**METHODS**

**Study Design**

This study was designed to determine what variables predict the cadence at which a walk transitions to a run in children. To address this question we used data collected at the University of Massachusetts, Amherst’s Physical Activity and Health Laboratory to develop a model that predict the cadence at the walk to run transition from an easily measurable set of demographic and metabolic variables.

**Participants**

Some questions

1. When was the data collected and what was the primary outcome?
2. Has the data already been described elsewhere?
3. How were the participant’s recruited?
4. How many minutes were the children asked to walk at a constant pace?
5. How was “run” versus “walk” defined?
6. What is run cadence?

The dataset contained 122 participants who met transitioned from walk to run and maintained their running pace through the duration of their final stage. (50% female) between the ages of 7 and 20. Participants were guided to walk at a constant pace for \_\_\_ minutes. At the end of this interval, the speed was increased by 0.5 mph. The increase in speed was continued until the individual transitioned from walk to run. The experiment concluded at the walk to run transition interval. Of the 122 participants, 69 were able to attain the walk to run interval and only these data were used for model development. A number of the originally available variables on these 69 participants had linear dependencies. The list of independent variables considered for model development after accounting for these relationships are in Table 1 below. The dependent variable for all models was the step cadence of the participant in their final stage of the study (the stage in which they first began running); this is called the walk-to-run transition cadence.

Table 1 - List of considered independent variables and their explanations.

|  |  |
| --- | --- |
| Independent Variable | Explanation |
| Sex | Male or Female |
| Age | Age of participant in years |
| Height | Height of participant in cm |
| Weight | Weight of participant in kg |
| Waist | Waist circumference of participant in cm |
| BMI | Body Mass Index of participant |
| BMI percentile | The percentile of the participant’s BMI |
| BMI z-score | The standardized BMI score for the participant based on age and gender[[1]](#footnote-1) |
| Obesity Classification | 85th percentile BMI classified overweight, 95th percentile BMI classified obese |
| Tanita | Tanita Body Impedence Analysis Measure |
| Walk VO2 |  |
| Run VO2 |  |
| Run METS Youth1 |  |
| Run METS Youth 3 |  |
| Walk METS Youth 3 |  |

The 69 participants are a mix of males and females which are treated identically in model development. Physical attributes differ minimally in youth vice adulthood, and the mean critical characteristics to this study are within a standard deviation across genders (see Table 2).

Table 2 - Subject characteristic table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | Age  (mean ± sd) | BMI  (mean ± sd) | Walk-to-Run Transition Cadence (mean ± sd) |
| Males | 37 | 15.0 ± 3.88 | 22.6 ± 6.01 | 156 ± 10.2 |
| Females | 32 | 14.9 ± 3.70 | 21.9 ± 4.66 | 158 ± 8.92 |

**Mathematical Models**

Before constructing any regression models, two sets of highly correlated (greater than 90%) independent variables were discovered; waist circumference strongly correlates with weight while BMI percentile correlates with BMI z-score. Waist circumference and BMI percentile were removed from consideration because weight is an easier and more practical measure for an individual to attain and because BMI z-score is a more rigorous representation of a person’s BMI by accounting for age and gender.

We then used the leaps[[2]](#footnote-2) package in R[[3]](#footnote-3) to compare regression models for every possible subset of factors, selecting the model with the lowest BIC. All factors in the selected model were significant (have a p-value less than .01) while meeting the necessary linear regression assumptions. The final regression model was:

The independent variables in the model provide To aid users in determining the pace an individual will transition from walk to run, we built a shiny application accessible at the following webpage: <https://dustyturner.shinyapps.io/KidsStep/>. This allows the user to input their age, gender, height, and weight and they will receive a 95% confidence interval at the pace in which they will transition from walk to run. Based off the user’s inputs, the application calculates an individual’s BMI and uses the AGD[[4]](#footnote-4) package in R to determine their BMIz-score.



**Other Considered Models**

Linear regression was not the only type of model considered for this research. A k-means clustering approach feeding a Gaussian mixture model was also considered but did not provide the predictive power of the regression model based on BIC. A survivability model was also considered, but its BIC was also worse than the regression model used.

What Diana Recommended for Us

**METHODS**

**Study Design**

This study was designed to determine what variables predict the cadence at which a walk transitions to a run in children. To address this question we used data collected at the University of Massachusetts, Amherst’s Physical Activity and Health Laboratory to develop \_\_\_\_ models that predict the cadence at the walk to run transition from an easily measurable set of demographic and metabolic variables. We also performed a survival analysis

**Participants**

Some questions for Elroy

1. When was the data collected and what was the primary outcome?
2. Has the data already been described elsewhere?
3. How were the participant’s recruited?
4. How many minutes were the children asked to walk at a constant pace?
5. How was “run” versus “walk” defined?
6. What is run cadence?

The dataset contained 122 participants (x% female) between the ages of X and X.

Participants were guided to walk at a constant pace for \_\_\_ minutes. At the end of this interval, the speed was increased by 0.5 mph. The increase in speed was continued until the individual transitioned from walk to run. The experiment concluded at the walk to run transition interval. Of the 122 participants, 69 were able to attain the walk to run interval and only these data were used for model development. The list of independent variables used for model development were sex, age (years), race, height (cm), weight (kg), waist (cm), body mass index (BMI) (kg/m2), BMI percentile, BMI z-score, and obesity classification (85th percentile BMI classified overweight, 95th percentile BMI classified obese). The dependent variable for all models was run cadence.

Include a “subject characteristic” table with means and SD like in our PBRC paper.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | Age | BMI | Run Cadence |
| Males |  |  |  |  |
| Females |  |  |  |  |

**Mathematical Models**

**Here put in the model you used and how you arrived at a final model. Don’t put the final model, only the process of how you arrived at the final model. For example: Regression models were developed that included x,y,z as covariates and u as the dependent variable. The LASSO algorithm was used to determine which model terms to retain. All statistical modeling was performed using R (R Core Team) You can copy and paste from some of our other articles.**

**App development**

**Describe the process of building the R Shiny App**

**Results**

**Mathematical models**

**Put all findings (what LASSO picked and why it picked it) and the final model, adj R2, signicance of terms etc.**

**R Shiny App**

**Put in a screen shot and where to find the app (URL).**

\_\_\_

Methods as of 17 July 2018

Data

The data used for model building consisted of 122 children and 36 attributes. The children were asked to walk for \_\_\_ minutes before increasing the speed .5 miles per hour. This continued until the children reached a speed in which they transitioned from walk to run and finished their last stage. Only 69 of the 122 participants were able to complete this and will be used in the final analysis.

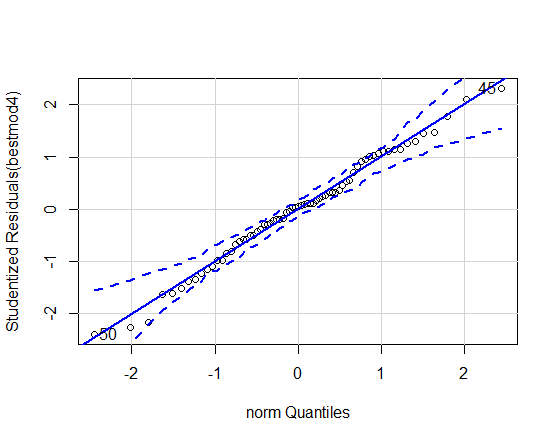
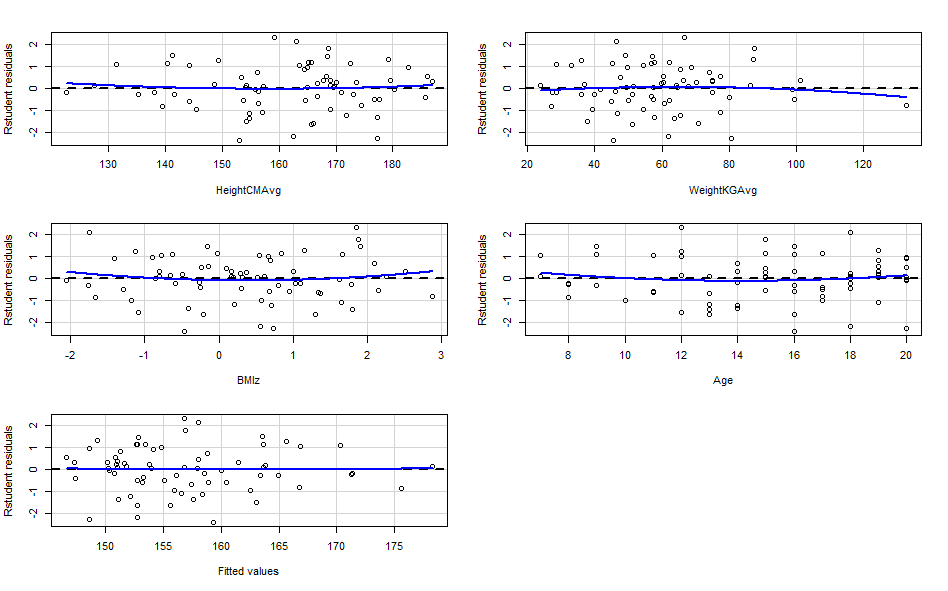
The attributes about the children are listed below. Since, the goal of this model is to find an easily measurable method to predict when a child will transition from walk to run, we did not considers factors that were difficult to measure or were highly collinear with other factors. The final list of factors which we considered for this analysis are in bold below.

Id, **Sex**, **Age**, Agecat, **Race**, **HeightCMAvg**, **WeightKGAvg**, WaistAvg, SittingHeightAvg, Leglength, Tanita.Avg, **BMIcont**, BMIperc, **BMIz**, **Obesecat**, LastWalk, FirstRun, LastFullStage, Transitioned, Transitioned\_FullStage, Walk\_Stage, Walk\_Speed, Walk\_Cadence, Walk\_VO2mlkgmin, Walk\_METSAdult, Walk\_METSYouth1, Walk\_METSYouth2, Walk\_METSYouth3, Run\_Stage, Run\_Speed, **Run\_Cadence**, Run\_VO2mlkgmin, Run\_METSAdult, Run\_METSYouth1, Run\_METSYouth2, Run\_METSYouth3

Model building

In order to determine the best model, we used a best subsets selection method from the leaps[[5]](#footnote-5) package in R. Leaps finds the best model of all sizes based off of BIC. We selected the four factor model below based off it having the lowest BIC of each size model and due to its simplicity. All factors have a P Value of <.01.

Assumptions of independence, normality, linearity, and heteroscedasticity are met. The plots below show evidence to support this claim. Plots are created using the car[[6]](#footnote-6) package in R.



To aid users in determining when they will transition from walk to run, we built a shiny application accessible at the following webpage: <https://dustyturner.shinyapps.io/KidsStep/>. This allows the user to input their age, gender, height, and weight and they will receive a 95% confidence interval at the pace in which they will transition from walk to run. Based off the user’s inputs, the application calculates an individual’s BMI and uses the AGD[[7]](#footnote-7) package in R to determine their BMIz score.

UPDATE as of 2 April 2018

Purpose: Predict the cadence at which a child will go from a walk to a run.

Methods: All statistical analysis was performed in the statistical package, R (R Core Team 2013).

Data Cleaning:

1. Only consider kids who have transitioned to running and completed the full state: Transitioned\_FullStage==1
2. Based on phone conversation and email correspondence, we considered only the following factors:
   1. Dependent Variable: Run\_Cadence
   2. Independent Variables: Sex, Age, Race, HeightCMAvg, WeightKGAvg, WaistAVE, BMIcont, BMIperc, BMIZ, Obesecat
3. There were no linear dependencies
4. There were several highly correlated independent variables >90%:
   1. WaistAvg and WeightKGAvg
   2. BMIperc and BMIz

We removed WaistAvg and BMIperc due to our perceive measurability (WaistAvg) and the future ability for the model to handle the independent variable (BMIper).

1. Final features considered in model selection: Age, Sex, Race, Obesecat, HeightCMAvg, WeightCMAvg, WeightKGAvg, BMIcont, BMIz

Linear Model

1. We created our ‘best’ linear model using the “Leaps” package in R.
   1. This model considers every subset of p predictors and returns the ‘best’ model for each number of predictors (1 to p-1).
   2. The best model for each size are selected based off of performance in AIC/BIC/CIC/DIC (all will rank models of equal size the same so specific criteria is irrelevant).
   3. We chose the best model of each size based off of BIC, where the lowest BIC represents the best model with the fewest variables when compared with other models – if the client does not like this choice, we can re-look the analysis.
   4. BIC: -28.60957
   5. R^2: .514
   6. Model meets all 4 linear modeling assumptions: Linearity, Normality, Heteroscedasticity, Independence

We also propose a second model where we began the model building process with all predictors. We did not consider interpretability nor practicality of use.

Data Cleaning:

1. Only consider kids who have transitioned to running and completed the full state: Transitioned\_FullStage==1
2. Factors in Model:
   1. Dependent Variable: Run\_Cadence
   2. Independent Variables: Age, HeightCMAvg, WeightKGAvg, WaistAvg, SittingHeightAvg, Leglength, Tanita.Avg, BMIcont, BMIperc, BMIz, LastWalk, FirstRun, LastFullStage, Walk\_Speed, Walk\_Cadence, Walk\_VO2mlkgmin, Walk\_METSAdult, Walk\_METSYouth1, Walk\_METSYouth2, Walk\_METSYouth3, Run\_Speed, Run\_Cadence, Run\_VO2mlkgmin, Run\_METSAdult, Run\_METSYouth1, Run\_METSYouth2, Run\_METSYouth3, Sex, Agecat, Race, Obesecat, Walk\_Stage, Run\_Stage
3. There were several linear dependencies. We removed the predictors in ***bold italics***
   1. ***Leglength***, HeightCMAvg, SittingHeightAVG
   2. FirstRun, ***LastFullStage***
   3. LastWalk, ***Walk\_Speed***
   4. ***Walk\_METSAdult***, Walk\_VO2mlkgmin
   5. ***Run\_Speed***, FirstRun
   6. ***Run\_METSAdult***, Run\_VO2mlkgmin
4. There were several highly correlated independent variables >90%: We removed the predictors in ***bold italics***
   1. ***HeightCMAv*** with SittingHeightAVG
   2. WeightKGAvg with ***WaistAVG***,
   3. ***WaistAVG***, WeightKGAvg, BMIcont
   4. BMIcont with ***WaistAvg***
   5. ***BMIperc*** with BMIz
   6. ***Lastwalk*** with FirstRun
   7. ***Walk\_METSYouth1*** with Run\_METSYouth1
   8. ***Walk\_METSYouth2*** with Walk\_METSYouth3
   9. ***Run\_METSYouth2*** with Run\_METSYouth3
5. We removed the following due to our perceive measurability and the future ability for the model to handle the independent variable.
   1. FirstRun
   2. Walk\_Cadence
   3. Agecat
   4. Walk\_Stage
   5. Run\_Stage
6. Final features considered in model selection: Age, WeightKGAvg, HeightAvg, Tanita.Avg, BMIcont, BMIz, Walk\_VO2mlkgmin, Walk\_METSYouth3, Run\_Cadence, Run\_VO2mlkgmin, Run\_METSYouth1, Run\_METSYouth3, Sex, Race, Obesecat

Linear Model

1. We created our ‘best’ linear model using the “Leaps” package in R.
   1. This model considers every subset of p predictors and returns the ‘best’ model for each number of predictors (1 to p-1).
   2. The best model for each size are selected based off of performance in AIC/BIC/CIC/DIC (all will rank models of equal size the same so specific criteria is irrelevant).
   3. We chose the best model of each size based off of BIC, where the lowest BIC represents the best model with the fewest variables when compared with other models – if the client does not like this choice, we can re-look the analysis.
   4. BIC: -25.45861
   5. R^2: .4912
   6. Model meets all 4 linear modeling assumptions: Linearity, Normality, Heteroscedasticity, Independence

Models Below OUTDATED

Purpose: Predict the cadence at which a child will go from a walk to a run.

Data Cleaning:

1. In order to conduct our analysis, we selected rows at which the child began to run. This was the fastest pace for all data as we understand that this is the point where kids went from walking to running.
2. There were two rows (one subject: ID# 68152) which contained missing information about sitting height.
   1. We did not include this subject in our analysis.
   2. However, after conclusion of our analysis, some of our models did not include Sitting Height as a predictor that impacted the cadence in which kids went from walking to running.
   3. Presently we have decided to keep this individual out of all models for consistency, but I do not feel there is a mathematical or statistical reason that this is necessary.
   4. The biggest factor for consideration of re-inserting this data point is: Does the fact we do not have data for this person reflect some sort of unique trait about this person?
3. Features used in model selection: 10 features and 1 response out of 21 total variables.
   1. Used Cadence\_stepsmin (1)
   2. Sex, Age\_years, Race, HeightCMAvg, WeightKGAvg, WaistCMAvg, leglengthCM, Tanita.Avg\_percentbodyfat, BMI\_rawscore, Obese\_status (10)
   3. Due to very high correlation with BMI\_rawscore, we removed BMI\_percentile and BMI\_zscore (2)
   4. Due to very high correlation, we only used HeightCMAvg and not SittingHeightCMAvg (1)
   5. Removed Agecat because it was a less granular version of age.
   6. Removed id as it is unique to each individual (1)
   7. Stage/TreadmillSpeed\_MPH because they are a different version of the Cadence\_stepsmin (2)
   8. Removed V02mlkgmin/METSYouth2/METSAdult because they require an invasive measurement and note practically useful for prediction (2)

Model 1: Linear Regression

1. We created a ‘best subsets’ regression of the above predictors using all the data. Best subsets were selected by lowest Bayesian Information Criterion (BIC) which takes into consideration error, number of samples, and features
2. The best model is below. (5 predictors)
   1. Cadence\_stepsmin~Sex+Age\_years+WaistCMAvg+Racelimited+Tanita.Avg\_percentbodyfat
   2. BIC is -7.789.
   3. Leave one out cross validation Root square mean error (RMSE) is 18.95053.
3. Compare with second best model 6 predictors:
   1. Cadence\_stepsmin~Sex+Age\_years+WaistCMAvg+Racelimited+Tanita.Avg\_percentbodyfat +obeselimited(only ‘obese’)
   2. b. BIC is -7.6769
   3. c. Leave one out cross validation RMSE is 18.73
4. If we limit to only 3 predictors (for simplicity):
   1. Cadence\_stepsmin~Age\_years+WaistCMAvg+Racelimited
   2. BIC is -3.1176
   3. Leave one out cross validation RMSE is 19.967

Model 2: Survival Analysis

1) Best Subsets is not available for Survival Analysis so we used a ‘purposeful selection’ method of arriving at the best model. We tried all factors univariately. All factors that had a p-value below .25, we retained for the ‘saturated model.’ Also, compared with a ‘full model’ and factors that we would have removed from the univariate analysis – if significant in the full model, we retained to begin.

2) This may not be as relevant, because the model coefficients should be interpreted as ‘for every one unit increase in feature x is betax times more likely to ‘run’ than before.

3) The ‘best’ model according to the purposeful selection is:

Survival time ~ Sex+Age\_years+Race2+HeightCMAvg+WeightKGAvg+WaistCMAvg+ Tanita.Avg\_percentbodyfat+BMI\_rawscore+Obese\_status2

Model 3: Cluster Analysis

1. We wanted to see if there is any underlying structure in the static variables (those that don’t change in a subject over short duration). Clustering at high dimensions loses some of its benefit, so we looked at variables with high covariance and eliminated a few of them from our cluster analysis. The highest covariance was between waist and BMI (.958). Waist also had a high covariance with Weight (.919), so we eliminated it. BMI was strongly correlated with Tanita and weight as well (.83 and .889 respectively) so that was also eliminated. Finally, we took out Height as it correlated strongly with Age and leglength (.845 and .947, respectively).
2. We were then left with four primary factors to cluster: Age, leglength, Weight, and Tanita. We looked at a pair-wise depiction of these variables and only one pairing (Weight to Leglength) seemed to have some cluster aspects. However, after further analysis, this secondary cluster was only males as they had the larger height elements.
3. We used a GMM (Gaussian Mixture Model) in order to find the best possible model with the optimal number of clusters based on a BIC metric. The model assigns a vector of probabilities associated to each observation indicating the posterior probability of belonging to the respective clusters.
4. After doing this clustering, we then put the clusters into a linear regression model, using their assigned cluster as a factor. The results of this were inconclusive in that the cluster did not help in predicting run cadence time in children.

1. The AGD package in R was used for determining the BMI Z-score (cite). [↑](#footnote-ref-1)
2. Thomas Lumley based on Fortran code by Alan Miller (2017). leaps: Regression Subset Selection. R package version 3.0. https://CRAN.R-project.org/package=leaps [↑](#footnote-ref-2)
3. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/. [↑](#footnote-ref-3)
4. Van Buuren S (2018). AGD: Analysis of Growth Data. R package version 0.39, <URL: https://CRAN.R-project.org/package=AGD>

   https://CRAN.R-project.org/package=AGD>. [↑](#footnote-ref-4)
5. Thomas Lumley based on Fortran code by Alan Miller (2017). leaps: Regression Subset Selection. R package version 3.0. https://CRAN.R-project.org/package=leaps [↑](#footnote-ref-5)
6. John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition.

   Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion [↑](#footnote-ref-6)
7. Van Buuren S (2018). AGD: Analysis of Growth Data. R package version 0.39, <URL:

   https://CRAN.R-project.org/package=AGD>. [↑](#footnote-ref-7)