Synthesizing Statistical and Mathematical Models to Address Non-Positivity

Paul Zivich
Assistant Professor
Department of Epidemiology
University of North Carolina at Chapel Hill

May 14, 2026

Acknowledgments[1]

Funding: USA NIH: K01AI177102

Conflicts of Interest: None

Acknowledgments: Stephen Cole, Jess Edwards, Bonnie Shook-Sa, Eric Lofgren

Disclaimer: All errors are my own

Publication: Zivich et al. *J Epidemiol Community Health* 2026;80:352-356

pzivich@unc.edu pzivich pausalz@bsky.social

[1] Footnotes are for asides or references

Motivating Example

Question: What is the mean systolic blood pressure (SBP) in mm Hg among children and adolescents aged 2-17 in the United States?

Parameter: $\mu := E[Y]$

- Y_i: SBP for unit i
- $E[\cdot]$: expected value function

Data: NHANES 2017-2018

Motivating Example

Systematic Error: SBP is missing for 44% of children

- R_i: indicator of SBP being observed

Extent of missingness means that estimate is sensitive to choices about missing data[2]

A complete-case analysis assumes marginal exchangeability

$$E[Y] = E[Y \mid R = 1]$$

[2] While a relatively simple descriptive question, we can consider causal inference to be a missing data problem and thus link the methods used here for confounding

A More Plausible Assumption

SBP is missing completely at random conditional on X

- X_i: age in years

$$E[Y \mid X = x] = E[Y \mid X = x, R = 1] \text{ for all values of } x$$

Under this assumption, μ is identified as[3]

$$E[Y] = E\left[E(Y \mid X = x, R = 1)\right]$$

or a weighted average of the complete-case means within strata of X

[3] Estimate with direct maximum likelihood estimator, g-computation

Implication of Exchangeability

Positivity: all unique patterns of X have a nonzero probability of having Y being measured

$$\Pr(R = 1 \mid X = x) > 0 \text{ for all } x \in \mathcal{X}$$

Every unique age in the population has a non-zero probability of having Y observed

- Statement about population
- Deterministic or structural positivity

Why Positivity Matters

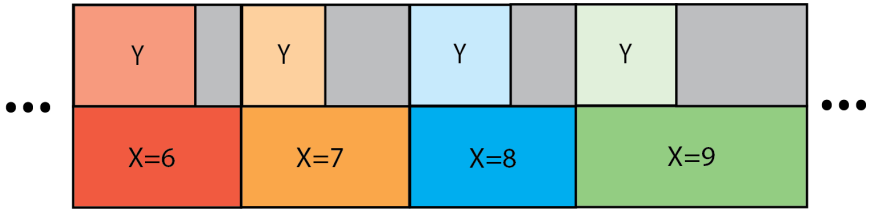
Formal: following definition of conditional expectation

$$E[Y \mid X = x, R = 1] = \frac{E[Y R\, I(X = x)]}{\Pr(R = 1 \mid X = x) \Pr(X = x)}$$

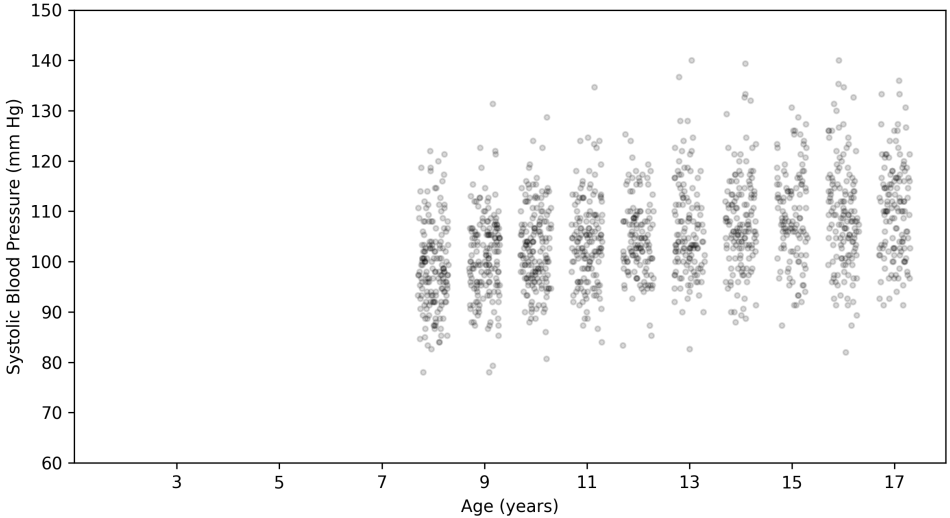
Intuitive: need some with $X = x$ to have Y observed at least sometimes so can 'stand-in' for those without Y measured

So, *positivity* comes along with *exchangeability*

Positivity for NHANES



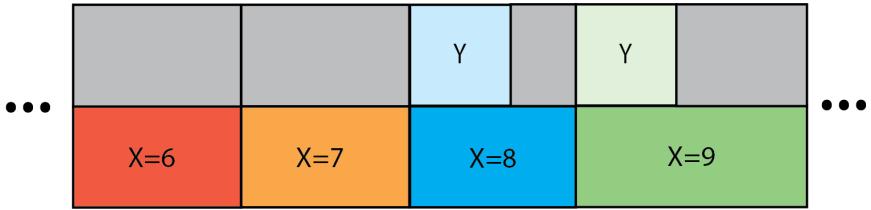
NHANES Data


Systolic Blood Pressure (mm Hg)
150
140
130
120
110
100
90
80
70
60
3
5
7
9
11
13
15
17
Age (years)

Nonpositivity by Age in NHANES

In NHANES 2017-2018,

- "BP is measured on participants 8 years and older"[4]
- $\Pr(R = 1 \mid X = x) = 0$ for $x \in \{2, 3, 4, 5, 6, 7\}$



[4] https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/BPX_J.htm

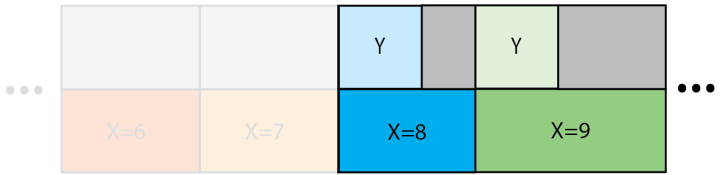
Addressing Positivity Violations

Option 1: Change Parameter

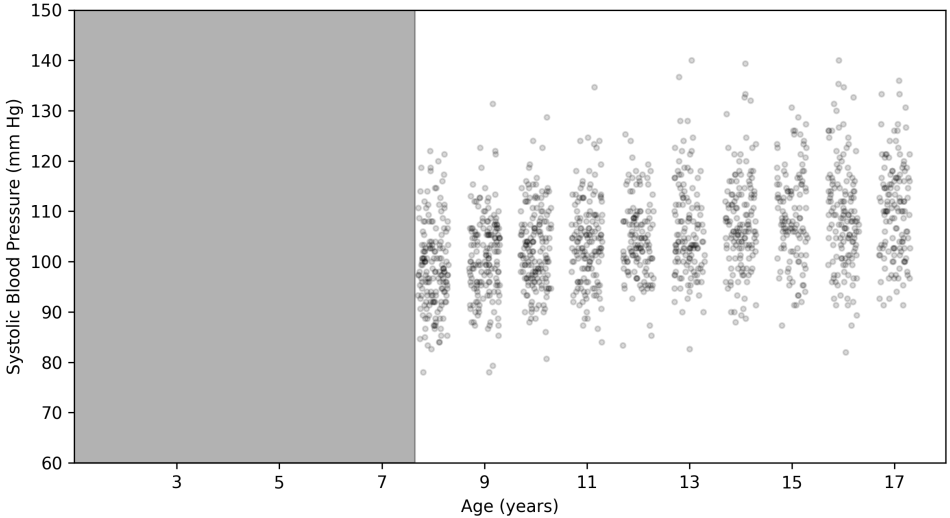
Parameter (revised): mean SBP for those 8 or older

$$E[Y \mid X \geq 8]$$

Question: what is the mean systolic blood pressure (SBP) in mm Hg among children and adolescents aged 8-17 in the US?



NHANES Data

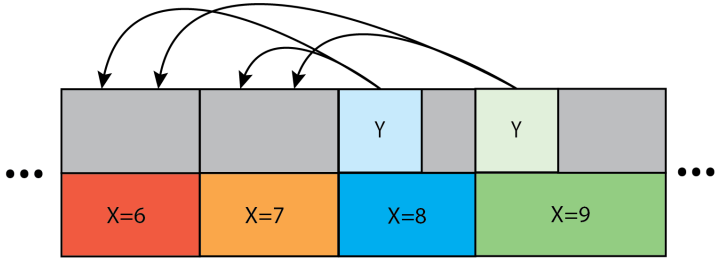

Systolic Blood Pressure (mm Hg)
150
140
130
120
110
100
90
80
70
60
3
5
7
9
11
13
15
17
Age (years)

Option 2. Extrapolate from Statistical Model

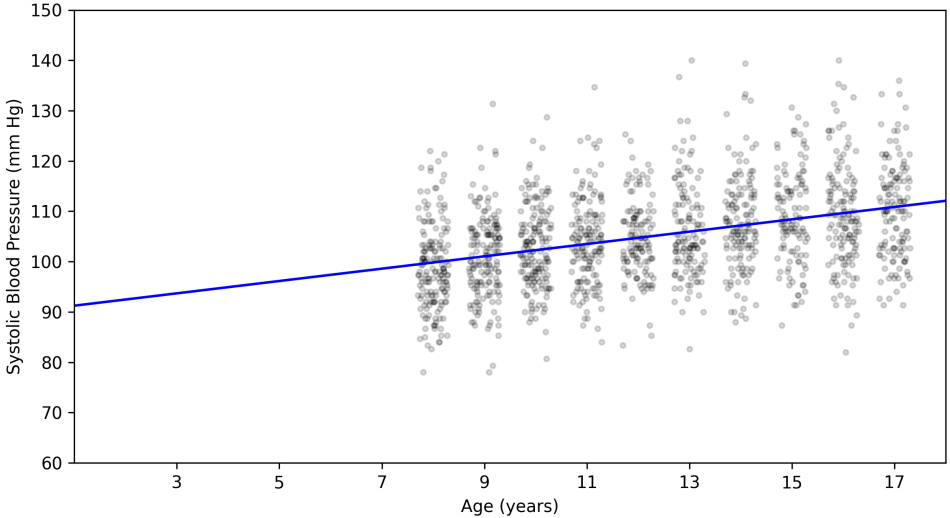
Fit a parametric model

$$E[Y \mid X \geq 8, R = 1] = \beta_0 + \beta_1 X$$

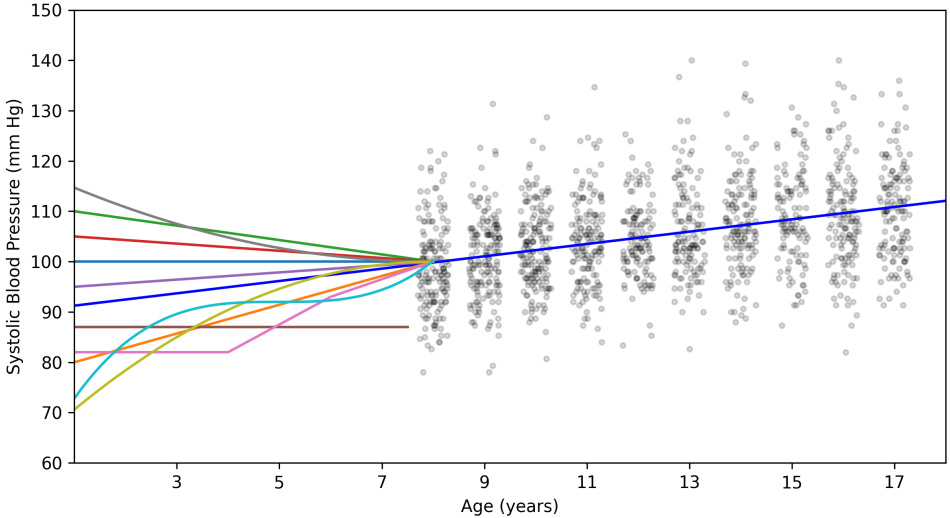
then fill in missing SBP with model predictions



NHANES Data


Systolic Blood Pressure (mm Hg)
150
140
130
120
110
100
90
80
70
60
3
5
7
9
11
13
15
17
Age (years)

NHANES Data


Systolic Blood Pressure (mm Hg)
150
140
130
120
110
100
90
80
70
60
3
5
7
9
11
13
15
17
Age (years)

2. Extrapolate

Use of this model has two assumptions:

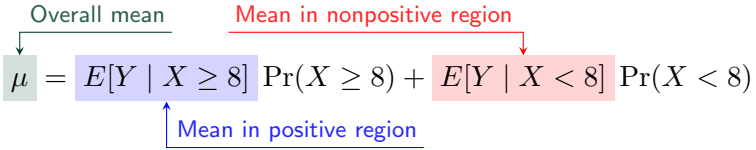
a. Linear relationship with age for those 8-17
b. Linear relationship extends to those 2-7

(a) is in-principle checkable

(b) is not checkable under nonpositivity

Option 3: Bounds and Synthesis Modeling

Divide parameter of interest in parts



Here,

$$E[Y \mid X \geq 8] = E[E(Y \mid X, R = 1, X \geq 8) \mid X \geq 8]$$

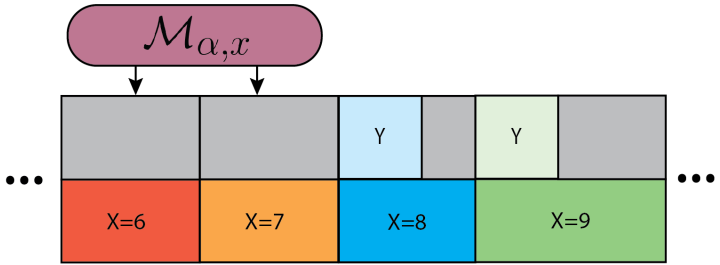
from conditional exchangeability in positive region

Option 3: Bounds and Synthesis Modeling

Available data leaves us stuck with $E[Y \mid X < 8] = ?$

To progress, will appeal to external information for

a. Bounds
b. Sensitivity Analysis
c. Point Identification



Option 3a: Bounds

In our context, know that SBP must be a (finite) positive number, so

$$E[Y \mid X = x] \in (0, \infty) \text{ for all } x \in \{2, 3, 4, 5, 6, 7\}$$

Context allows us to further narrow the plausible region to

$$E[Y \mid X = x] \in [140, 200] \text{ for all } x \in \{2, 3, 4, 5, 6, 7\}$$

as these boundary values would constitute medical emergencies

Therefore,

$$50 + \Pr(X \geq 8) \left\{ E[Y \mid X \geq 8] - 50 \right\} \leq \mu \leq 200 + \Pr(X \geq 8) \left\{ E[Y \mid X \geq 8] - 200 \right\}$$

Option 3b: Sensitivity Analysis

Rather than evaluate at the extremes, evaluate

$$\mu = E[E(Y \mid X, R = 1, X \geq 8) \mid X \geq 8] \Pr(X \geq 8) + q \Pr(X < 8)$$

where $q \in [50, 200]$

Assess how μ varies as a function of the possible means for the nonpositive region

- May help to further contextualize the bounds

Option 3c: Leverage a Mathematical Model

Previous only provide range of equally plausible values

- Can leverage more specific information when available
- Published distributions of SBP by age, gender, height[5]

Use this additional information to construct a mathematical model that produces a probability distribution $\mathcal{M}_{\alpha,x}$ for age x given parameters α, and assume

$$E[Y \mid X = x] \in \mathcal{M}_{\alpha,x} \text{ for all } x \in \{2, 3, 4, 5, 6, 7\}$$

[5] Flynn et al. 2017 *Pediatrics* 140(3):e20171904

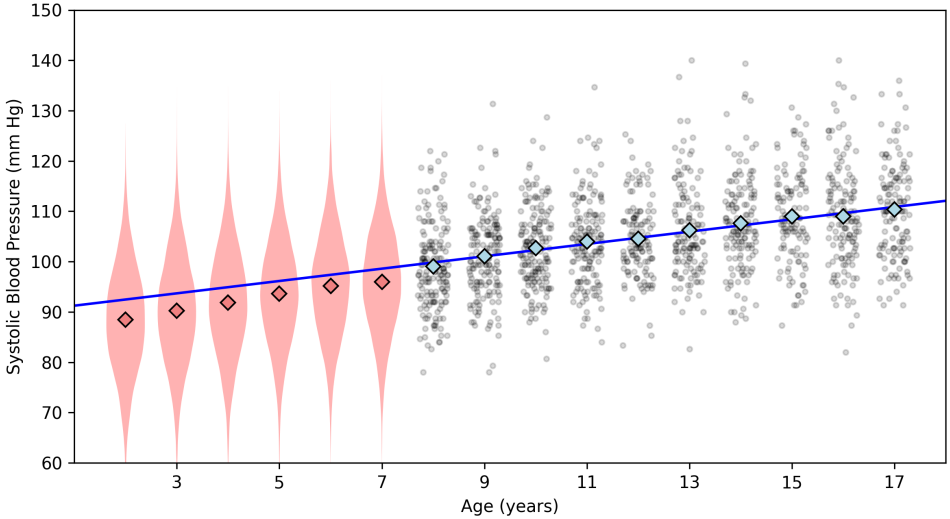
Option 3c: Synthesis of Statistical and Mathematical Models

Then can combine statistical and mathematical models for identification

$$\mu = E[E(Y \mid X, R = 1, X \geq 8) \mid X \geq 8] \Pr(X \geq 8) + E[E(\mathcal{M}_{\alpha,x} \mid X, X < 8) \mid X < 8] \Pr(X < 8)$$

Build $\mathcal{M}_{\alpha,x}$ from published age-, gender-, and height-specific SBP distributions

NHANES Data


Systolic Blood Pressure (mm Hg)
150
140
130
120
110
100
90
80
70
60
3
5
7
9
11
13
15
17
Age (years)

Estimators for Option 3

Option 3a/b: estimate $E[E(Y \mid X, R = 1, X \geq 8) \mid X \geq 8]$ using statistical modeling, [6] set q, solve for μ.[7]

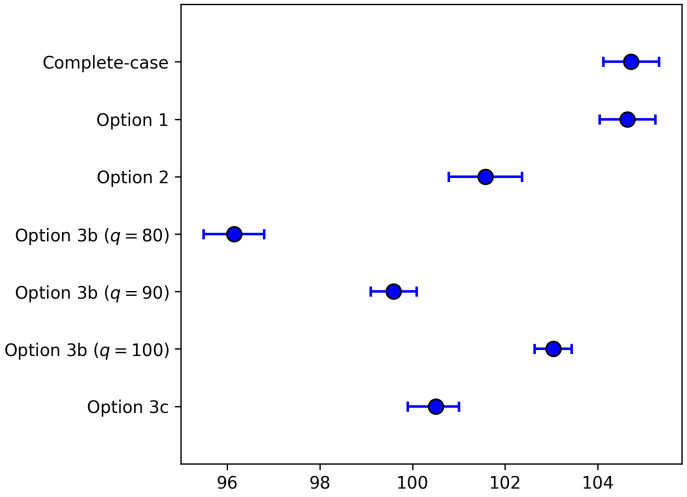
Option 3c: Monte Carlo procedure to capture uncertainty

1. Resample with replacement observed data
2. Estimate a statistical model for $E(Y \mid X, R = 1, X \geq 8)$ and then impute Y for observations with $X \geq 8$
3. For those $X < 8$, impute Y using a random draw from the mathematical model covariate-specific distributions
4. Take the mean of the imputed Y for all observations

[6] Here, I use g-computation for ease but could also use IPW or AIPW. The publication applies all three

[7] Use the pointwise Confidence Intervals which are conservative, Richardson et al. *Stat Sci* 2014;29(4):596-618

Results by Method



	Mean	95% CI
Complete-case	105	(104, 105)
Option 1	105	(104, 105)
Option 2	102	(101, 102)
Option 3b ($q = 80$)	96	(95, 97)
Option 3b ($q = 90$)	100	(99, 100)
Option 3b ($q = 100$)	103	(103, 103)
Option 3c	100	(100, 101)

Conclusions

Summary

Positivity is often neglected as an assumption

- But it remains vital for identification of many parameters

Restricting parameters to positive regions often reasonable

- But may limit utility of research
- Extrapolation premised on strong assumptions
- Synthesis modeling allows for exploration and weakening of extrapolation assumptions

Future Directions

Compare statistical and mathematical models for positive region

- May provide some assurance for mathematical model
- Regularize statistical model for efficiency

Exchangeability (and thus positivity) is used to address other systematic errors

- Approach can be used to address nonpositivity in these settings[8]

Analysis of mathematical models and improving their robustness[9]

[8] Originally proposed in the context of transporting trials: Zivich et al. *Epidemiology* 2024;35(1):23-31, Zivich et al. *JRSSA* 2025;188(1):158-180

[9] Some work on formalizing identification with mathematical models: Zivich *arXiv:2511.01960*

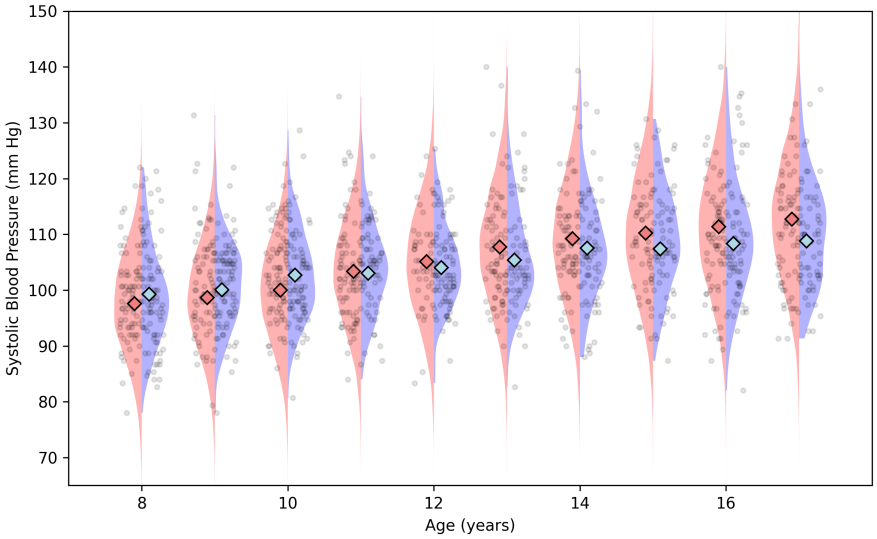
Thank You!

Publication: Zivich et al. *J Epidemiol Community Health* 2026;80:352-356

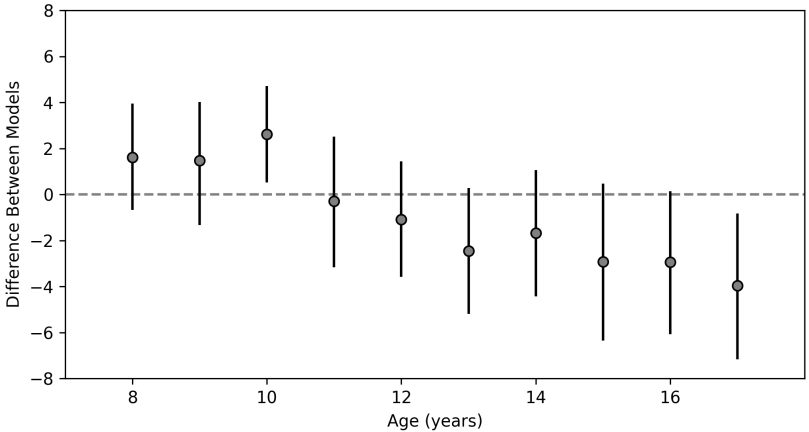
pzivich@unc.edu pzivich pausalz@bsky.social

Appendix

Comparing Statistical and Mathematical Models



Comparing Statistical and Mathematical Models[10]



[10] Intervals are pointwise 95% confidence intervals, so anti-conservative for overall comparison