Novel pediatric height outlier detection for electronic health records: machine learning with monotonic Bayesian Additive Regression Trees

RA Sparapani, BQ Teng, J Hilbrands,
R Pipkorn, MB Feuling, PS Goday
Journal of Pediatric Gastroenterology and Nutrition 2002

This research was supported by
the Children's Wisconsin Foundation through the generous support of the LAMBS fund and Abraham Family Fund

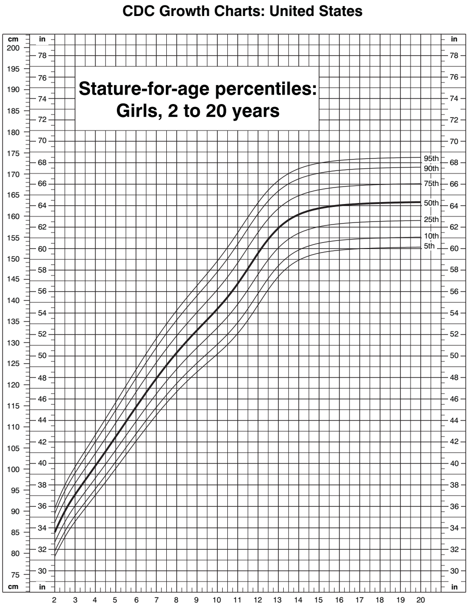
Outline

- Motivation: a clinical application in chronically ill children potentially compounded by malnutrition
- What is monotonic BART (mBART)
- Nonparametric outlier detection and monotonic advantages
- Nonparametric marginal effect estimation
- Returning to the real data example

Chronically ill children and potential height outliers

- This data is from the electronic health records (EHR) of a large children's health care system
- Chronically ill children are often at high risk for malnutrition
- Typically this is assessed by comparison to Centers for Disease Control (CDC) growth chart benchmarks
- CDC inputs are age, gender, height and weight
- Age and gender are extremely reliable
- However, height and weight are prone to outliers and there is practically NO quality control for these measures i.e., the ground truth of height outliers is largely unknown
- There are about an order of magnitude more height (3%) than weight outliers (0.2%) per measurement (Phan et al. 2020 Scientific Reports)
- Determining malnourishment requires height outlier detection
- Furthermore, this method should be robust to weight outliers that are harder to identify but thankfully less prevalent
- Proposed EHR height outlier removal methods are either too simplistic or too complex to implement (such as Phan et al.)

https://www.cdc.gov/growthcharts



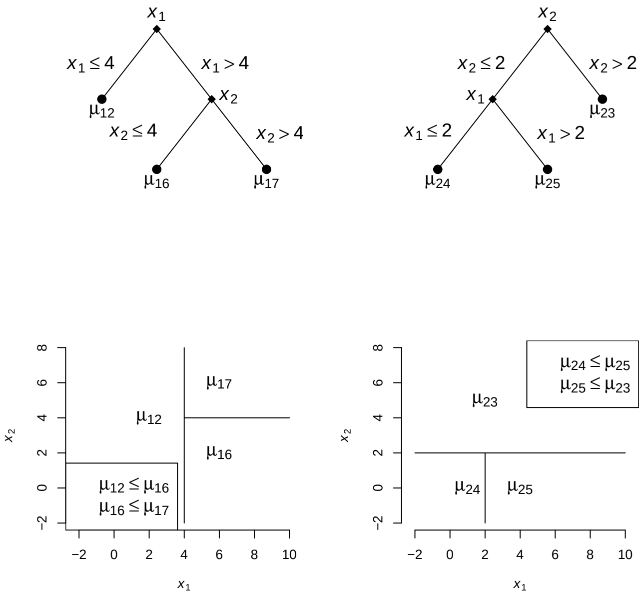
Motivating Example: Growth Chart Outliers

- The US Centers for Disease Control and Prevention (CDC) as well as the World Health Organization (WHO) have developed growth charts for childhood development: height by age, weight by age, body mass index by age and weight by height
- Here we will focus on height, y_t, by age in months, $t = 24, \ldots, 215$ (2 to 17 years old)
- The CDC uses the LMS method via natural cubic splines (Cole and Green 1992 *Statistics in Medicine*)
- Three parameters estimated by penalized maximum likelihood the Box-Cox power transformation, L_t; the mean, M_t; and the coefficient of variation, S_t

$$z_t = \left\{ \begin{array}{cc} \frac{-1+(y_t/M_t)^{L_t}}{L_t S_t} & L_t \neq 0 \\ \frac{\log(y_t/M_t)}{S_t} & L_t = 0 \end{array} \right\} \sim \mathrm{N}(0,\ 1)$$

- CDC/WHO guidelines say values of $z_t < -6$ or $z_t > 6$ are outliers but this will catch only the most extreme outliers
- Regardless of the exact cutoff, this outlier method is called Standard Deviation Scores (SDS), i.e., Height SDS

Monotonic example: increasing in x_1 and x_2



Monotonic BART (mBART)

Chipman et al. 2021 *Bayesian Analysis*

- $f \overset{\text{prior}}{\sim} \text{mBART}$
- A function f is monotone with respect to x_j if f satisfies $f(\ldots, x_{j-1}, x_j + \Delta x, x_{j+1}, \ldots) \geq f(\ldots, x_{j-1}, x_j, x_{j+1}, \ldots)$ for all $\Delta x > 0$ (increasing/nondecreasing) or for all $\Delta x < 0$ (decreasing/nonincreasing)
- Constraint Conditions for Tree Monotonicity
  A tree function $g(x; \mathcal{T}, \mathcal{M})$ will be monotone in coordinate x_j if the leaf value of each of its terminal node regions is
  (a) not greater than the minimum level of all of its above-neighbor regions with respect to x_j and
  (b) not less than the maximum level of all of its below-neighbor regions with respect to x_j

Monotonic BART (mBART)

Chipman et al. 2021 *Bayesian Analysis*

- The leaf prior for BART $\quad \mu_j|\mathcal{T} \overset{\text{prior}}{\sim} \mathrm{N}\left(0,\ \sigma_\mu^2\right)$
- Consider the simplest case of two monotonic leaves in mBART (relying on the results of Azzalini 1985 *Scand J Stat*)

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \overset{\text{prior}}{\sim} \mathrm{N}_2\left(\vec{0}_2,\ c^2\sigma_\mu^2 I_2\right) \mathrm{I}(\mu_1 < \mu_2) \text{ where } c^2 = \frac{\pi}{\pi - 1} \approx 1.47$$

Equivalent to skew Normal marginals where $\mathrm{V}[\mu_1] = \mathrm{V}[\mu_2] = \sigma_\mu^2$

$$\mu_1 \overset{\text{prior}}{\sim} \phi\left(\frac{\mu_1}{c\sigma_\mu}\right)\Phi\left(\frac{-\mu_1}{c\sigma_\mu}\right) \qquad \mathrm{E}[\mu_1] = \frac{-\sigma_\mu}{\sqrt{\pi - 1}}$$

$$\mu_2 \overset{\text{prior}}{\sim} \phi\left(\frac{\mu_2}{c\sigma_\mu}\right)\Phi\left(\frac{\mu_2}{c\sigma_\mu}\right) \qquad \mathrm{E}[\mu_2] = \frac{+\sigma_\mu}{\sqrt{\pi - 1}}$$

BART vs. mBART priors

Default BART prior settings				
$\boldsymbol{\alpha = 0.95, \beta = 2}$				
Number of leaves	1	2	3	4+
Prior probability	0.05	0.55	0.27	0.13
Default mBART prior settings				
$\boldsymbol{\alpha = 0.25, \beta = 0.8}$				
Number of leaves				
Comparable with BART due to a different sampling approach				

Nonparametric outlier detection

- *Monotonicity provides additional robustness to outliers since f can't just go up before an outlier and back down after (or vice versa)*
- We have *population* predictions of the form $\hat{y}_{ij} = \mathrm{E}\left[y_{ij}\right] = \mu + \hat{f}(x_{ij})$ where $j = 1, \ldots, n_i$ (recall, μ is just a constant roughly centering the population)
- *But these expectations are biased except for the average child*
- We need to adjust these up or down for a given subject
- So let $m_i = \mathrm{median}_j(y_{ij} - \hat{y}_{ij})$ (median rather than mean to be robust to outliers)
- Now, we make *personalized* predictions $\tilde{y}_{ij} = m_i + \hat{y}_{ij}$
- We define the *relative error* of these as $d_{ij} = (y_{ij} - \tilde{y}_{ij})/\tilde{y}_{ij}$
- Outliers are defined as $|d_{ij}| > \delta$ where δ can be determined from the Receiver Operating Characteristic (ROC) curve
- And the discriminating performance of the method is assessed by the area under the ROC curve

Returning to the real data example

- Constructed two independent cohorts of chronically ill children
  - 2-8 years old
  - measured at least every 120 days on average
  - followed for at least 2 years
- Training cohort: 1376 children with height outliers unknown
  39491 measurements: 28.7/child on average
- Validation cohort: 318 children
  7378 measurements: 23.2/child on average
  manually reviewed to determine height outliers
  however, the *ground truth* is fallible
  i.e., retrospective: we can't just re-measure the child's height
- Heights in the Training cohort fit with mBART to
  age, gender, race/ethnicity and weight

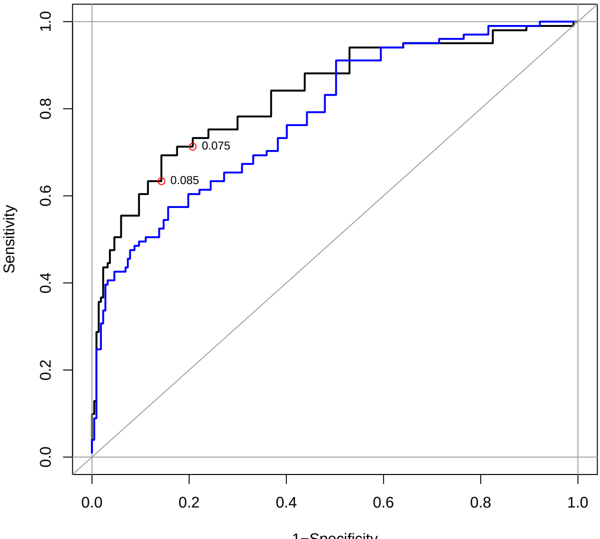
Returning to the real data example

- Outlier detection conducted for the Validation cohort
- The area under the Receiver Operating Characteristic (ROC) curve was excellent: 0.841
- By comparison, if you use the height SDS by age growth chart, the area is only 0.776
- Based on ROC curve, two relative error cutoffs considered Aggressive, 0.075; and Conservative, 0.085

Real data summary

	Training		Validation	
	1376		318	
Children	n	(%)	n	(%)
Female	594	(43.2%)	132	(41.5%)
White	783	(56.9%)	189	(59.4%)
Black	313	(22.7%)	66	(20.8%)
Other	280	(20.3%)	63	(19.8%)
Children with outliers	Unk.		101	(31.8%)
Measurements	m		m	
Height (cm)	39491		7378	
	Mean	(SD)	Mean	(SD)
Measurements/child	28.7		23.2	
First visit age 2	86.4	(8.8)	84.5	(6.6)
Last visit age 5	111.3	(8.4)	109.5	(9.4)
mBART R^2	82.2%		75.3%	

Receiver Operating Characteristic curve (AUC): mBART (0.841) vs. SDS (0.776)



Aggressive cutoff 0.075

- B: mBART outlier detection
- C: clinical review ground truth
- Outlier: 0 (False), 1 (True)

	B=0	B=1	
C=0	TN=172	FP=45	M=217
C=1	FN=29	TP=72	Q=101
	N=201	P=117	T=318

$$\text{Sensitivity or Recall} = \mathbf{P}[\boldsymbol{B} = \mathbf{1}|\boldsymbol{C} = \mathbf{1}] = \frac{\boldsymbol{TP}}{\boldsymbol{Q}} = \frac{\mathbf{72}}{\mathbf{101}} = \mathbf{0.713}$$

$$\text{Specificity} = \mathbf{P}[\boldsymbol{B} = \mathbf{0}|\boldsymbol{C} = \mathbf{0}] = \frac{\boldsymbol{TN}}{\boldsymbol{M}} = \frac{\mathbf{172}}{\mathbf{217}} = \mathbf{0.793}$$

$$\text{PPV or Precision} = \mathbf{P}[\boldsymbol{C} = \mathbf{1}|\boldsymbol{B} = \mathbf{1}] = \frac{\boldsymbol{TP}}{\boldsymbol{P}} = \frac{\mathbf{72}}{\mathbf{117}} = \mathbf{0.615}$$

$$\text{NPV} = \mathbf{P}[\boldsymbol{C} = \mathbf{0}|\boldsymbol{B} = \mathbf{0}] = \frac{\boldsymbol{TN}}{\boldsymbol{N}} = \frac{\mathbf{172}}{\mathbf{201}} = \mathbf{0.856}$$

Conservative cutoff 0.085

- B: mBART outlier detection
- C: clinical review ground truth
- Outlier: 0 (False), 1 (True)

	B=0	B=1	
C=0	TN=186	FP=31	M=217
C=1	FN=37	TP=64	Q=101
	N=223	P=95	T=318

$$\text{Sensitivity or Recall} = \mathbf{P}[\boldsymbol{B} = 1|\boldsymbol{C} = 1] = \frac{\boldsymbol{TP}}{\boldsymbol{Q}} = \frac{\mathbf{64}}{\mathbf{101}} = \mathbf{0.634}$$

$$\text{Specificity} = \mathbf{P}[\boldsymbol{B} = 0|\boldsymbol{C} = 0] = \frac{\boldsymbol{TN}}{\boldsymbol{M}} = \frac{\mathbf{186}}{\mathbf{217}} = \mathbf{0.857}$$

$$\text{PPV or Precision} = \mathbf{P}[\boldsymbol{C} = 1|\boldsymbol{B} = 1] = \frac{\boldsymbol{TP}}{\boldsymbol{P}} = \frac{\mathbf{64}}{\mathbf{95}} = \mathbf{0.674}$$

$$\text{NPV} = \mathbf{P}[\boldsymbol{C} = 0|\boldsymbol{B} = 0] = \frac{\boldsymbol{TN}}{\boldsymbol{N}} = \frac{\mathbf{186}}{\mathbf{223}} = \mathbf{0.834}$$

Aggressive cutoff: targeted smoothing BART with monotonic weight

Starling et al. Annals of Applied Statistics 2020

- B: mBART outlier detection
- C: clinical review ground truth
- Outlier: 0 (False), 1 (True)

	B=0	B=1	
C=0	TN=165	FP=52	M=217
C=1	FN=27	TP=74	Q=101
	N=192	P=126	T=318

$$\text{Sensitivity or Recall} = \mathbf{P}[B=1|C=1] = \frac{TP}{Q} = \frac{74}{101} = 0.732$$

$$\text{Specificity} = \mathbf{P}[B=0|C=0] = \frac{TN}{M} = \frac{165}{217} = 0.760$$

$$\text{PPV or Precision} = \mathbf{P}[C=1|B=1] = \frac{TP}{P} = \frac{74}{126} = 0.587$$

$$\text{NPV} = \mathbf{P}[C=0|B=0] = \frac{TN}{N} = \frac{165}{192} = 0.859$$

Aggressive cutoff: females only

- B: mBART outlier detection
- C: clinical review ground truth
- Outlier: 0 (False), 1 (True)

	B=0	B=1	
C=0	TN=70	FP=20	M=90
C=1	FN=11	TP=31	Q=42
	N=81	P=51	T=132

$$\text{Sensitivity or Recall} = \mathbf{P}[B = 1|C = 1] = \frac{TP}{Q} = \frac{31}{42} = 0.738$$

$$\text{Specificity} = \mathbf{P}[B = 0|C = 0] = \frac{TN}{M} = \frac{70}{90} = 0.778$$

$$\text{PPV or Precision} = \mathbf{P}[C = 1|B = 1] = \frac{TP}{P} = \frac{31}{51} = 0.608$$

$$\text{NPV} = \mathbf{P}[C = 0|B = 0] = \frac{TN}{N} = \frac{70}{81} = 0.864$$

Aggressive cutoff: non-whites only

- B: mBART outlier detection
- C: clinical review ground truth
- Outlier: 0 (False), 1 (True)

	B=0	B=1	
C=0	TN=60	FP=19	M=79
C=1	FN=11	TP=39	Q=50
	N=71	P=58	T=129

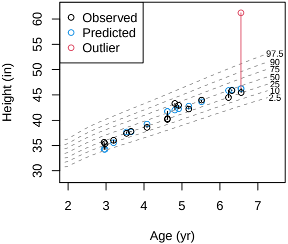
$$\text{Sensitivity or Recall} = \mathbf{P}[B = 1|C = 1] = \frac{TP}{Q} = \frac{39}{50} = 0.780$$

$$\text{Specificity} = \mathbf{P}[B = 0|C = 0] = \frac{TN}{M} = \frac{60}{79} = 0.759$$

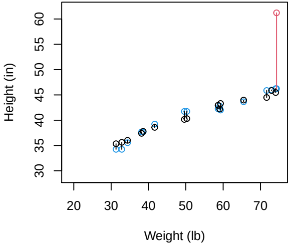
$$\text{PPV or Precision} = \mathbf{P}[C = 1|B = 1] = \frac{TP}{P} = \frac{39}{58} = 0.672$$

$$\text{NPV} = \mathbf{P}[C = 0|B = 0] = \frac{TN}{N} = \frac{60}{71} = 0.845$$

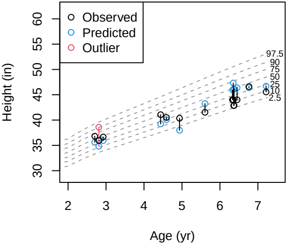
True Positives



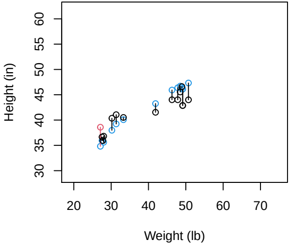
(a) Sex F, Race W



(b) Sex F, Race W

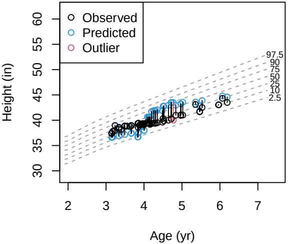


(c) Sex F, Race B

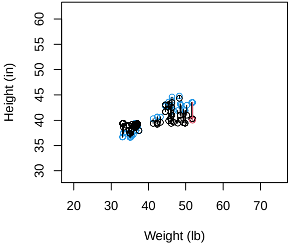


(d) Sex F, Race B

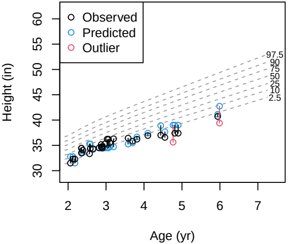
False Positives



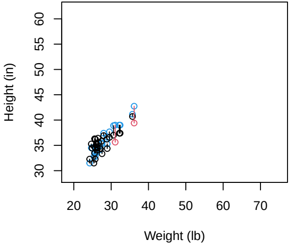
(a) Sex M, Race W



(b) Sex M, Race W

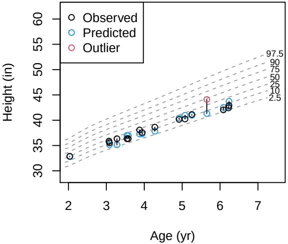


(c) Sex M, Race B

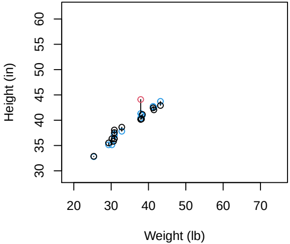


(d) Sex M, Race B

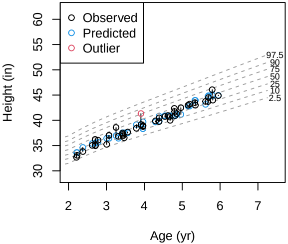
False Negatives



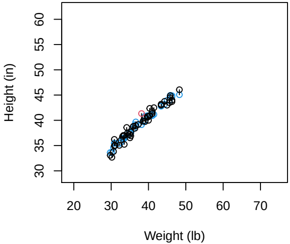
(a) Sex F, Race W



(b) Sex F, Race W



(c) Sex M, Race W

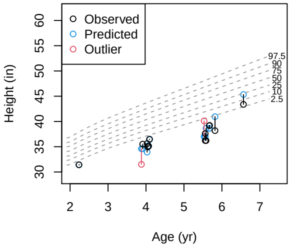


(d) Sex M, Race W

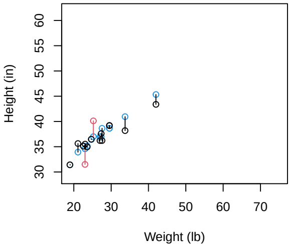
Conclusions

- We constructed our new outlier detection methodology based on nonparametric machine learning via monotonic BART
- This automated method's performance was deemed to be adequate via an indpendent validation cohort
- Modern methodology leads to a simply-tuned single rule as opposed to complex simultaneous tuning of multiple rules that have been proposed based on classic methods
- For EHR heights/weights, the ground truth is unknown prospective corrections are rarely performed and retrospective attempts to identify outliers manually are fallible

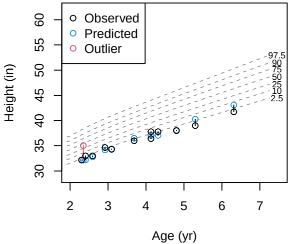
True Positives



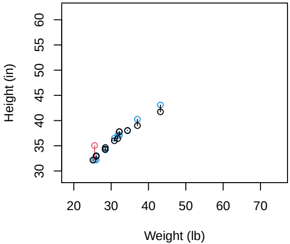
(a) Sex M, Race O



(b) Sex M, Race O

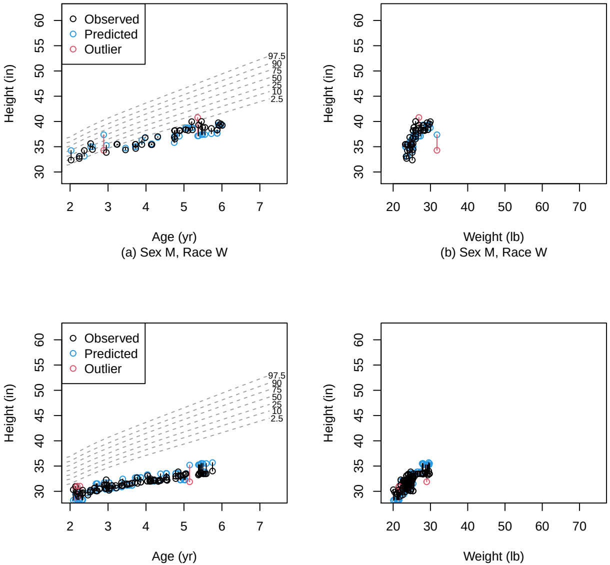


(c) Sex M, Race W



(d) Sex M, Race W

False Positives



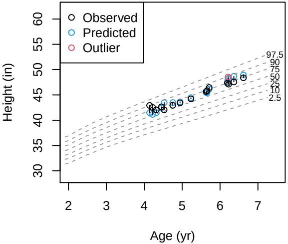
(a) Sex M, Race W

(b) Sex M, Race W

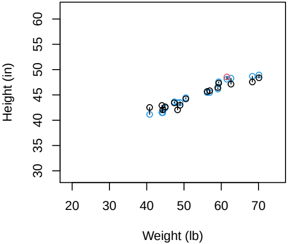
(c) Sex M, Race W

(d) Sex M, Race W

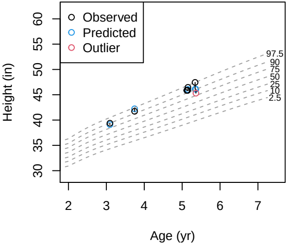
False Negatives



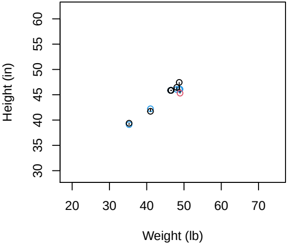
(a) Sex M, Race B



(b) Sex M, Race B



(c) Sex F, Race W



(d) Sex F, Race W