were able to predict smoking addiction in young adults with higher accuracy than previous research. By focusing on the most informative variables, the model can discern meaningful patterns within the data, leading to more precise predictions and improved performance on unseen or new data. This selective approach prevents the model from being overwhelmed by noise or irrelevant information, allowing it to focus on capturing the underlying relationships that drive the outcome of interest. For example, researchers found that applying Elastic Net regularization to classifiers based on clinical notes reduced the number of features selected by more than a thousandfold, making these classifiers more easily interpretable and maintaining performance (Marafino, John Boscardin, & Adams Dudley, 2015).

Furthermore, the inclusion of irrelevant variables in the modeling process can introduce bias and adversely affect the estimation of model parameters. Additionally, extraneous variables may introduce noise or confounding factors, leading to skewed parameter estimates and potentially misleading conclusions (Kerkhoff & Nussbeck, 2019). By excluding such variables through proper selec-

tion techniques, we can ensure that the model's estimates remain unbiased and reflects the true underlying relationships in the data, increasing the ecological validity of study results and models produced.

Lastly, a well-selected set of variables enhances the model's predictive performance and contributes to its stability and reliability (Arjomandi-Nezhad, Guo, Pal, & Varagnolo, 2023; Fox et al., 2017). Models built on a carefully chosen subset of variables are less susceptible to overfitting, where the model simply memorizes the data rather than learning meaningful patterns. Avoiding overfitting leads to more robust models that generalize better and are less prone to erratic behavior or unexpected deviations, which may lead to harmful classifications (e.g., classifying an individual as having a particular disorder when they do not; Cateni, Colla, & Vannucci, 2010; Heinze, Wallisch, & Dunkler, 2018).

Put simply, variable selection is indispensable in the realm of machine learning. It serves as a cornerstone for improving computational efficiency, enhancing model accuracy and generalization, reducing bias in parameter estimation, and fostering the stability and reliability of the resulting models. As such, behavioral and health researchers must employ rigorous techniques and considerations during the variable selection process to ensure the models' and conclusions' effectiveness and generalizability.

Applications of Variable Selection Methods Understanding the appropriate contexts for applying different variable selection methods is crucial for researchers to make informed decisions. Below we outline scenarios where each of the five methods discussed in this tutorial – LASSO, Elastic Net, genetic algorithm (GA), support vector machines (SVM), and random forest – can be most effectively utilized.

LASSO is particularly effective for datasets with a large number of predictors, especially when many predictors are thought to be irrelevant or redundant (Tibshirani, 1996). It is often used in clinical research for identifying key biomarkers from extensive genetic data or in psychological students for selecting significant psychological traits that predict mental health outcomes (Chu et al., 2024; Wettstein et al., 2023). However, LASSO is constrained by degrees of freedom requirements, so, if researchers' data contains more predictors than observations, this approach would be infeasible.

Elastic Net is best suited for datasets with highly correlated predictors. It combines the strengths of both LASSO and Ridge regression, which makes it most suitable for complex datasets with multicollinearity. This method is applied in epidemiology to study the impact of multiple, correlated environmental exposures on health outcomes and in social sciences to analyze survey data where multiple questions pertaining to a given latent construct are often correlated (Han & Dawson, 2021; Pratik et al., 2022).

The genetic algorithm is ideal for complex optimization problems where traditional methods may fail to find the global optimum. It is flexible and can be adapted to various types of models and data structures. If researchers believe there may be strong interactions between variables, this approach may be most

appropriate. In fact, GA has been used in the behavioral and health sciences to explore variable selection when interactions between numerous behavioral variables are present, or hypothesized to be present, in the data (Adams, Bello, & Dumancas, 2015; Basarkod, Sahdra, & Ciarrochi, 2018; Gan & Learmonth, 2016; Moore et al., 2017; Yukselturk, Ozekes, & Türel, 2014).

SVMs are highly effective for classification problems with high-dimensional (where there are more predictors than observations) data. They are robust to overfitting, especially when an advanced kernel function (discussed in more detail later) are used. They are often used in medical diagnosis for classifying patients based on medical imaging data (Becker, Werft, Toedt, Lichter, & Benner, 2009; Fernandez, Caballero, Fernandez, & Sarai, 2011) and in classification studies such as predicting dementia (Battineni, Chintalapudi, & Amenta, 2019).

Random forest performs particularly well when data have a mix of variable types or complex interactions. It handles large datasets well, provides measures of variable importance, and is less prone to overfitting than some other approaches due to the ensemble approach. Random forest has been applied to educational psychology to assess student related outcomes (Alamri et al., 2021; El Haouij et al., 2018; Tan, Main, & Darolia, 2021). Within the health sciences, researchers have used random forest to predict cases of COVID-19, predict risk for adverse health effective, and identify longitudinal predictors of health (Cafri, Li, Paxton, & Fan, 2018; Iwendi et al., 2020; Loef et al., 2022).

Our Chosen Variable Selection Techniques Researchers have a variety of variable selection methods available to them, and many are freely available to researchers in R packages. Perhaps the most widely applicable and easy-to-use R package for variable selection is the relatively new *FSinR* package (Aragón-Royón, Jiménez-Vílchez, Arauzo-Azofra, & Benítez, 2020), which contains a large number of filter and wrapper methods widely used in the literature for both classification and regression models that are available in the R caret package (Kuchirko, Bennet, Halim, Costanzo, & Ruble, 2021). A short, non-exhaustive list of other easy-to-use R packages for variable selection is cited here for the reader's convenience (Calcagno & Mazancourt, 2010; Genuer, Poggi, & Tuleau-Malot, 2010; Kursa & Rudnicki, 2010; Strobl, Malley, & Tutz, 2009; Trevino & Falciani, 2006; Wehrens & Franceschi, 2012).

The five techniques utilized in this paper were chosen for a variety of reasons. First and foremost, LASSO and Elastic Net are arguably the most popular modern variable selection techniques within the behavioral sciences. The implementations used in this tutorial come from the *glmnet* R package (Friedman, Hastie, & Tibshirani, 2010; Tay, Narasimhan, & Hastie, 2023). Social psychology researchers have used such techniques to create better environments that promote prosocial environments for children (Chu et al., 2024), and health researchers have used them to model the progression of Alzheimer's disease (Liu, Cao, Gonçalves, Zhao, & Banerjee, 2018). Implementations of SVM and random forest were chosen because of their strength as classification algorithms and because they can handle more complex data types (e.g., mixed variable types or

non-linearly separable). The SVM implementation comes from the *penalizedSVM* package (Becker et al., 2009) while the random forest implementation comes from the *Boruta* package (Kursa & Rudnicki, 2010). Lastly, the GA was chosen (1) to introduce the reader to the concept of metaheuristic approaches to variable selection and (2) because it has been shown to outperform more common methods like LASSO and Elastic Net across a variety of different data conditions (Bain, Shi, Boness, & Loeffelman, 2023). The GA implementation comes from the *GA* package (Scrucca, 2013, 2017). Note that while this paper includes core code snippets, the accompanying Open Science Framework (OSF) repository provides the complete code and data necessary to replicate all analyses. The repository link is provided in the data availability section.

A Motivating Example This tutorial uses the assessment of misophonia as an example through which we illustrate each technique. Individuals with misophonia experience strong, negative, emotional responses to specific sounds (i.e., triggers Wu, Lewin, Murphy, & Storch, 2014). The original data sample consisted of undergraduate students ($N = 343$) at a large southwestern university. Participants were predominately white (76.7%), female (69.7%), and students (96.5%) ranging from ages 18 to 36 ($M = 18.96$, $SD = 1.7$). The dataset contains 106 independent variables related to both direct characteristics of misophonia and related characteristics, as well as one self-report binary diagnosis variable. It is available to the reader on the accompanying OSF repository linked in the availability of data and materials section of this paper. Since misophonia is still not fully understood (i.e., formal diagnostic criteria have not been set, and researchers are still trying to determine the most important symptoms), this dataset is an illustrative example of variable selection. Some symptoms may be unimportant for, or not predictive of, a true misophonia diagnosis. One should note that this dataset does not contain any missing data, as it was handled *a priori* using listwise deletion. In addition, one should note that the group sizes are unbalanced (16.5% diagnosed, 83.5% not). This presents additional complexity and is one reason why we have chosen to evaluate the methods using both accuracy and F-score. For more information on the larger previously published dataset from which this data was selected and the background on misophonia, see the work of Norris et al. (2022).

The Importance of Cross Validation. Model overfitting is a common problem for implementing variable selection techniques (see Figure 1). If a model is built too closely to the specifications of a specific dataset (i.e., it is not robust to changes in the data), it is considered overfit. Alternatively, a model can be underfitted where it is built in such a way that it is too generalizable and does not create accurate or meaningful predictions. Researchers need to be cautious of overfitting and underfitting to ensure that they build models that can accurately generalize to new data while making meaningful and accurate predictions.

Cross-validation is one common way to help researchers increase generalizability in a meaningful way (i.e., protect against overfitting). In cross-validation, the model is built on (or, in the case of this tutorial, variables are selected from)

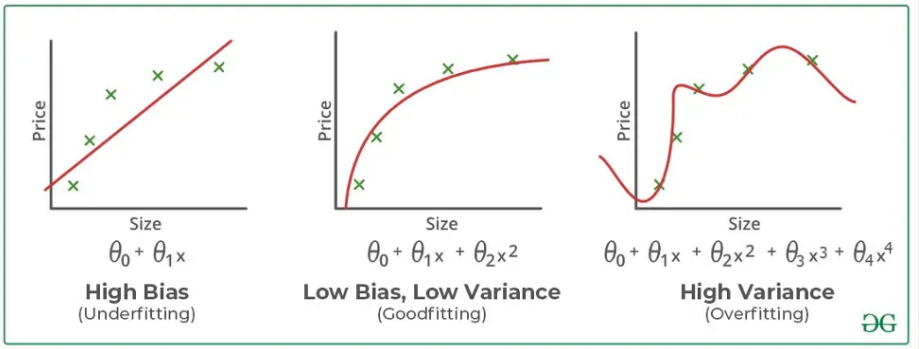


Figure 1. The leftmost graph illustrates an underfit model on a small amount of data. The middle figure illustrates a fit that balances both bias and variance leading to good fit. The rightmost graph illustrates an overfit model. Figure obtained from Geeks for Geeks (*ML | Underfitting and Overfitting*, 2017)

a different set of data than it is evaluated. Although this can occur through the collection of two different datasets, this is typically done by dividing one dataset into parts. One can do this division in many ways, and this paper implements holdout cross-validation, which occurs when one splits the data into two sets (test and training sets) before conducting any analyses. Typically, 70% of the data is used for the training set in holdout cross-validation, and the remaining 30% is used for the test dataset. The code for how we performed holdout cross-validation can be found in the companion code on OSF. For additional information on the importance of cross-validation and alternative approaches to cross-validation, see the helpful tutorials cited here (Ghojogh & Crowley, 2023; Song, Tang, & Wee, 2021).

2 Methods

2.1 Logistic Regression Models

Logistic regression is a widely used statistical model for binary classification problems and models the probability that a given observation (e.g., a set of participants' responses to a given questionnaire), belongs to a particular category. The equation for logistic regression is :

$$P(Y = 1|\mathbf{X}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m)}} \tag{1}$$

Here, $P(Y = 1|\mathbf{X})$ represents the probability that the participant belongs in class 1 given their response matrix ($\mathbf{X}$). The intercept term (β_0), is the value of the log-odds when all predictor variables are zero. The coefficients ($\beta_1, \ldots, \beta_m$) associated with each of the predictor variables ($x_1, \ldots, x_{\Updownarrow}$) represent the change in the log-odds of the dependent variable for a one-unit change in the corresponding predictor variable for a total of m predictors.

Regularization Techniques Two of the techniques discussed in this paper, LASSO and Elastic Net, are regularization techniques. Regularization is a common method used to combat issues of overfitting found in models estimated with maximum likelihood estimation (like logistic regression). Each regularization technique works to combat overfitting by intentionally introducing a small amount of bias into the model such that a generic regularization function, within the context of classification, takes the following form:

$$L^{Reg}(\boldsymbol{\beta}) = L^{logistic}(\boldsymbol{\beta}) - \lambda \mathscr{P}(\boldsymbol{\beta}) \tag{2}$$

where L^{Reg} is the penalized optimization function, $L^{logistic}$ is the negative log likelihood, λ is a regularization parameter (i.e., a tuning parameter), and $\mathscr{P}$ is a penalty function that will vary across the regularization technique. The goal of regularization is to find the optimal balance between bias (generalizability of the model) and variance (specific model fit Helwig, 2017). The magnitude of the lambda (λ) penalty determines this balance. A larger lambda will lead to a sparser and more generalizable model. One popular technique utilized to determine the value of the lambda parameter is cross-validation. As mentioned above, cross-validation occurs when the data is split into multiple subsets, the model is developed (i.e. trained) on a subset, and evaluated (i.e., validated) on another. This process is iterative, allowing for the selection of the lambda penalty that minimizes prediction error across different subsets.

One optimal model, in the context of this paper, is one that produces the most accurate classifications. Accuracy can be calculated using the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP is the number of individuals who were correctly classified as having a diagnosis of misophonia, TN is the number of individuals who were correctly classified as not having a diagnosis of misophonia, FP is the number of individuals who were classified as having a diagnosis but did not truly have a diagnosis in the labeled data, and FN are the number of individuals who were incorrectly classified as not having a diagnosis when a diagnosis was present in the labeled data. It is worth noting that accuracy may not be the best optimization criteria given the unbalanced nature of the data (i.e., the number of observations in class 0 is much larger than the number in class 1). In practice, researchers may want to use a weighted accuracy or an F-score in their own research, depending on the relative importance of a false positive versus a false negative. For example, a clinician attempting to predict suicide attempts may prioritize a false positive (i.e., saying the individual is likely to attempt suicide when they do not actually attempt) over a false negative (i.e., saying the individual will not attempt when they actually will). Non-weighted accuracy was included for ease of explanation. However, we will also evaluate each model in terms of an F-score to illustrate the differences between these metrics. The equation for calculating an F-score is seen below.

$$F1 = \frac{TP}{TP + .5(FP + FN)} \tag{4}$$

The F1 score is a measure of a model's ability to balance precision (accuracy of positive predictions) and recall (correct identification of positive instances). The equation provided modifies the traditional F1 score by scaling the sum of false positives (FP) and false negatives (FN) by 0.5, reducing their weight in the final score. This adjustment can be useful when false positives and false negatives are not equally important or should be penalized less.

LASSO. LASSO (Tibshirani, 1996) is one of the penalized regression techniques that perform variable selection. LASSO can handle data with multicollinearity, be applied to various types of data (e.g., continuous, categorical, mixed type), and is adaptable to sparse data (i.e., multiple predictors have zero or near-zero coefficients; Foucart, Tadmor, & Zhong, 2023; Mendez-Civieta, Aguilera-Morillo, & Lillo, 2021). The parameter estimates (i.e., the $\boldsymbol{\beta}$ coefficients) for LASSO can be obtained by maximizing the penalized log-likelihood function:

$$L^{LASSO}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \mathbf{x}_i \boldsymbol{\beta} - log(1 + e^{\mathbf{x}_i \boldsymbol{\beta}})] - \lambda \sum_{j=1}^{m} |\beta_j| \tag{5}$$

where $L^{LASSO}(\boldsymbol{\beta})$ is is the loss function and is comprised of two summations. The first summation represents the logistic regression log likelihood and n is the number of observations in the data, y_i represents the actual binary outcome of the i-th observation, $\mathbf{x}_i$ is the vector of predictor variables for the i-th observation, $\boldsymbol{\beta}$ is the vector of coefficients (including the intercept term), and $log(1 + e^{\mathbf{x}_i \boldsymbol{\beta}})$ is the log of the logistic function denominator, which ensures that the probabilities are correctly bounded between 0 and 1. The second summation is the LASSO penalty (or the ℓ_1 regularization term) which adds a penalty proportional to the absolute value of the coefficients and m is the number of predictors in the initial model. Here λ is the regularization hyperparameter that controls the degree of shrinkage such that larger values lead to the selection of fewer variables and $\sum_{j=1}^{m} |\beta_j|$ is the sum of the absolute values of the coefficients for all predictor variables (note that the summation begins at 1, indicating that the intercept, β_0, is excluded from regularization and must be included in the final. For a more detailed discussion of LASSO, see Tibshirani's (1996) paper . Regularization techniques are useful for variable selection because they add a penalty for large coefficients, effectively shrinking less important variables towards zero and thus eliminating them from the model. This helps in improving model interpretability and preventing overfitting, particularly in scenarios with a large number of predictors or multicollinearity.

As with all methods, researchers may be interested in the recommended sample size LASSO. One conservative estimate suggests that researchers should have 10 observations per candidate variable (e.g., with 10 variables, a researcher would need 100 observations; Peduzzi, Concato, Feinstein, & Holford, 1995; Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996). However, this recommendation is made more generally for regression, and thus, does not generalize as specifically to regularization techniques where not all variables are included in the final model. Recent simulation studies have investigated the performance of LASSO in

small sample sizes (e.g., 50 – 100 participants) and found that methods perform well (Bain et al., 2023; Kirpich et al., 2018; Wen et al., 2019).

To utilize LASSO for variable selection, we use the `cv.glmnet()` function from the *glmnet* package in R (Friedman et al., 2010). More information on the hyperparameters of the function can be found in Table 1. This function determines the magnitude of lambda through a k-fold cross-validation approach.

```
lasso.model <- cv.glmnet(x = predTrain,
    y = outcomeTrain, type.measure = "class",
    alpha=1, family="binomial", nfolds = 10)
```

Through this model, we can obtain the chosen lambda value. To obtain a full list of all evaluated lambda values, use `lasso.model$lambda`. One can also plot the k-fold cross-validation procedure to obtain λ using `plot(lasso.model)` (Figure 2). There are two lambda values that are particularly of interest. The first can be obtained with `lasso.model$lambda.min`. This lambda value is responsible for producing the model with minimal cross-validated error. The second can be obtained with `lasso.model$lambda.1se`, or the 1se rule. This lambda value is responsible for producing the model that has a cross-validated error within one standard error of the minimum. There are advantages to each. Breiman and colleagues (2017) as well as Chen and Yang (2021) suggest that researchers should use the 1se rule to select lambda to reduce the instability of the model while maintaining a parsimonious model. However, this gain in stability comes with a loss in accuracy (an increase in misclassification error of one standard error). In addition, some research has shown that the 1se rule performs poorly in regression (Chen & Yang, 2021) as opposed to a classification tree, so we used the value that minimized cross-validation error (lambda min). To obtain our lambda min value, specify `lasso.model$lambda.min`. Using this specified lambda value, we can build a LASSO model using the `glmnet()` function with the following code, which will produce the coefficients as seen in Table 2.

```
lasso.model.min <- glmnet(x = predTrain,
    y = outcomeTrain, alpha=1,
    family="binomial",
    lambda = lasso.model$lambda.min)
```

Out of the original 106 predictor variables, only 16 were selected via LASSO, thus a sparse model has been obtained. It is important to examine what variables were selected by the model to ensure that they are theoretically justified. Ideally researchers would make this decision about all variables, however, for the sake of space within this paper, we have chosen to only examine two of the 16 selected items. One selected item, MQ4 reads: "In comparison to other people, I am sensitive to the sound of people making nasal sounds." As nasal and throat sounds are often thought to be triggers for those with misophonia, this item makes theoretical sense to be a predictor of the diagnosis. For another selected item, S5_7, participants were asked, "Please rate your typical reaction to the following stimuli, if produced by another person: Throat clearing." This item is

Table 1. Hyperparameters of the `cv.glmnet()` function and their corresponding definitions.

Parameter	Description
x	A matrix of predictor (or input) variables.
y	The vector containing the response (or outcome) variable.
type.measure	The optimization measure to be used within the internal cross-validation procedure. By setting this to "class" misclassification error is optimized.
alpha	The Elastic Net mixing hyperparameter. Because the same function is used to implement ridge, LASSO, and Elastic Net, the value for alpha determines which regularization technique is run. Alpha is constrained between 0 and 1, with a value of 0 implementing ridge regression, 1 implementing LASSO regression, and anything in between implementing an Elastic Net regression.
family	The type of regression to be implemented. By setting this hyperparameter to "binomial" an MLE regression is implemented.
nfolds	The number of partitions implemented in the internal k-fold cross-validation.

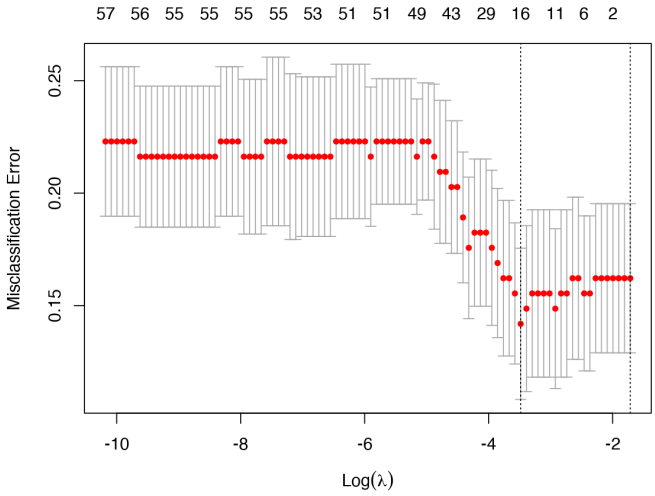


Figure 2. Cross-validated estimate of the mean squared prediction error for LASSO as a function of the log λ. The upper axis indicates the number of non-zero coefficients in the regression model at the given log λ. The dashed vertical line illustrates the location of the CV minimum and the one standard error rule locations for λ.

theoretically justifiable for the same reason as above, reactions to throat sounds are a symptom of the disorder.

The coefficient estimates obtained through a LASSO approach are biased by the nature of the algorithm (Yarkoni & Westfall, 2017), and thus research

Table 2. A table containing the variables selected by the LASSO model and their corresponding estimated coefficients. The table also includes the full item for that particular variable. If items share a common stem, we have grouped them together.

Variable	Coefficient	Full Item
(Intercept)	-5.741	
MQ4	0.154	In comparison to other people, I am sensitive to the sound of people making nasal sounds (e.g., inhale, exhale, sniffing, etc.).
Once you are aware of the sound(s), because of the sound(s), how often do you:		
MQ11	0.039	Cover your ears?
MQ12	0.137	Feel anxious or distressed?
MQ13	0.112	Become sad or depressed?
MQ17	0.045	Become physically aggressive?
Please rate your typical reaction to the following stimuli, if produced by another person:		
S5_7	0.087	Throat clearing
S5_24	-0.032	Car engine
S5_25	0.318	Clock ticking
S5_31	0.159	Pacing
S5_32	0.024	Nail biting
S5_35	0.123	Strong smells
S5_36	0.089	Seeing someone chew gum
Please indicate your level of agreement to the following statements:		
S5_56	0.124	I can feel physical pain if I cannot avoid a sound.
S5_57	0.419	Sometimes in response to sounds I feel rage that is difficult to control.
S5_75	0.252	Some sounds have caused me to use violence towards myself or others.
S5_78	-0.018	It does not matter who is making the sounds, my reactions are the same.

recommends recalculating them using a standard regression before interpreting the coefficients of the model. To do that, one could use the following code.

```
selected <- trainDat %>% select(MQDX, MQ4, MQ11,
    MQ12, MQ13,MQ17, S5_7,S5_24, S5_25, S5_31,
    S5_32, S5_35, S5_32,S5_35, S5_36, S5_56,
    S5_57, S5_75, S5_78)
logistic.model <- glm(MQDX ~ .,
    family=binomial(link = "logit"),
    data = selected)
```

In this code, we first use the `select()` function from the *dplyr* package to select only the variables with non-zero coefficients in the lasso.model.min as well as our outcome variable, MQDX (Wickham et al., 2023). We then use these variables to build a standard logistic regression model using the `glm()` function.

A comparison of the biased coefficients obtained from the LASSO model and the corrected coefficients obtained in the standard logistic model can be seen in Table 3. Obtaining the predicted classification prior to calculating accuracy is crucial. Accuracy values (Equation 3) are then determined using the coefficients estimated from both the LASSO model (incorrectly biased) and the logistic model. The following code can be used to obtain the accuracy values from the logistic model as well as the F-score from the model. Note that the F-score is obtained using the `F1_Score()` function from the *MLmetrics* package (Yan, 2024).

Table 3. A table containing the variables selected by the LASSO model and the coefficient estimates obtained directly from the LASSO model as well as the re-estimated (non-biased) coefficients obtained by creating a typical logistic model using the selected variables.

Variable	LASSO Estimate	Logistic Estimate
(Intercept)	-5.741	-8.802
MQ4	0.154	0.361
MQ11	0.039	0.480
MQ12	0.137	0.760
MQ13	0.112	-0.193
MQ17	0.045	0.143
S5_7	0.087	-0.057
S5_24	-0.032	-1.154
S5_25	0.318	1.051
S5_31	0.159	0.538
S5_32	0.024	0.245
S5_35	0.123	0.110
S5_36	0.089	0.285
S5_56	0.124	0.387
S5_57	0.419	0.867
S5_75	0.252	0.186
S5_78	-0.018	-0.600

```
pp.logistic <- predict(logistic.model,
    data.frame(predTest),
    type = "response")
pc.logistic <- ifelse(pp.logistic > .5, 1, 0)
a.logistic <- mean(outcomeTest == pc.logistic)
```

```
f1.logistic <- F1_Score(pc.logistic, outcomeTest)
```

In the first line of code, using the `predict()` function, the `logistic.model` object and our `predTest` data (reminder that this is the holdout sample created during cross-validation earlier) we can create our predictions. By specifying `type = "response"`, the function will return predicted probabilities. In our second line, the predicted probabilities are transformed into predicted classes such that if the probability of them belonging to class 1 is at least 0.5, they are assigned to class 1 otherwise class 0. The third line calculates accuracy. The value obtained using the coefficient estimates from the LASSO model is an accuracy score of 0.86. The value obtained using the coefficient estimates from the logistic model is 0.89. The F-score for both the LASSO model and the logistic model is 0.92. Note that the accuracy changes across models, but the F-score remains the same. This indicates that the models likely differ only in their true negative results, as that measure is not included in the calculation of the F-score.

Despite the strong performance of LASSO on this data, LASSO does have limitations (Algamal & Lee, 2015). First, it is unable to select more variables than there are observations. Second, LASSO will select a single variable in the presence of multicollinearity regardless of that variable's predictive capacity. Zou and Hastie (2005) proposed a new regularization technique called Elastic Net to combat these first two limitations.

Elastic Net. Elastic Net differs from LASSO through the use of an additional penalty to the regression equation. Elastic Net implements both the ℓ_1 penalty, or the LASSO penalty, and the ℓ_2 penalty, or the ridge penalty, to the regression equation. With the inclusion of both penalties, the optimization function for Elastic Net is as follows:

$$L^{ElasticNet}(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i \mathbf{x}_i \boldsymbol{\beta} - log(1 + e^{\mathbf{x}_i \boldsymbol{\beta}}] - \lambda_1 \sum_{j=1}^{m} \beta_j^2 - \lambda_2 \sum_{j=1}^{m} |\beta_j| \quad (6)$$

The first summation represents the log likelihood and is exactly the same as was seen in Equation 4. The second summation is new to the reader as it is the ridge penalty, which adds a penalty proportional to the squared value of the coefficients (Hoerl & Kennard, 1970). Here λ_1 is the regularization hyperparameter that controls the degree of shrinkage such that larger values lead to the selection of fewer variables. The third summation is the LASSO penalty, which only differs from Equation 4 in that we now use λ_2 (instead of just λ) to denote the regularization hyperparameter that controls the degree of shrinkage from the LASSO penalty. The values for λ_1 and λ_2 can be equal or can be set to different values to allow differential application of the penalties. By incorporating the ridge penalty, Elastic Net can select multiple correlated variables while removing irrelevant ones (Algamal & Lee, 2015). For more on the ridge penalty, see work by McDonald (2009). This makes Elastic Net more suitable than LASSO for datasets with highly correlated predictors, such as dummy-coded variables.

Sample size considerations should also be made when researchers are considering using Elastic Net. The recommendations are similar to those for LASSO in

that conservative estimates suggests that researchers should have 10 observations per candidate variable similar to LASSO (Peduzzi et al., 1995, 1996). However, this recommendation comes from the general regression literature, and thus, may not hold with regularization. Recent simulation studies have investigated the performance of Elastic Net in small sample sizes (e.g., 50 – 100 participants) and found that methods perform well (Bain et al., 2023; Kirpich et al., 2018; Wen et al., 2019).

We can obtain our `lambda.min` value using the `cv.glmnet()` function, just as we did for LASSO. However, we change the value for alpha from alpha = 1 to alpha = 0.5. We can then use this value to build our final Elastic Net model (the second piece of code below). We can then use the variables with non-zero coefficients from our final Elastic Net model (`en.model.min`) to build a standard logistic regression model (`logistic.en.model`) to get unbiased coefficients, as was done for LASSO. Two of the selected items include MQ11, and S5_11. MQ11 reads, "Once you are aware of the sound(s), because of the sound(s), how often do you actively avoid certain situations, places, things, and/or people in anticipation of the sound(s)." Individuals with misophonia are known to employ a variety of coping strategies (including avoidance) to deal with their triggering sounds, so this variable makes sense theoretically. S5_11 reads, "Please rate your typical reaction to the following stimuli, if produced by another person: Repetitive barking." This item is interesting, because some research has found that not all sounds must be human made to be triggers for individuals with misophonia, for example, this is a sound most often made by dogs, not people. However, it is theoretically sound.

```
elasticNet <- cv.glmnet(x = predTrain,
    y = outcomeTrain, type.measure = "class",
    alpha=0.5, family="binomial", nfolds = 10)
en.model.min <- glmnet(x=predTrain y=outcomeTrain,
alpha=0.5, family="binomial",
    lambda = elasticNet$lamda.min)
selected <- trainDat %>% select(MQDX, MQ4, MQ11,
    MQ12, MQ13, MQ15, MQ16, MQ17, S5_2, S5_7, S5_11,
    S5_24, S5_25, S5_27, S5_31, S5_32, S5_35, S5_36,
    S5_38, S5_40, S5_42, S5_53, S5_56, S5_57, S5_68,
    S5_74, S5_75, S5_78, S5_82)
logistic.en.model <- glm(MQDX ~.,
    family=binomial(link = "logit"),
    data = selected)
```

Coefficient estimates from the Elastic Net model and unbiased coefficients from a standard logistic model can be seen in Table 4. An accuracy of 0.88 was obtained using the coefficient estimates from the Elastic Net model, while an accuracy of 0.80 was obtained using the coefficient estimates from the logistic model. The F-score obtained using the coefficient estimates from the Elastic Net model is 0.94, while the logistic model produces an F-score of 0.88. The code below illustrates how to obtain the 0.80 accuracy value and 0.88 F-score from

the unbiased logistic regression model. The only change the reader would need to make to obtain the estimates from the final Elastic Net model instead would be to substitute `en.model.min` for `logistic.en.model`.

```
pp.en.logistic <- predict(logistic.en.model,
    data.frame(predTest), type = "response")
pc.en.logistic <- ifelse(pp.logistic > .5, 1, 0)
a.en.logistic <- mean(outcomeTest == pc.logistic)
f1.en.logistic <- F1_Score(pc.en.logistic,
    outcomeTest)
```

Table 4. A table containing the variables selected by the Elastic Net model and their corresponding estimated coefficients obtained directly from the Elastic Net model as well as the coefficients estimated by implementing a logistic model (non-biased coefficients).

Variable	Elastic Net Estimate	Logistic Estimate
(Intercept)	-5.796	-1732.902
MQ4	0.133	15.606
MQ11	0.046	20.788
MQ12	0.119	81.502
MQ13	0.103	-6.765
MQ15	0.057	101.386
MQ16	0.075	5.368
MQ17	0.086	145.615
S5_2	0.023	-6.618
S5_7	0.091	12.560
S5_11	-0.051	-190.866
S5_24	-0.095	-39.021
S5_25	0.262	31.171
S5_27	0.031	94.479
S5_31	0.165	58.559
S5_32	0.092	97.889
S5_35	0.147	56.132
S5_36	0.088	38.234
S5_38	0.036	30.691
S5_40	0.048	111.532
S5_42	0.032	42.788
S5_53	-0.020	12.132
S5_56	0.190	94.590
S5_57	0.259	-87.586
S5_68	0.081	126.088
S5_74	0.018	5.356
S5_75	0.197	-17.256
S5_78	-0.089	-101.315
S5_82	0.033	-2.060

Elastic Net also has some limitations. Namely, it may struggle with datasets containing many more variables than observations, it is sensitive to outliers, and, given that it is designed for linear relationships, it may not capture complex non-linear relationships between predictors and the response variable effectively (Wang, Cheng, Liu, & Zhu, 2014).

Genetic Algorithm (GA) Unlike LASSO and Elastic Net, which utilize internal regression models as embedded methods, the genetic algorithm (GA) operates as a wrapper method. As a reminder, this means that the user must specify which model it should use (i.e., a user could wrap the GA around a logistic regression model or something more complex like a random forest or SVM, depending on the nature of their data). As mentioned, wrapper methods each follow their own algorithmic strategy to explore potential solutions (i.e., potential sets of variables to select). One wrapper method that may be familiar to readers is stepwise regression, which builds a model iteratively by either adding or removing variables based on a given criteria (e.g., Akaike Information Criteria; AIC). It uses a greedy approach, selecting the variable at each step that yields the greatest immediate improvement in the chosen criterion (e.g., the largest decrease in AIC). The GA also operates a greedy algorithm; however, its search strategy differs.

Instead of adding or removing a single variable (as is done in stepwise regression), the GA, inspired by the principles of natural selection and evolution, mimics the process of biological evolution to refine potential solutions iteratively. Through crossover, mutation, and selection mechanisms, the GA explores and evolves a population of potential solutions over successive generations, gradually improving the overall quality of solutions. Figure 3 illustrates the general structure of the genetic algorithm, depicting its iterative process of generating, evaluating, and evolving solutions. Each iteration refines the population, guiding the search towards promising regions of the solution space.

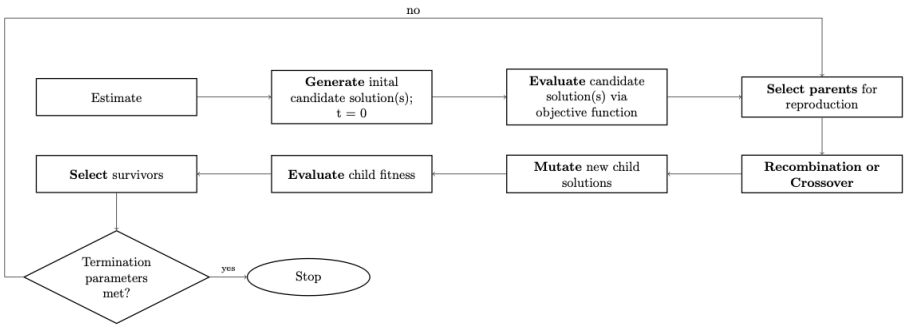


Figure 3. The basic algorithmic steps of the Genetic Algorithm.

For a comprehensive understanding of the genetic algorithm and its application to variable selection, interested readers are encouraged to refer to the work by Bain et al. (2023). Their research provides detailed insights into the underlying principles, implementation strategies, and practical considerations associated with the GA's use in solving two-group classification problems.

There are no accepted sample size recommendations for the GA for variable selection. The required sample size depends heavily on the complexity of the underlying model, the number of predictors, and the strength of the signals. As a rough guideline, samples sizes in the range of 100-500 are often used, but larger samples may be necessary for high-dimensional problems (Cateni et al., 2010; Leardi, 2000).

For this paper, logistic regression is chosen as the model around which the GA will wrap. The optimization function used in this paper is the Hubert and Arabie (1985) Adjusted Rand Index (ARI). ARI is a measure of agreeability between predicted classifications and true (or known) classifications and can be calculated in the following way:

$$ARI = \frac{RI - RI_{Expected}}{max(RI) - RI_{Expected}} \tag{7}$$

$$RI = \frac{a+d}{a+b+c+d} \tag{8}$$

$$RI_{Expected} = \frac{2(a+b)(a+c)}{(a+b+c+d)^2} \tag{9}$$

Here, a is the number of pairs of individuals (or observations) that are in the same class in both the true labels and the predicted labels, b is the number of pairs of individuals that are in the same class in the true labels but are in different classes in the predicted labels, c is the number of pairs of individuals that are in different classes in the true labels but are in the same class in the predicted labels, and d is the number of pairs of individuals that are in different classes in both the true labels and predicted labels.

The implementation of the GA used in this tutorial comes from the `ga()` function in the GA package (Scrucca, 2013, 2017). To implement the GA, the following code can be run:

```
ga.solution <- ga(fitness = function(vars)
   gaOpt(vars=vars, IV.train=data.frame(predTrain),
        DV.train=outcomeTrain),
        type = "binary", nBits = ncol(predTrain),
        names = colnames(predTrain), seed = 123456,
        run=5
)
```

Here, we set type to binary to indicate that we want binary representations of decision variables. This hyperparameter may need to change depending on the nature of the variables of interest. Second, we set `nBits` to be equal to

the number of predictor variables to indicate that all variables in the dataset could be selected. A seed is set for reproducibility. The run hyperparameter has been set to five, indicating that the algorithm should terminate if there is no improvement in the optimization function after five iterations. Note that one of the parameters in this function is the `gaOpt()` function. The `gaOpt()` function is a self-defined, user-specified function that could take on a different name. However, regardless of the name, the function must be passed as a hyperparameter in the `ga()` function. The R code needed to implement this optimization function with a logistic regression model can be seen below. For more information on the hyperparameters of the GA function and their default values, see Table 5.

Table 5. A table containing the hyperparameters of the `ga()` function and their corresponding definitions and default values.

Parameter	Description
fitness	The hyperparameter containing the optimization function is passed. No default is set.
type	The type of ga that needs to be run is dependent upon the nature of the outcome variable. "binary" is selected.
crossover	The type of crossover performed. The default for a binary implementation is found via the `ga_Crossover()` function.
popSize	An R function to generate the initial population. To access available functions, run `ga_Population()`.
pcrossover	The probability of crossover, default of 0.8 is used.
pmutation	The probability of mutation, default of 0.1 is used.
elitism	The number of best fitted chromosomes to survive at the end of each generation, default of `max(1, round(popSize*0.05))` is used.
nBits	A value specifying the number of bits in a potential solution, set equal to the number of predictors.
names	The variable names.
maxIter	The maximum number of iterations to run before the GA search is halted, default of 100 is used.
keepBest	A logical argument specifying if best solutions at each iteration should be saved, default FALSE.
seed	A number allowed to control randomness for reproducibility.
run	The number of consecutive generations that can occur without any improvement before the GA is halted, default is modified from maxiter to 5.

```
gaOpt <- function(vars, IV.train, DV.train){
varNames <- colnames(IV.train)
selectedVarNames <- varNames[vars == "1"]
gaSolutionData <- IV.train[,selectedVarNames]
    gaDat <- cbind(gaSolutionData, DV.train)
    gaMod <- glm(DV.train ~ ., family = "binomial",
        data = gaDat)
```

```
gaProbabilities <- predict(gaMod, IV.train,
        type = "response")
gaPredictedClasses <-
        ifelse(gaProbabilities >= .8, 1, 0)
ari <- adjustedRandIndex(gaPred, DV.train)
return(ari)
}
```

The `glm()` function is the same function we used to calculate logistic regression models previously. The `adjustedRandIndex()` function comes from the *mclust* package (Scrucca, Fop, Murphy, & Raftery, 2016). The `gaOpt()` function takes us through the steps of finding the ARI for the selected subset of variables. First, the names of all candidate variables are acquired, then the names of the variables selected by the GA are found, and we select only those columns from our train data. Since we had previously removed the dependent variable (the misophonia diagnosis) from the dataset, we must recombine our selected variables and our outcome variable into one matrix (line 5 above, here called `gaDat`). Next, the logistic regression model is built using these selected variables. Then the predicted probabilities are obtained, transformed into predicted classes (such that an individual is given a positive misophonia diagnosis if their probability of diagnosis is at least .8, which was chosen because only about 20% of our sample belongs to class 1). Finally, the ARI of the model is calculated and returned to the `ga()` function. To view the selected subset of variables from the `ga()` function, one calls, `ga.solution@solution[1,]`. Note, the returned solution (given by `ga.solution@solution`) contains many potential subsets of variables, but by referencing only the first row (using the indexing `[1,]`), the optimal subset of variables as determined by the GA can be accessed. Two of the selected items include item MQ18 and S5_3. Variable MQ18 reads, "Once you are aware of the sound(s), because of the sound(s), how often do you become physically aggressive" which is theoretically justifiable as individuals with misophonia are known to have disproportional, often violent, reactions to their triggers. Variable S5_3 reads,"Please rate your typical reaction to the following stimuli, if produced by another person: Swallowing," which is justifiable as it pertains to throat noises.

```
allVarNames <- colnames(predTrain)
selectedVarNames <-
    allVarNames[ga.solution@solution[1,]==1]
selectedVars <-
    data.frame(predTest[,selectedVarNames],
    outcomeTest)
ga.model <- glm(outcomeTest~., family="binomial",
data=selectedVars)
```

Since the `ga()` function does not have a specified method for model building, but rather simply returns a list of variable selections, one must first build a model to obtain an accuracy value for the selected variables. Given that the internal model we specified was a logistic regression model, it makes sense to use a simple

logistic model, which can be built using the following code. The coefficients from this model can be seen in Table 6. After building the model, an accuracy and F-score can be obtained using the following code:

```
p <- predict(ga.model, newx = predTest)
c <- ifelse(p >= .8, 1,0)
accuracy <- mean(c == outcomeTest)
f1 <- F1_Score(c,outcomeTest)
```

Table 6. A table containing the variables selected by the GA and their corresponding estimated coefficients in the logistic regression model.

Variable	Coefficient	Variable	Coefficient
(Intercept)	72.896	S5_42	-8.713
MQ4	-7.952	S5_45	8.668
MQ6	-3.454	S5_46	-10.066
MQ8	-5.707	S5_49	-1.448
MQ17	-6.154	S5_50	5.708
MQ18	21.166	S5_51	-0.198
S5_2	9.767	S5_52	-8.592
S5_3	-3.703	S5_53	-2.791
S5_4	-0.814	S5_55	-13.721
S5_6	3.907	S5_57	3.470
S5_7	-18.858	S5_58	-2.086
S5_8	1.915	S5_60	-16.660
S5_9	14.258	S5_62	4.408
S5_10	10.305	S5_63	5.143
S5_11	-11.589	S5_64	-0.372
S5_12	43.946	S5_65	4.143
S5_13	-36.764	S5_66	-8.781
S5_18	10.628	S5_68	-13.752
S5_19	2.410	S5_69	7.001
S5_20	-6.446	S5_72	12.343
S5_21	-0.442	S5_73	19.055
S5_23	3.657	S5_76	-9.715
S5_25	2.824	S5_77	-4.178
S5_26	1.490	S5_78	5.347
S5_27	4.120	S5_79	-7.315
S5_31	-4.880	S5_81	7.567
S5_32	-14.408	S5_83	-4.304
S5_33	-11.206	S5_84	-1.235
S5_38	5.880	S5_86	5.970
S5_41	11.462	S5_87	-14.504

Accuracy and F-score values of 1 are obtained, indicating a perfect fit, as with the past models built in this tutorial. Current literature indicates that the GA is prone to overfitting (Frohlich, Chapelle, & Scholkopf, 2003; Leardi, 2000;

Loughrey & Cunningham, 2005), suggesting the model would not fit quite as well if a new sample was collected, despite the accuracy of the model fit for the test sample used in this tutorial.

2.2 Support Vector Machines

Support Vector Machines (SVM) are a class of supervised learning models widely employed in classification and regression tasks (Fernandez et al., 2011; Karatzoglou, Meyer, & Hornik, 2006). SVMs operate by finding the optimal hyperplane that maximizes the margin between different classes of data points. By maximizing the margin between classes, SVM achieves good generalizability and is robust to outliers (Singla & Shukla, 2020; Xu, Caramanis, & Mannor, 2009). SVM can handle both linearly separable and non-linearly separable data by using a kernel function that artificially projects the original data into a higher-dimensional space (Karatzoglou et al., 2006).

Elastic SCAD SVM SVM, by itself, is a classification algorithm. However, researchers have created implementations of SVM that simultaneously perform classification and variable selection (Becker et al., 2011; Bierman & Steel, 2009; Tharwat & Hassanien, 2019). This tutorial uses an approach like LASSO and Elastic Net in that it selects variables via the addition of a penalty that comes from the *penalizedSVM* package (Becker et al., 2011). The penalty utilized in this tutorial is the Elastic smoothly clipped absolute deviation (SCAD) penalty, which when included in an SVM, reads:

$$SVM_{ESCAD} = min_{b,w}[sign(\mathbf{w}^T\mathbf{x} + \mathbf{b}) + \sum_{j=1}^{p} \mathscr{P}_{SCAD}\lambda_1(\mathbf{W}_j) + \lambda_2\|w\|_2^2] \quad (10)$$

where λ_1 controls the degree of shrinkage applied by the SCAD ($\mathscr{P}_{SCAD}\lambda_1(\mathbf{W}_j)$) penalty and λ_2 controls the degree of shrinkage applied by the Elastic Net ($\lambda_2\|w\|_2^2$) penalties. Higher values of either λ increase the degree of shrinkage applied by their given penalty. For more information on the SCAD penalty, see work by Becker et al. (2011). Just as with Elastic Net, the λ_1 and λ_2 values can be equal or set individually to differentially apply the penalties. The initial part of the equation ($sign(\mathbf{w}^T\mathbf{x} + \mathbf{b})$ is the base equation for an SVM where $\mathbf{w}$ is the weight vector, $\mathbf{x}$ is the input feature vector, $\mathbf{b}$ is the bias term vector, $sign(.)$ is the sign function, which returns +1 if the argument is positive, -1 if negative, and 0 if zero. All hyperparameters are set to default values in this tutorial. In addition, data needs to be restructured for this function. For a clearer understanding of the additional hyperparameters in the `svmfs()` function, see Table 7.

Generally, research shows that SVMs improve as sample sizes increase (Bain et al., 2023). However, some research has shown that sample sizes as small as 80 produce adequate classification models (average RMSEA below 0.01; Figueroa, Zeng-Treitler, Kandula, & Ngo, 2012), though the required size may increase as

Table 7. A table containing the hyperparameters of the `svmfs()` function as well as their corresponding definitions.

Parameter	Description
x	Matrix of the input or predictor variables where the columns are the variables, and the rows are the observations.
y	A numerical vector of class labels, -1, 1.
fs.method	The feature (or variable) selection method. Available methods include 'scad', '1norm' used for LASSO, 'DrHSVM' for Elastic Net, and 'scad+L2'; for Elastic SCAD.
bounds	For an interval grid search a list of values for lambda1 and lambda 2 must be provided to the model.
grid.search	The inner validation method used to obtain the values for lambda1 and lambda2.
inner.val.method	Whether or not the plots of DIRECT algorithm should be shown.
show	Specification of how hyperparameters should be recoded or if no recoding should occur.
parms.coding	By specifying a seed, the results become reproducible. It is included here for the sake of those readers following along.
seed	Matrix of the input or predictor variables where the columns are the variables, and the rows are the observations.

models become more complex (Guo, Graber, McBurney, & Balasubramanian, 2010). We are aware of no sample size recommendations exist for a penalized SVM such as this. The `svmfs()` function can be applied in the following manner.

```
Bounds <- t(data.frame(log2lambda1=c(-10, 10),
    log2lambda2=c(-10,10)))
colnames(bounds)<-c("lower", "upper")
svm.model <- svmfs(x=predTrain, y = svmTrainOutcome,
    fs.method = "scad+L2", bounds=bounds,
    grid.search = "interval", inner.val.method = "cv",
    show = "none", parms.coding = "none",
    seed=123456)
```

The output of the model created using the `svmfs()` function has its own nomenclature that requires explanation. First, rather than referring to the coefficients as coefficients, the model uses the `w` parameter (coming from the term beta weight). The `b` parameter illustrates the intercept of the SVM hyperplane and can be thought of like the `b0` of a regression model. The `xind` parameter tells the user the index (or column location) of the variables selected in the dataset. The full output can be seen in Table 8. Two items selected by this model were MQ16 and S5_66. Variable MQ16 reads, "Once you are aware of the sound(s), because of the sound(s), how often do you have violent thoughts" and S5_66 reads, "Some sounds are so unbearable that I have shouted at people for making them, to make them stop". Both of these items are related to typical responses to triggers by those with misophonia and therefore make theoretical sense.

To examine the accuracy of this model, the same predict function can be used as was implemented previously, but the outputted predictions will require some restructuring, as they come in the form of a factor with underlying numeric values 1 and 2 and they need to have numeric values of 0 and 1. The Elastic SCAD SVM model obtained an accuracy of 0.83 and an F-score of 0.91. The code required to calculate that accuracy and F values are below.

Table 8. A table containing all calculated coefficients of all variables in the Elastic SCAD SVM.

Variable	Coefficient	Variable	Coefficient
(Intercept)	-1.209	S5_38	0.003
MQ3	0.002	S5_39	0.003
MQ5	0.003	S5_40	0.001
MQ8	0.003	S5_41	-0.002
MQ11	0.002	S5_42	0.002
MQ12	0.003	S5_43	0.002
MQ16	0.003	S5_49	0.002
MQ17	0.003	S5_53	-0.003
MQ18	0.002	S5_55	0.003
S5_1	0.001	S5_56	0.007
S5_2	0.005	S5_57	0.005
S5_7	0.006	S5_59	0.003
S5_10	-0.003	S5_65	-0.005
S5_11	-0.002	S5_66	0.001
S5_13	-0.001	S5_68	0.004
S5_24	-0.005	S5_69	0.001
S5_25	0.006	S5_72	0.001
S5_26	0.002	S5_74	0.005
S5_28	0.002	S5_75	0.007
S5_31	0.005	S5_78	-0.008
S5_32	0.005	S5_82	0.005
S5_35	0.005	S5_83	0.006
S5_37	0.002	S5_85	0.003

```
esvm.predictions <- predict(svm.model,
    newdata = svmTestPreds)
esvm.predictions.formatted <-
    as.numeric(esvm.predictions$pred.class)-1
esvm.accuracy <-
    mean(esvm.predictions.formatted == outcomeTest)
esvm.f1 <- F1_Score(esvm.predictions.formatted,
    outcomeTest)
```

Limitations of SVM include the researcher's selection of the kernel function, computation time, and dimension constraints. By default, the `svmfs()` function

utilizes a linear kernel function. Since the kernel is chosen a priori by the researcher, an optimal function must be used for optimal results. SVM models are computationally more expensive than a simpler classification technique (e.g., logistic regression) and will take longer to compute. SVM models face the same degree of freedom problem as LASSO and Elastic Net, which are limited by the number of observations. As such, an ideal dataset for SVM would contain more observations than variables.

2.3 Tree Based Models

Random Forest Another powerful classifier is a decision (or classification) tree (Breiman et al., 2017; Strobl et al., 2009). An example can be seen in Figure 4. From this decision tree, it can be concluded that anyone whose score on variable S5_57 is less than 3 and score on variable S5_60 is less than 3 does not qualify for a misophonia diagnosis. Decision trees are not only powerful classifiers, but they also produce an output that is easy to interpret. However, decision trees are prone to overfitting – so much so that overfitting is almost guaranteed (Bengio, Delalleau, & Simard, 2010). One of the most efficient ways to avoid overfitting is by using multiple trees (i.e., creating a random forest). Random forest creates many decision trees using a randomly selected subset of the data to create each individual tree. The results of all trees are then aggregated to predict the desired outcome. Some major benefits of a random forest classifier are that it can be used with an outcome variable that has any number of levels (Brieuc, Waters, Drinan, & Naish, 2018), meaning that unlike logistic regression, which only works with binary variables, random forest could handle a variable with 3, 4, or even 10 different levels. However, these trees are only used for classification, meaning that they do not perform variable selection. Thus, researchers have had to adapt the classifier to perform variable selection. The utilization of random forest in the *Boruta* package performs well in many different conditions (Kursa & Rudnicki, 2010), and, therefore, is the implementation demonstrated in this tutorial.

The Boruta package contains a series of functions pertaining to variable selection techniques using different measures of importance to select the variables. A measure of importance simply indicates a given variable's value to the model's overall strength. The more useful variables, meaning that they are stronger predictors of the outcome variable, are deemed more important and thus are more likely to be selected than those of lesser importance (i.e., less predictive power). Note that in this paper, mean decreased accuracy is the metric used to calculate variable importance. The Boruta package also has its own sample size suggestions. The original paper implementing the the package states that for typical problems, samples of 5-200 are often sufficient, assuming the number of true predictors is not extremely small compared to the total (Kursa & Rudnicki, 2010). It notes that as problems get more complex, the sample size should increase.

A simple regression formula statement is used to run the model: `outcome predictors`. Because all predictors will be used, a shortcut can be implemented using a period (`.`) in place of predictors as seen in the code below. If not, all variables were to be included in the model, the user would need to type all the

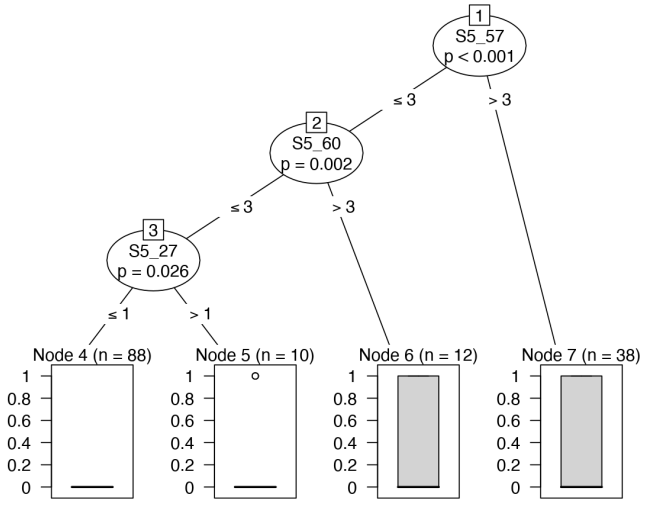


Figure 4. An example of a decision tree built on the misophonia data using the `ctree()` function.

relevant predictors names in the formula statement concatenated with addition symbols (+). Knowing this, the model can then be built using the following code:

```
set.seed(123456)
boruta.model <- Boruta(as.factor(MQDX) ~. ,
    data=trainDat)
```

The `Boruta()` function classifies variables as either important, unimportant, or of tentative importance. Regarding the misophonia dataset, 15 were deemed important, 74 were deemed unimportant, and the remaining 17 were placed in the tentative category. For a list of all variables that were classified in each category and a visualization of the `boruta.model` output, see Table 9. Figure 5 illustrates the variability of the importance score calculated for each variable during the Boruta process and their ultimate classification. A model can be built using either a) all variables that were not deemed unimportant (non-rejected variables) or b) only the confirmed important variables. For the purpose of this tutorial, only variables that have been confirmed important are included in the model. Two items that were confirmed important are MQ16 and S5_59. Variable MQ16 was justified in the SVM section as it pertains to having violent thoughts. S5_59 reads "If I cannot avoid certain sounds I feel helpless." Helplessness is often associated with anxiety (i.e., learned helplessness) which is often co-diagnosed with misophonia, and as such, this variable is theoretically justified.

This model is then built using the `randomForest()` function since Boruta implements a random forest model internally. The model is built in the following way.

Table 9. A table containing the classifications of importance for each variable as determined by the `Boruta()` function. Note that the implementation of Boruta used in this paper utilized mean decreased accuracy as the metric to calculate variable importance.

Variables	Items
Confirmed Important	MQ12, MQ13, MQ16, S5_3, S5_34, S5_35, S5_39, S5_40, S5_53, S5_56, S5_57, S5_59, S5_60, S5_67, S5_75
Rejected	MQ1, MQ2, MQ3, MQ4, MQ5, MQ6, MQ7, MQ8, MQ10, MQ11, MQ14, MQ15, MQ18, MQ20, S5_1, S5_4, S5_6, S5_7, S5_8, S5_9, S5_10, S5_11, S5_12, S5_13, S5_14, S5_15, S5_16, S5_17, S5_19, S5_20, S5_23, S5_24, S5_26, S5_28, S5_29, S5_30, S5_32, S5_33, S5_36, S5_37, S5_41, S5_42, S5_43, S5_44, S5_45, S5_46, S5_47, S5_48, S5_49, S5_50, S5_51, S5_52, S5_54, S5_55, S5_58, S5_64, S5_65, S5_66, S5_68, S5_70, S5_71, S5_72, S5_73, S5_74, S5_76, S5_77, S5_78, S5_79, S5_80, S5_82, S5_83, S5_84, S5_86, S5_87
Tentative	MQ17, MQ19, S5_2, S5_5, S5_18, S5_21, S5_22, S5_25, S5_27, S5_31, S5_38, S5_61, S5_62, S5_63, S5_69, S5_81, S5_85

```
set.seed(123456)
finalBoruta <- getConfirmedFormula(boruta.model)
selectedModel <- randomForest(finalBoruta,
    data=trainDat)
```

The predictive accuracy of the random forest model can be calculated using the `predict()` function, just as it has been for other models. An accuracy of .88 was obtained for this model and an F-score of 0.93. The algorithm may not perform well with highly unbalanced classifications or in situations where a given level contains a very small number of classifications.

2.4 Comparing All Models

For a comparison of the accuracy values and F-scores obtained by all techniques implemented in this tutorial, see Table 10. From this, we can state that the GA produced the most accurate model. However, there was no difference in the accuracy of the LASSO non-biased (e.g., the standard regression model built using variable selected via the LASSO), Boruta, and Elastic Net models. Depending on the purpose of your model, you may want to use a performance metric other than accuracy. Within the context of our motivating example, it may be worth examining the following:

- **Sensitivity**: Given the individual truly has misophonia, how likely is the classifier to realize that?
- **Specificity**: Given the individual truly does not have misophonia, how likely is the classifier to realize that?

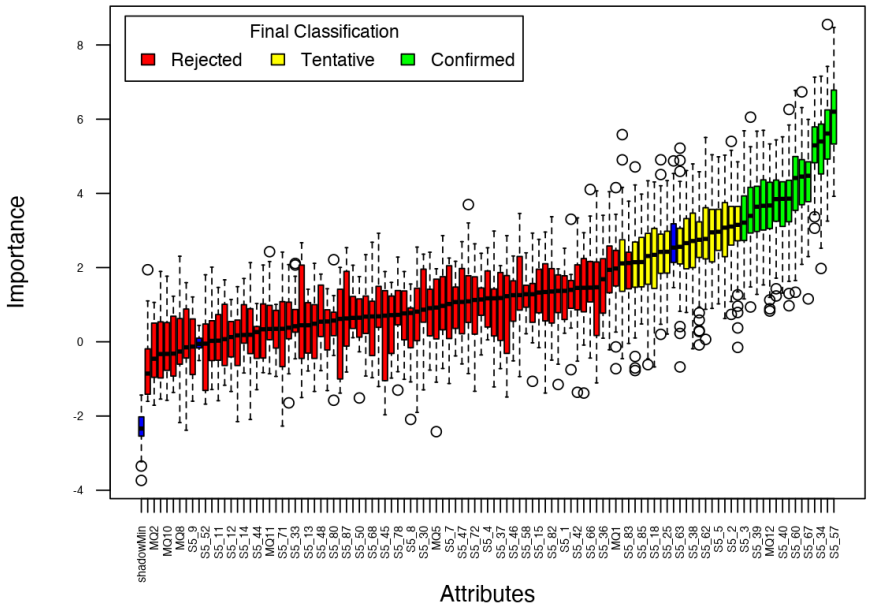


Figure 5. A plot containing the Z-score transformed estimates of variable importance scores for each variable in the `Boruta()` model. Blue boxplots correspond to minimal, average, and maximum Z-scores of a shadow attribute. Red and green boxplots represent Z-scores of rejected and confirmed attributes respectively. Note that the implementation of Boruta used in this paper utilized mean decreased accuracy as the metric to calculate variable importance.

- **Positive predictive value**: Given the classifier claims the individual to have misophonia, how likely is it that the individual really has misophonia?
- **Negative predictive value**: Given the classifier claims the individual does not have misophonia, how likely is it that the individual really does not have the disease?

While accuracy serves as a useful general indicator of model performance, it can be misleading, particularly when dealing with unbalanced datasets where one class is significantly more prevalent than the other. In such cases, a model can achieve high accuracy by simply predicting the majority class, even if it performs poorly on the minority class. Therefore, it's essential to consider alternative performance metrics that provide a more nuanced understanding of a model's strengths and weaknesses. For instance, sensitivity (the true positive rate) measures the proportion of actual positives that are correctly identified, while specificity (the true negative rate) quantifies the proportion of actual neg-

Table 10. A table containing the predictive accuracy values obtained by all models built in this tutorial paper. Methods are listed such that the accuracy values are ordered from least accurate to most accurate. Significance is determined relative to the previous model (i.e., Elastic SCAD SVM was determined to have a statistically significant better accuracy than Elastic Net non-biased) according to a McNemar's Chi-squared test with continuity correction. Note significant differences were not evaluated for F-scores.

Method	**Cross-validated Accuracy**	**Cross-validated F-Score**
Elastic Net non-biased	0.797	0.881
Elastic SCAD SVM	0.828**	0.905
LASSO	0.859**	0.918
Elastic Net	0.875	0.938
Boruta	0.875	0.930
LASSO non-biased	0.891	0.916
GA	1***	1

Note: * $p < .05$, ** $p < .01$, *** $p < .0001$

atives that are correctly classified. These metrics are crucial when the cost of misclassification differs for each class, such as in medical diagnosis where failing to identify a true case (low sensitivity) can have more severe consequences than a false positive (low specificity). Precision reflects the proportion of predicted positives that are actually positive, while recall is synonymous with sensitivity. Another valuable metric is the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve, which comprehensively measures a model's ability to discriminate between classes across various thresholds. Researchers should carefully consider their research question's specific goals and context to select the most appropriate performance metrics, ensuring a balanced and insightful evaluation of their models.

Given that our example pertains to diagnosis, it is possible that one may favor sensitivity over specificity in that we want to minimize the number of missed cases. However, it is also possible that we would want to minimize the number of false diagnoses to save individuals the cost of unnecessary intervention. A confusion matrix (discussed briefly in Appendix B) might be useful. Alternatively, one could use the AUC of the ROC curve. One should carefully consider these factors when deciding on the performance metric by which to evaluate a model.

Examining the selected variables reveals interesting method-dependent patterns. Elastic SCAD SVM selected many more variables than LASSO, but had a worse accuracy. Given this outcome, it may not be ideal to use all variables selected by Elastic SCAD SVM in this dataset. There was only one variable (S5_57) that was selected by all five methods. So, there is a clear method effect on the variables that are deemed to be important. Within the context of our example, we could interpret this to mean that the question, "Sometimes in response to sounds, I feel rage that is difficult to control," is an incredibly important predictor for misophonia and may capture a defining characteristic of the disorder. Beyond improving predictive accuracy, understanding why certain

variables are deemed important can provide valuable insights into the underlying mechanisms or factors driving the outcome of interest. This insight could guide future research exploring the role of emotional regulation in misophonia and potentially inform the development of targeted interventions. Furthermore, the identification of unexpected or previously overlooked variables as important predictors can spark new research questions and hypotheses. This iterative process of variable selection, model building, and hypothesis generation can lead to a more nuanced and comprehensive understanding of complex phenomena. By carefully examining the selected variables, researchers can generate hypotheses, refine theoretical models, and ultimately gain a deeper understanding of complex human behavior and health outcomes.

3 Discussion

This tutorial provided an overview and a practical guide for the implementation of LASSO (Friedman et al., 2010), Elastic Net (Friedman et al., 2010), a genetic algorithm (Scrucca, 2013, 2017), Elastic SCAD SVM (Becker et al., 2009), and random forest via Boruta (Kursa & Rudnicki, 2010) in R v. 4.2.1. Proper analysis of the output as well as comparisons on the predictive accuracy of each method are also discussed. More information on R, other useful machine learning software, and some of these functions were provided in the Appendices. Lastly, an OSF project containing all code implemented in this tutorial, additional code the reader may find useful, and the data used is available. For a full link to the project, see the availability of data and materials section of this paper.

Variable selection allows researchers to find parsimonious models that are also good predictive or classifying models. Given R's increasing popularity among researchers due to the software's free and open access nature, it is valuable to the field to provide more guidance on the variable selection methods available in R. In addition, the extent to which some of these methods overfit data should not be ignored when implementing them on real-world data. Suppose a researcher is concerned with creating a generalizable model. In that case, it is recommended that the results be validated not only through some form of cross-validation but also through the collection of a new sample. Through this tutorial, we aim to push the field towards more transparent guidelines and standardization for the use of variable selection techniques and machine learning in psychological research.

While variable selection offers numerous advantages, it's crucial to acknowledge its potential ethical implications, particularly in sensitive applications like clinical diagnosis or risk assessment (Obermeyer, Powers, Vogeli, & Mullainathan, 2019). If biased or incomplete data is used for training, variable selection algorithms can perpetuate and even amplify existing societal biases, leading to unfair or discriminatory outcomes (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021). For example, if a dataset used to predict criminal recidivism is skewed towards certain demographics, the selected variables might unfairly target individuals from those groups, even if the variables are not causally related to re-

cidivism. Similarly, in clinical diagnosis, relying on variables that are correlated with social determinants of health rather than underlying biological mechanisms could result in misdiagnosis or inadequate treatment for marginalized populations (Vyas, Eisenstein, & Jones, 2020). Therefore, researchers must carefully consider the potential for bias in their data and strive to develop and implement variable selection techniques that prioritize fairness and equity.

Beyond enhancing model performance, variable selection holds significant potential for translational impact in the social and health sciences. By identifying the most influential predictors, researchers can better understand the underlying mechanisms driving complex phenomena, leading to more effective interventions, treatments, and public health strategies. For instance, in personalized medicine, variable selection can help tailor treatments to individual patients based on their unique genetic, environmental, and lifestyle factors . Identifying key risk factors for chronic diseases through variable selection in public health can inform targeted prevention programs and resource allocation strategies. Moreover, in developing psychological interventions, variable selection can aid in identifying the most effective treatment components and tailoring therapies to specific patient needs and characteristics (Vyas et al., 2020). By focusing research and interventions on the most impactful variables, variable selection can contribute more effective and efficient solutions to pressing social and health challenges.

There are many ways a researcher can define accuracy. When interested in classification, an optimal model is one with minimal classification error, as we have highlighted throughout this tutorial (Huang, 2015). However, previous research notes that if classification is not the goal, minimal error can be conceptualized as selecting variables with the highest relevance to the given outcome (Peng, Long, & Ding, 2005). With this in mind, it is important that variables are not falsely discovered (i.e., a variable that is not relevant is selected; Type I error in selection). An interested reader is pointed to the *knockoff* package (Candés, Fan, Janson, & Lv, 2018) and work by Zimmermann, Baillie, Kormaksson, Ohlssen, and Sechidis (2024). Another important aspect of variable selection, especially for the applied researcher, is the stability of a model (i.e., how robust a particular model is to small changes in the data). We discussed one way to address this concern through the concept of cross-validation, however, there are additional ways one might go about addressing this concern (Bommert & Lang, 2021; Nogueira, Sechidis, & Brown, 2018). The field would benefit from additional tutorial papers discussing the balance of these issues with accuracy to help guide the applied researcher.

Many additional R packages will perform variable selection using random forest as well as SVMs, but only one of each was demonstrated in this tutorial. The demonstrated methods in the current tutorial were selected because they are commonly used in the psychological sciences, are powerful techniques for classification (e.g., diagnosing individuals with misophonia) and variable selection, and are all freely available to researchers in R. In a similar vein, we have included only five machine learning methods here but many more exist, and additional tutorials should be provided to applied researchers about how best to implement

them following research demonstrating each algorithm's performance to indicate which algorithm is best for addressing certain research questions. For the interested reader, a comparison of the performance of each method demonstrated in this tutorial can be found in Bain et al. (2023).

4 Conclusion

This tutorial presented an overview and a practical guide for implementing five variable selection techniques: LASSO (Friedman et al., 2010), Elastic Net (Friedman et al., 2010), a genetic algorithm (Scrucca, 2013, 2017), Elastic SCAD SVM (Becker et al., 2009), and random forest via Boruta (Kursa & Rudnicki, 2010) in R. Proper analysis of the output as well as comparisons on the predictive accuracy of each method are also discussed. More information on R, other useful machine learning software, and some of these functions were provided in the Appendices. Lastly, an OSF project containing all code implemented in this tutorial, additional code the reader may find useful, and the data used is available. For a full link to the project, see the availability of data and materials section of this paper.

This paper highlighted the increasing availability of large and complex datasets in the social and health sciences, requiring a move beyond traditional variable selection techniques like stepwise regression. This tutorial demonstrates that modern machine learning methods offer powerful and accessible alternatives for identifying the most informative variables, improving model accuracy, and gaining a deeper understanding of complex phenomena. By embracing these advancements and continuing to explore the ethical and interpretive dimensions of variable selection, researchers can enhance the rigor, reproducibility, and, ultimately, the translational impact of their work. We encourage readers to consult the documentation for each method for further examples and details. The user is to refer to each method's full documentation for additional examples and details. We hope that this tutorial makes these methods more easily accessible to the everyday psychological researcher, opens doors to applications of variable selection in new areas, and leads to a decreased presence of less powerful methods (e.g., stepwise selection) in the literature.

Availability of Data and Materials

The accompanying code and data utilized in this tutorial can be found here: https://osf.io/pr6j8/?view_only=c778e322f1d54429990067580e615afb. Additional supplementary information such as a glossary of key terms, R package recommendations, etc. are also available through OSF.

Authors' Contributions

CMB conducted all analyses and drafted the manuscript; DS contributed to manuscript draft and recreation from its original form, and supervised manuscript

preparation. YMB contributed to manuscript draft and recreation from its old form. Regarding the data utilized here, LEE conceived of the study design of the project and supervised all aspects of funding, participant recruitment, and data collection while JEN aided in the original study design, led all data collection and data preprocessing; JEL supervised data analysis and manuscript preparation. All authors contributed significantly to manuscript preparation.

References

Adams, L. J., Bello, G., & Dumancas, G. G. (2015). Development and Application of a Genetic Algorithm for Variable Optimization and Predictive Modeling of Five-Year Mortality Using Questionnaire Data. *Bioinformatics and Biology Insights*, *9s3*, BBI.S29469. doi: https://doi.org/10.4137/BBI.S29469

Alamri, L. H., Almuslim, R. S., Alotibi, M. S., Alkadi, D. K., Ullah Khan, I., & Aslam, N. (2021). Predicting Student Academic Performance using Support Vector Machine and Random Forest. In *Proceedings of the 2020 3rd International Conference on Education Technology Management* (pp. 100–107). New York, NY, USA: Association for Computing Machinery. doi: https://doi.org/10.1145/3446590.3446607

Algamal, Z. Y., & Lee, M. H. (2015). Applying Penalized Binary Logistic Regression with Correlation Based Elastic Net for Variables Selection. *Journal of Modern Applied Statistical Methods*, *14*(1), 168–179. doi: https://doi.org/10.22237/jmasm/1430453640

Amene, E., Hanson, L. A., Zahn, E. A., Wild, S. R., & Döpfer, D. (2016). Variable selection and regression analysis for the prediction of mortality rates associated with foodborne diseases. *Epidemiology and Infection*, *144*(9), 1959–1973.

Aragón-Royón, F., Jiménez-Vílchez, A., Arauzo-Azofra, A., & Benítez, J. M. (2020). *FSinR: an exhaustive package for feature selection.* arXiv.

Arjomandi-Nezhad, A., Guo, Y., Pal, B. C., & Varagnolo, D. (2023). *A Model Predictive Approach for Enhancing Transient Stability of Grid-Forming Converters.* arXiv.

Bain, C., Shi, D., Boness, C. L., & Loeffelman, J. (2023). *A Simulation Study Comparing the Use of Supervised Machine Learning Variable Selection Methods in the Psychological Sciences.* PsyArXiv. doi: https://doi.org/10.31234/osf.io/y53t6

Barceló, P., Monet, M., Pérez, J., & Subercaseaux, B. (2020). Model interpretability through the lens of computational complexity. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 15487–15498). Red Hook, NY, USA: Curran Associates Inc.

Basarkod, G., Sahdra, B., & Ciarrochi, J. (2018). *Body Image-Acceptance and Action Questionnaire-5: An Abbreviation Using Genetic Algorithms* (Tech. Rep.).

Battineni, G., Chintalapudi, N., & Amenta, F. (2019). Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM). *Informatics in Medicine Unlocked*, *16*, 100200. doi: https://doi.org/10.1016/j.imu.2019.100200

Becker, N., Toedt, G., Lichter, P., & Benner, A. (2011). Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data. *BMC Bioinformatics*, *12*(1), 138. doi: https://doi.org/10.1186/1471-2105-12-138

Becker, N., Werft, W., Toedt, G., Lichter, P., & Benner, A. (2009). penalizedSVM: a R-package for feature selection SVM classification. *Bioinformatics*, *25*(13), 1711–1712. doi: https://doi.org/10.1093/bioinformatics/btp286

Bengio, Y., Delalleau, O., & Simard, C. (2010). Decision Trees Do Not Generalize to New Variations. *Computational Intelligence*, *26*(4), 449–467. doi: https://doi.org/10.1111/j.1467-8640.2010.00366.x

Bierman, S., & Steel, S. (2009). Variable Selection for Support Vector Machines. *Communications in Statistics - Simulation and Computation*, *38*(8), 1640–1658. doi: https://doi.org/10.1080/03610910903072391

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, *97*(1), 245–271. doi: https://doi.org/10.1016/S0004-3702(97)00063-5

Bommert, A., & Lang, M. (2021). stabm: Stability Measures for Feature Selection. *Journal of Open Source Software*, *6*(59), 3010. doi: https://doi.org/10.21105/joss.03010

Bourdès, V., Bonnevay, S., Lisboa, P., Defrance, R., Pérol, D., Chabaud, S., ... Négrier, S. (2010). Comparison of Artificial Neural Network with Logistic Regression as Classification Models for Variable Selection for Prediction of Breast Cancer Patient Outcomes. *Advances in Artificial Neural Systems*, *2010*, 1–11. doi: https://doi.org/10.1155/2010/309841

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. New York: Chapman and Hall/CRC. doi: https://doi.org/10.1201/9781315139470

Brieuc, M. S. O., Waters, C. D., Drinan, D. P., & Naish, K. A. (2018). A practical introduction to Random Forest for genetic association studies in ecology and evolution. *Molecular Ecology Resources*, *18*(4), 755–766. doi: https://doi.org/10.1111/1755-0998.12773

Cafri, G., Li, L., Paxton, E. W., & Fan, J. (2018). Predicting risk for adverse health events using random forest. *Journal of Applied Statistics*, *45*(12), 2279–2294. doi: https://doi.org/10.1080/02664763.2017.1414166

Calcagno, V., & Mazancourt, C. D. (2010). **glmulti** : An *R* Package for Easy Automated Model Selection with (Generalized) Linear Models. *Journal of Statistical Software*, *34*(12). doi: https://doi.org/10.18637/jss.v034.i12

Candés, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *80*(3).

Cateni, S., Colla, V., & Vannucci, M. (2010). Variable Selection through Genetic Algorithms for Classification Purposes. In *Artificial Intelligence and Applications.* Innsbruck, Austria: ACTAPRESS. doi: https://doi.org/10.2316/P.2010.674-080

Chen, Y., & Yang, Y. (2021). The One Standard Error Rule for Model Selection: Does It Work? *Stats*, *4*(4), 868–892. doi: https://doi.org/10.3390/stats4040051

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), e000262. doi: https://doi.org/10.1136/fmch-2019-000262

Chu, M., Fang, Z., Mao, L., Ma, H., Lee, C.-Y., & Chiang, Y.-C. (2024). Creating A child-friendly social environment for fewer conduct problems and more prosocial behaviors among children: A LASSO regression approach. *Acta Psychologica*, *244*, 104200. doi: https://doi.org/10.1016/j.actpsy.2024.104200

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282. doi: https://doi.org/10.1111/j.2044-8317.1992.tb00992.x

El Haouij, N., Poggi, J.-M., Ghozi, R., Sevestre-Ghalila, S., Jaïdane, M., Poggi Jean-Michel, J.-M., & El Haouij, N. (2018). Random forest-based approach for physiological functional variable selection for driver's stress level classification. *Statistical Methods & Applications*. doi: https://doi.org/10.1007/s10260-018-0423-5

Engebretsen, S., & Bohlin, J. (2019). Statistical predictions with glmnet. *Clinical Epigenetics*, *11*(1), 123. doi: https://doi.org/10.1186/s13148-019-0730-1

Fernandez, M., Caballero, J., Fernandez, L., & Sarai, A. (2011). Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM). *Molecular Diversity*, *15*(1), 269–289. doi: https://doi.org/10.1007/s11030-010-9234-9

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 8. doi: https://doi.org/10.1186/1472-6947-12-8

Foucart, S., Tadmor, E., & Zhong, M. (2023). On the Sparsity of LASSO Minimizers in Sparse Data Recovery. *Constructive Approximation*, *57*(2), 901–919. doi: https://doi.org/10.1007/s00365-022-09594-1

Fox, E. W., Hill, R. A., Leibowitz, S. G., Olsen, A. R., Thornbrugh, D. J., & Weber, M. H. (2017). Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. *Environmental Monitoring and Assessment*, *189*(7), 316. doi: https://doi.org/10.1007/s10661-017-6025-0

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for

Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, *33*(1), 1–22.

Frohlich, H., Chapelle, O., & Scholkopf, B. (2003). Feature selection for support vector machines by means of genetic algorithm. In *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence* (pp. 142–148). Sacramento, CA, USA: IEEE Comput. Soc. doi: https://doi.org/10.1109/TAI.2003.1250182

Gan, C. C., & Learmonth, G. (2016). *Developing an ICU scoring system with interaction terms using a genetic algorithm.* arXiv. doi: https://doi.org/10.48550/arXiv.1604.06730

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. doi: https://doi.org/10.1016/j.patrec.2010.03.014

Ghojogh, B., & Crowley, M. (2023). *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial.* arXiv. doi: https://doi.org/10.48550/arXiv.1905.12787

Gunn, H. J., Hayati Rezvan, P., Fernández, M. I., & Comulada, W. S. (2023). How to apply variable selection machine learning algorithms with multiply imputed data: A missing discussion. *Psychological Methods*, *28*(2), 452–471. doi: https://doi.org/10.1037/met0000478

Guo, Y., Graber, A., McBurney, R. N., & Balasubramanian, R. (2010). Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. *BMC Bioinformatics*, *11*(1), 447. doi: https://doi.org/10.1186/1471-2105-11-447

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(7-8), 1157–1182. doi: https://doi.org/10.1162/153244303322753616

Han, H., & Dawson, K. J. (2021). Applying elastic-net regression to identify the best models predicting changes in civic purpose during the emerging adulthood. *Journal of Adolescence*, *93*, 20–27. doi: https://doi.org/10.1016/j.adolescence.2021.09.011

Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection – A review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449. doi: https://doi.org/10.1002/bimj.201700067

Helwig, N. E. (2017). Adding bias to reduce variance in psychological results: A tutorial on penalized regression. *The Quantitative Methods for Psychology*, *13*(1), 1–19. doi: https://doi.org/10.20982/tqmp.13.1.p001

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67. doi: https://doi.org/10.2307/1267351

Huang, S. H. (2015). Supervised feature selection: A tutorial. *Artificial Intelligence Research*, *4*(2), p22. doi: https://doi.org/10.5430/air.v4n2p22

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. doi: https://doi.org/10.1007/BF01908075

Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pa-

supuleti, S., ... Jo, O. (2020). COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Frontiers in Public Health*, *8*. doi: https://doi.org/10.3389/fpubh.2020.00357

Jia, W., Sun, M., Lian, J., & Hou, S. (2022). Feature dimensionality reduction: a review. *Complex & Intelligent Systems*, *8*(3), 2663–2693. doi: https://doi.org/10.1007/s40747-021-00637-x

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software*, *15*(9). doi: https://doi.org/10.18637/jss.v015.i09

Kerkhoff, D., & Nussbeck, F. W. (2019). The Influence of Sample Size on Parameter Estimates in Three-Level Random-Effects Models. *Frontiers in Psychology*, *10*. doi: https://doi.org/10.3389/fpsyg.2019.01067

Kirpich, A., Ainsworth, E. A., Wedow, J. M., Newman, J. R. B., Michailidis, G., & McIntyre, L. M. (2018). Variable selection in omics data: A practical evaluation of small sample sizes. *PLOS ONE*, *13*(6), e0197910. doi: https://doi.org/10.1371/journal.pone.0197910

Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 202–207). Portland, Oregon: AAAI Press.

Kok, B. C., Choi, J. S., Oh, H., & Choi, J. Y. (2021). Sparse Extended Redundancy Analysis: Variable Selection via the Exclusive LASSO. *Multivariate Behavioral Research*, *56*(3), 426–446. doi: https://doi.org/10.1080/00273171.2019.1694477

Kuchirko, Y., Bennet, A., Halim, M. L., Costanzo, P., & Ruble, D. (2021). The Influence of Siblings on Ethnically Diverse Children's Gender Typing Across Early Development. *Developmental Psychology*. doi: https://doi.org/10.1037/dev0001173.supp

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the **Boruta** Package. *Journal of Statistical Software*, *36*(11). doi: https://doi.org/10.18637/jss.v036.i11

Leardi, R. (2000). Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, *14*(5-6), 643–655. doi: https://doi.org/10.1002/1099-128X(200009/12)14:5/6¡643::AID-CEM621¿3.0.CO;2-E

Lenters, V., Vermeulen, R., & Portengen, L. (2018). Performance of variable selection methods for assessing the health effects of correlated exposures in case–control studies. *Occupational and Environmental Medicine*, *75*(7), 522–529. doi: https://doi.org/10.1136/oemed-2016-104231

Liu, X., Cao, P., Gonçalves, A. R., Zhao, D., & Banerjee, A. (2018). Modeling Alzheimer's Disease Progression with Fused Laplacian Sparse Group Lasso. *ACM Transactions on Knowledge Discovery from Data*, *12*(6), 65:1–65:35. doi: https://doi.org/10.1145/3230668

Loef, B., Wong, A., Janssen, N. A. H., Strak, M., Hoekstra, J., Picavet, H. S. J., ... Herber, G.-C. M. (2022). Using random forest to identify longitudinal

predictors of health in a 30-year cohort study. *Scientific Reports*, *12*(1), 10372. doi: https://doi.org/10.1038/s41598-022-14632-w

Loughrey, J., & Cunningham, P. (2005). Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets. In M. Bramer, F. Coenen, & T. Allen (Eds.), *Research and Development in Intelligent Systems XXI* (pp. 33–43). London: Springer London. doi: https://doi.org/10.1007/1-84628-102-4_3

Luo, J., Ren, S., Li, Y., & Liu, T. (2021). The Effect of College Students' Adaptability on Nomophobia: Based on Lasso Regression. *Frontiers in Psychiatry*, *12*. doi: https://doi.org/10.3389/fpsyt.2021.641417

Marafino, B. J., John Boscardin, W., & Adams Dudley, R. (2015). Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of Biomedical Informatics*, *54*, 114–120. doi: https://doi.org/10.1016/j.jbi.2015.02.003

McDonald, G. C. (2009). Ridge regression. *WIREs Computational Statistics*, *1*(1), 93–100. doi: https://doi.org/10.1002/wics.14

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, *54*(6), 115:1–115:35. doi: https://doi.org/10.1145/3457607

Mendez-Civieta, A., Aguilera-Morillo, M. C., & Lillo, R. E. (2021). Adaptive sparse group LASSO in quantile regression. *Advances in Data Analysis and Classification*, *15*(3), 547–573. doi: https://doi.org/10.1007/s11634-020-00413-8

ML | Underfitting and Overfitting. (2017, November). Retrieved 2024-07-10, from `https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/` (Section: Machine Learning)

Moore, J. H., Andrews, P. C., Olson, R. S., Carlson, S. E., Larock, C. R., Bulhoes, M. J., ... Armentrout, S. L. (2017). Grid-based stochastic search for hierarchical gene-gene interactions in population-based genetic studies of common human diseases. *BioData Mining*, *10*(1), 19. doi: https://doi.org/10.1186/s13040-017-0139-3

Nogueira, S., Sechidis, K., & Brown, G. (2018). On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research*, *18*(174), 1–54.

Norris, J. E., Kimball, S. H., Nemri, D. C., & Ethridge, L. E. (2022). Toward a Multidimensional Understanding of Misophonia Using Cluster-Based Phenotyping. *Frontiers in Neuroscience*, *16*. doi: https://doi.org/https://doi.org/10.3389/fnins.2022.832516

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. doi: https://doi.org/10.1126/science.aax2342

Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clin-*

ical Epidemiology, *48*(12), 1503–1510. doi: https://doi.org/10.1016/0895-4356(95)00048-8

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379. doi: https://doi.org/10.1016/s0895-4356(96)00236-3

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238. doi: https://doi.org/10.1109/TPAMI.2005.159

Pratik, S., Nayak, D., Prasath, R., & Swarnkar, T. (2022). Prediction of Smoking Addiction Among Youths Using Elastic Net and KNN: A Machine Learning Approach. In (pp. 199–209). doi: https://doi.org/10.1007/978-3-031-21517-9_20

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, *53*(4), 1–37. doi: https://doi.org/10.18637/jss.v053.i04

Scrucca, L. (2017). On some extensions to GA package: Hybrid optimisation, parallelisation and islands evolution. *R Journal*, *9*(1), 187–206. doi: https://doi.org/10.32614/rj-2017-008

Scrucca, L., Fop, M., Murphy, T., Brendan, & Raftery, A., E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, *8*(1), 289. doi: https://doi.org/10.32614/RJ-2016-021

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory Mediation Analysis via Regularization. *Structural equation modeling : a multidisciplinary journal*, *24*(5), 733–744. doi: https://doi.org/10.1080/10705511.2017.1311775

Shi, D., Shi, D., & Fairchild, A. J. (2023). Variable Selection for Mediators under a Bayesian Mediation Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–14. doi: https://doi.org/10.1080/10705511.2022.2164285

Singla, M., & Shukla, K. K. (2020). Robust statistics-based support vector machine and its variants: a survey. *Neural Computing and Applications*, *32*(15), 11173–11194. doi: https://doi.org/10.1007/s00521-019-04627-6

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, *5*(1), 32. doi: https://doi.org/10.1186/s40537-018-0143-6

Song, Q. C., Tang, C., & Wee, S. (2021). Making Sense of Model Generalizability: A Tutorial on Cross-Validation in R and Shiny. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920947067. doi: https://doi.org/10.1177/2515245920947067

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. doi: https://doi.org/10.1037/a0016973

Tan, L., Main, J. B., & Darolia, R. (2021). Using random forest analysis to identify student demographic and high school-level factors that predict college engineering major choice. *Journal of Engineering Education*, *110*(3), 572–593. doi: https://doi.org/10.1002/jee.20393

Tay, J. K., Narasimhan, B., & Hastie, T. (2023). Elastic Net Regularization Paths for All Generalized Linear Models. *Journal of Statistical Software*, *106*(1). doi: https://doi.org/10.18637/jss.v106.i01

Tharwat, A., & Hassanien, A. E. (2019). Quantum-Behaved Particle Swarm Optimization for Parameter Optimization of Support Vector Machine. *Journal of Classification*, *36*, 576–598. doi: https://doi.org/10.1007/s00357-018-9299-1

Thompson, B. (1995). Stepwise Regression and Stepwise Discriminant Analysis Need Not Apply here: A Guidelines Editorial. *Educational and Psychological Measurement*, *55*(4), 525–534. doi: https://doi.org/10.1177/0013164495055004001

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Trevino, V., & Falciani, F. (2006). GALGO: an R package for multivariate variable selection using genetic algorithms. *Bioinformatics*, *22*(9), 1154–1156. doi: https://doi.org/10.1093/bioinformatics/btl074

van Vuuren, C. L., van Mens, K., de Beurs, D., Lokkerbol, J., van der Wal, M. F., Cuijpers, P., & Chinapaw, M. J. M. (2021). Comparing machine learning to a rule-based approach for predicting suicidal behavior among adolescents: Results from a longitudinal population-based survey. *Journal of Affective Disorders*, *295*, 1415–1420. doi: https://doi.org/10.1016/j.jad.2021.09.018

Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *New England Journal of Medicine*, *383*(9), 874–882. doi: https://doi.org/10.1056/NEJMms2004740

Wang, L., Cheng, H., Liu, Z., & Zhu, C. (2014). A robust elastic net approach for feature learning. *Journal of Visual Communication and Image Representation*, *25*(2), 313–321. doi: https://doi.org/10.1016/j.jvcir.2013.11.002

Wehrens, R., & Franceschi, P. (2012). Meta-statistics for variable selection: The R package BioMark. *Journal of Statistical Software*, *51*(10). doi: https://doi.org/10.18637/jss.v051.i10

Wen, Q., Mustafi, S. M., Li, J., Risacher, S. L., Tallman, E., Brown, S. A., ... Wu, Y.-C. (2019). White matter alterations in early-stage Alzheimer's disease: A tract-specific study. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *11*, 576–587. doi: https://doi.org/10.1016/j.dadm.2019.06.003

Wettstein, A., Jenni, G., Schneider, I., Kühne, F., grosse Holtforth, M., & La Marca, R. (2023). Predictors of Psychological Strain and Allostatic Load in Teachers: Examining the Long-Term Effects of Biopsychosocial Risk and Protective Factors Using a LASSO Regression Approach. *In-*

ternational Journal of Environmental Research and Public Health, *20*(10), 5760. doi: https://doi.org/10.3390/ijerph20105760

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189. doi: https://doi.org/10.1111/j.1365-2656.2006.01141.x

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Software, P., & PBC. (2023). *dplyr: A Grammar of Data Manipulation.*

Wiegand, R. E. (2010). Performance of using multiple stepwise algorithms for variable selection. *Statistics in Medicine*, *29*(15), 1647–1659. doi: https://doi.org/10.1002/sim.3943

Wu, M. S., Lewin, A. B., Murphy, T. K., & Storch, E. A. (2014). Misophonia: Incidence, Phenomenology, and Clinical Correlates in an Undergraduate Student Sample: Misophonia. *Journal of Clinical Psychology*, *70*(10), 994–1007. doi: https://doi.org/10.1002/jclp.22098

Xu, H., Caramanis, C., & Mannor, S. (2009). Robustness and Regularization of Support Vector Machines. *Journal of Machine Learning Research 1*, *10*, 1485–1510.

Yan, Y. (2024). *MLmetrics: Machine Learning Evaluation Metrics.*

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: https://doi.org/10.1177/1745691617693393

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program. *European Journal of Open, Distance and E-Learning*, *17*(1), 118–133. doi: https://doi.org/10.2478/eurodl-2014-0008

Zimmermann, M. R., Baillie, M., Kormaksson, M., Ohlssen, D., & Sechidis, K. (2024). All that Glitters Is not Gold: Type-I Error Controlled Variable Selection from Clinical Trial Data. *Clinical Pharmacology & Therapeutics*, *115*(4), 774–785. doi: https://doi.org/10.1002/cpt.3211

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix A

Table 1. Demographic information for the sample used in the illustrative example.

Variable	n (%)
Age (Years)	M = 18.96, SD = 1.7
Gender	
Male	104 (30.3%)
Female	239 (69.7%)
Ethnicity	
White	263 (76.7%)
Black/African American	32 (9.3%)
Latino/Hispanic	46 (13.4%)
Asian/Asian American	28 (8.2%)
American Indian/Alaska Native	26 (7.6%)
Native Hawaiian/Other Pacific Islander	2 (0.6%)
Other	2 (0.6%)
Education	
Less than high school	2 (0.6%)
High school graduate	129 (37.6%)
Some years of college/university (no degree)	194 (56.6%)
Vocational training	2 (0.6%)
Associates degree	8 (2.3%)
Bachelor's degree	5 (1.5%)
Master's degree	1 (0.3%)

Appendix B

The random forest output contains different information than any other technique discussed in this paper because it performs a type of cross-validation internally through looking at something called Out of Bag error (OOB; sometimes referred to as the out-of-bag estimate). The OOB is an approach to measuring the prediction error of a random forest model or of other decision tree models. OOB error is the mean prediction error of a given sample, using only the trees which did not have that sample in their bootstrapped sample. This sounds potentially confusing, but it simply means that the OOB error is the average prediction error of a given sample of data when that sample of data is treated as a test sample rather than a train sample (i.e., a tree is evaluated on that data since it has yet to see it). OOB error is also used for other machine learning models implementing something called bootstrap aggregation (bagging). Bagging is the official term for only considering a random sample of the data when random forest creates each tree. It is unique in that it is a random sample that allows

for repetition, meaning that the records for a single participant could be represented more than once in the sample. For more on the theory behind bagging, see work by Ghojogh and Crowley (2023). In addition to the OOB error rate, the output provides a confusion matrix, something that is often used to discuss the performance of a classification method. A confusion matrix follows the form below:

Table 2. Confusion Matrix with Signal Detection Theory Terminology

	True 0	**True 1**
Predicted 0	Correct Rejection	Miss
Predicted 1	False Alarm	Hit

It is ideal to have a high number of both hits and correct rejections and a low number of both false alarms and misses. It is possible that one may wish to allow for more false alarms so as to decrease miss rates in some cases (e.g., a doctor would likely rather have a false positive screening for cancer than miss a cancer diagnosis). In other cases, one may want to minimize false alarms (e.g., in the court system, it is ideal to minimize the number of innocent people who are sent to jail). Thus, it is incredibly beneficial to understand each of these statistics when evaluating the performance of a classification model, as they both factor into calculating accuracy. The `randomForest()` output provides a classification error representing the proportion of a given class which has been misclassified (e.g., a true 0 that was classified as 1 or the reverse). For the model demonstrated, there is no classification error for either class since perfect accuracy occurred.