Using Machine Learning Methods to Identify Potential Construct Validity and Measurement Error Disparities in Health Outcomes from a U.S. National Survey

Morgan Earp, PhD
Lauren Rossen, PhD
Sarah Forrest, MPH
Trent Buskirk, PhD

2024 FCSM Research and Policy Conference

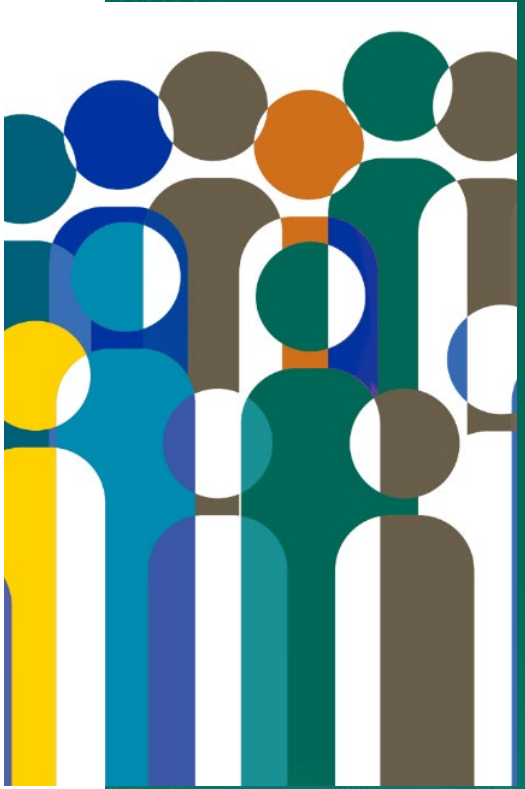
Background

- Measurement equity is critical for the accurate assessment of disparities in health outcomes
- Estimated disparities in the prevalence of a given health outcome can be affected by whether the outcome is measured or self-reported
- Objective: To assess potential measurement error inequities across sociodemographic characteristics for four common health outcomes collected in a national health survey

Research Questions

1. Do self-reporting errors vary across sociodemographic subgroups?
   - e.g., age, sex, race/ethnicity, education level, marital status, insurance coverage type, and poverty status

2. Are identified self-reporting errors significantly different from zero?



Methodology

- Recursive Partitioning for Modeling Survey Data (RPMS)
  - We used conditional linear regression trees (CLRT) via the `rpms` R package (Toth, 2022) to identify subgroups with larger differences between measured and self-reported health outcomes
- Model specification
  - Clustering by respondent
  - Permutations: 25,000
  - P-value: 0.0001

Methodology

- **`rpms` Package Highlights**:
  - Recursively partitions the dataset, fitting a specified least squares linear model on each node separately
  - Algorithm has an unbiased variable selection and accounts for complex sample design
  - Returns a tree that can be used for identifying key group differences in terms of means, but also intercepts and slopes
    - The intercept represents the measured prevalence
    - The slope represents the measurement error

Methodology

- **Data**:
  - National Health and Nutrition Examination Survey (NHANES) 2015-2016 through 2017-March 2020
    - ~5,000 persons per year
    - Household interviews and in-person health examinations
    - Includes both self-reported data (interview component) and measured data (examination component)
  - Did not include weights or survey design information
    - Objective was to quantify potential measurement error inequities in a sample, not to infer error magnitude for the target population

Methodology

- **Sociodemographic Predictors**:
  - Age group (18-34, 35-49, 50-64, 65 and over)
  - Sex (male or female)
  - Race/ethnicity (Hispanic, non-Hispanic Black, non-Hispanic White, non-Hispanic Other or Multiple Race)
  - Education (high school or less, greater than high school, missing/unknown)
  - Marital status (married or cohabitating; never married, widowed, divorced, or separated; missing)
  - Insurance status (public insurance, private insurance, uninsured, missing/unknown)
  - Ratio of income to poverty threshold (above poverty, below poverty, missing/unknown)

Methodology

- **Health Outcomes** (Measured* and Self-Reported)**:**
  1. Diabetes
  2. Hypertension
  3. High cholesterol
  4. Current smoking

* Measured health outcomes include self-reported prescription medication use, where respondents are requested to show the medication label to the interviewer

Methodology

1. Diabetes:
   - *Measured*: hemoglobin A1c (HbA1c) ≥6.5% or reported use of prescription medication for diabetes
   - *Self-report*: “have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?”

Methodology

2. Hypertension:
   - *Measured*: average blood pressure across up to three measurements ≥140 mmHg (systolic), ≥90 mmHg (diastolic), or reported use of prescription medication for hypertension
   - *Self-report*: “have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?”

Methodology

3. High cholesterol
   - *Measured*: total cholesterol of ≥240 mg/dL or reported use of cholesterol-lowering prescription medication
   - *Self-report*: “have you ever been told by a doctor or other health professional that your blood cholesterol level was high?”

Methodology

4. Current smoking
   - *Measured*: serum cotinine ≥11 ng/mL
   - *Self-report*: “have you ever smoked at least 100 cigarettes in your entire life?” and a response of “every day” to “do you now smoke cigarettes?”

Methodology

- Sample Data Setup for RPMS*:

Respondent ID	Age, years	Measure Type	Diabetes
1	22	0 (Measured)	1 (Yes)
1	22	1 (Self-reported)	1 (Yes)
2	64	0 (Measured)	1 (Yes)
2	64	1 (Self-reported)	0 (No)

*Recursive Partitioning for Modeling Survey Data (Toth, 2022)

Data includes additional predictors (not shown above): sex, race/ethnicity, education level, marital status, insurance coverage type, and poverty status

Results

- The CLRT models estimate differences in measured prevalence (β_0) and measurement error (β_1)
- β_1 is the number of percentage points, on average, that a subgroup tends to over– or under–report the specified health outcome relative to the measured value
  - e.g., β_1 = -1.0 indicates average under self-reporting by 1 percentage point for the given subgroup
- End nodes with significant measurement error were identified
  - $\beta_1 \neq 0$ at $p < 0.001$

Results

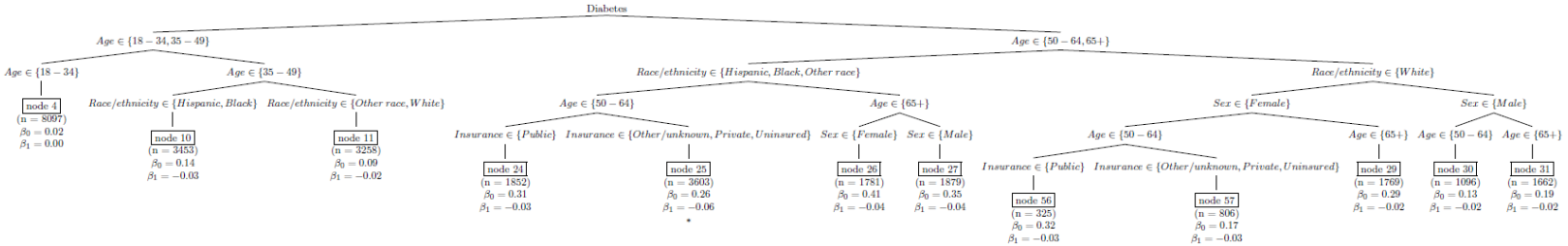
	Diabetes	Hypertension	High Cholesterol	Current Smoking
Measurement error direction	Underreport only	Underreport & overreport	Overreport only	Underreport only
Range in magnitude (%)	-0.4 to -6.4 Range: ~6%	-9.5 to 4.3 Range: ~14%	0.6 to 8.0 Range: ~7%	1.3 to 25.3 Range: ~24%
Number of significant end nodes	1	4	5	17
Modification variables	Age Sex Race/ethnicity Insurance	Age Sex Race/ethnicity Education Insurance	Age Sex Education Insurance Poverty	Age Sex Race/ethnicity Education Marital status Insurance Poverty

Results

- **Diabetes**:
  - Initial splits: age, race/ethnicity
  - Consistent underreporting across all subgroups
  - Significant **underreporting** by 6.4% for respondents ages 50-64 years who were Hispanic, non-Hispanic Black, or non-Hispanic other, and did not have public health insurance (node 25)

Results

Figure 1. Diabetes Tree Model

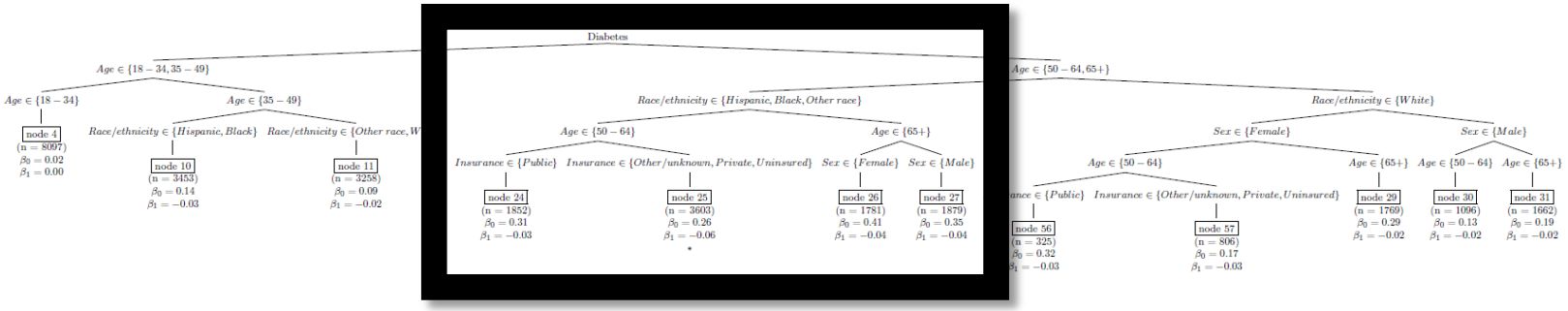


* indicates β_1 significantly different from 0, $p < 0.001$

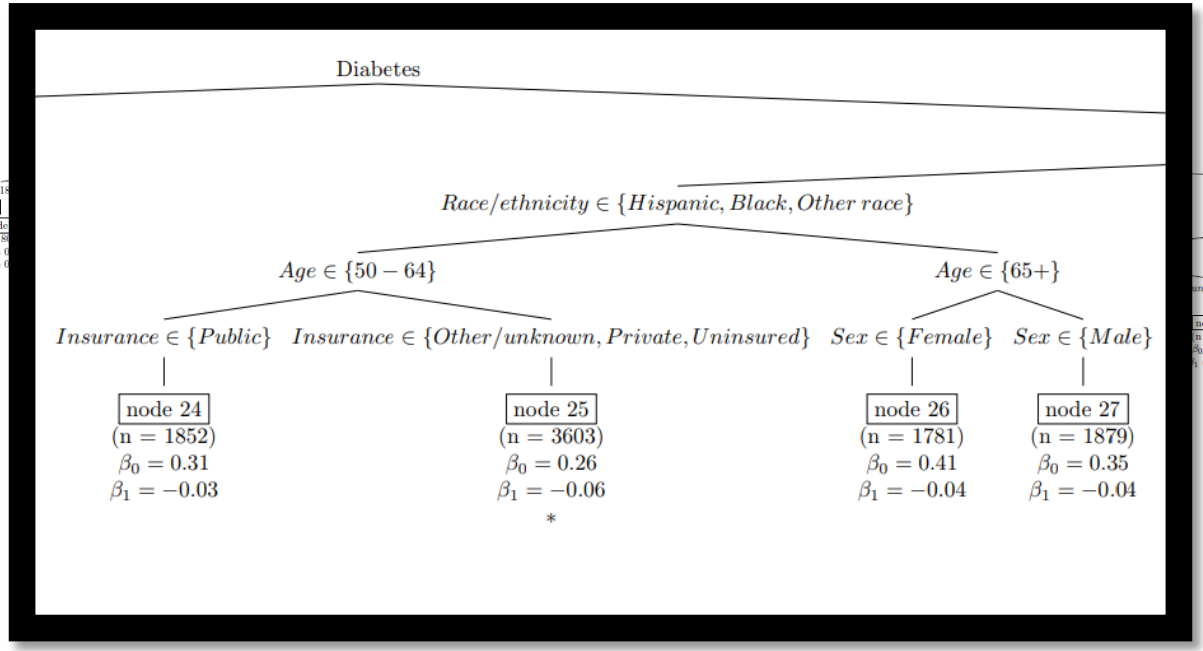
β_0 and β_1 are shown as proportions

Results

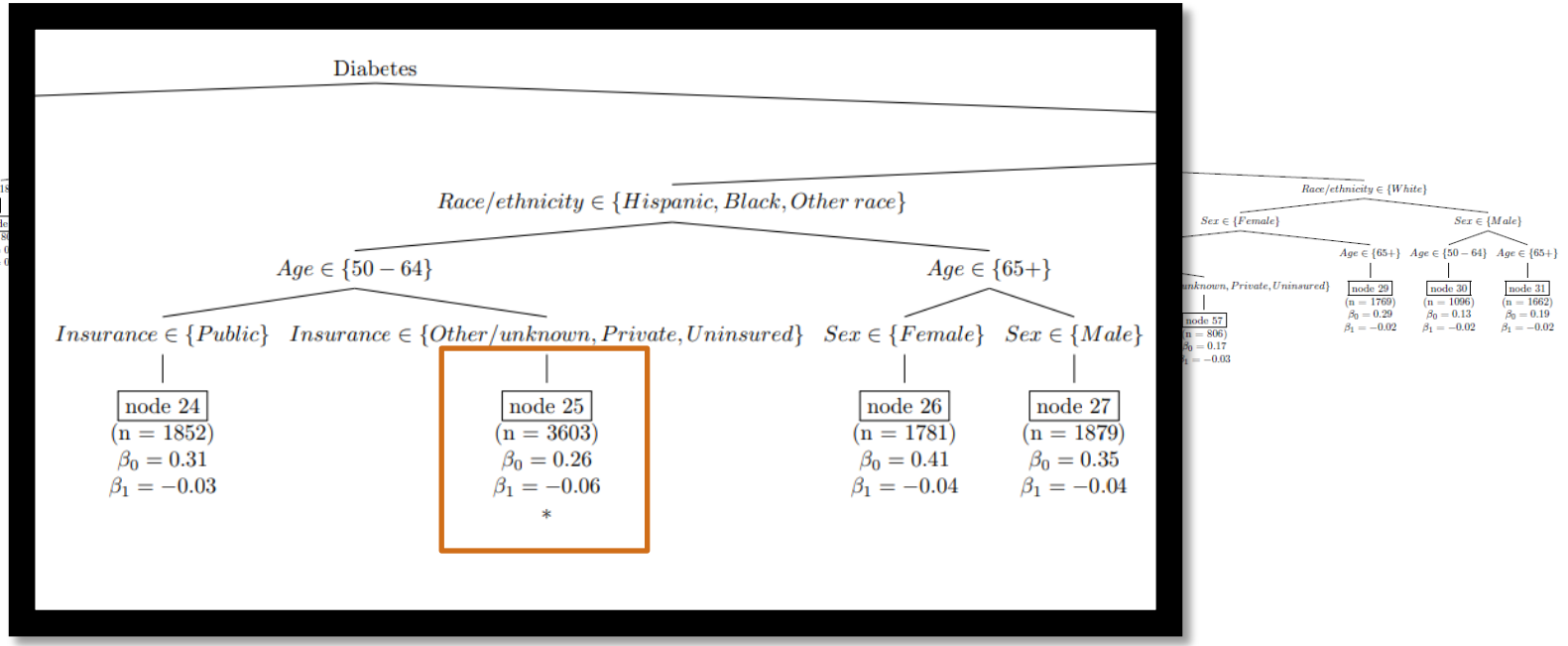
Figure 1. Diabetes Tree Model



Results


Diabetes
Race/ethnicity ∈ {Hispanic, Black, Other race}
Age ∈ {50 − 64}
Age ∈ {65+}
Insurance ∈ {Public}
Insurance ∈ {Other/unknown, Private, Uninsured}
Sex ∈ {Female}
Sex ∈ {Male}
node 24
(n = 1852)
β0 = 0.31
β1 = −0.03
node 25
(n = 3603)
β0 = 0.26
β1 = −0.06
node 26
(n = 1781)
β0 = 0.41
β1 = −0.04
node 27
(n = 1879)
β0 = 0.35
β1 = −0.04

Results

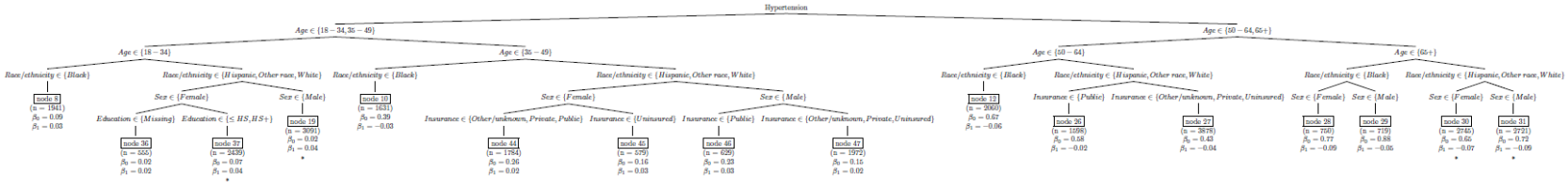

Diabetes
Race/ethnicity ∈ {Hispanic, Black, Other race}
Age ∈ {50 − 64}
Age ∈ {65+}
Insurance ∈ {Public}
Insurance ∈ {Other/unknown, Private, Uninsured}
Sex ∈ {Female}
Sex ∈ {Male}
node 24
(n = 1852)
β0 = 0.31
β1 = −0.03
node 25
(n = 3603)
β0 = 0.26
β1 = −0.06
*
node 26
(n = 1781)
β0 = 0.41
β1 = −0.04
node 27
(n = 1879)
β0 = 0.35
β1 = −0.04

Results

- **Hypertension**:
  - Initial splits: age, race/ethnicity
  - Younger respondents had lower prevalence and tended to overreport, while older respondents had higher prevalence and tended to underreport
  - Significant **underreporting** for respondents ages 65+ who were Hispanic, non-Hispanic White, or non-Hispanic other:
    - Males 8.8% (node 31); Females: 7.1% (node 30)
  - Significant **overreporting** for respondents ages 18-34 who were Hispanic, non-Hispanic White or non-Hispanic other:
    - Males: 3.9% (node 19); Females with non-missing education status: 4.3% (node 37)

Results

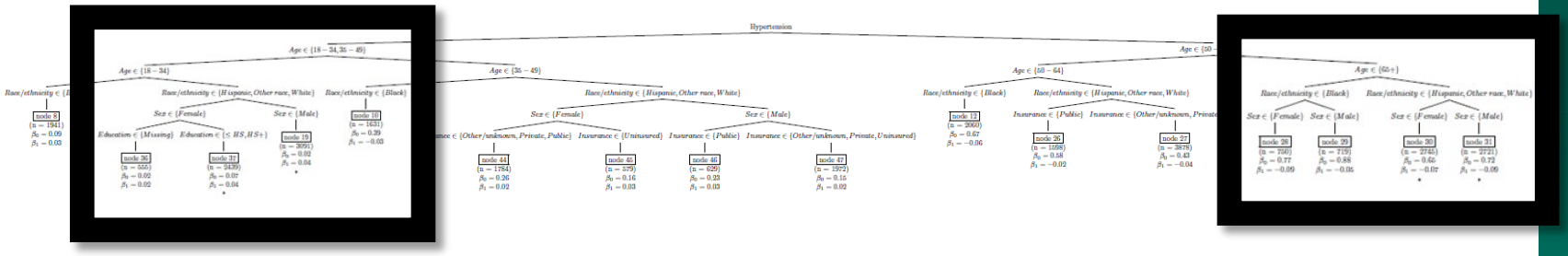
Figure 2. Hypertension Tree Model



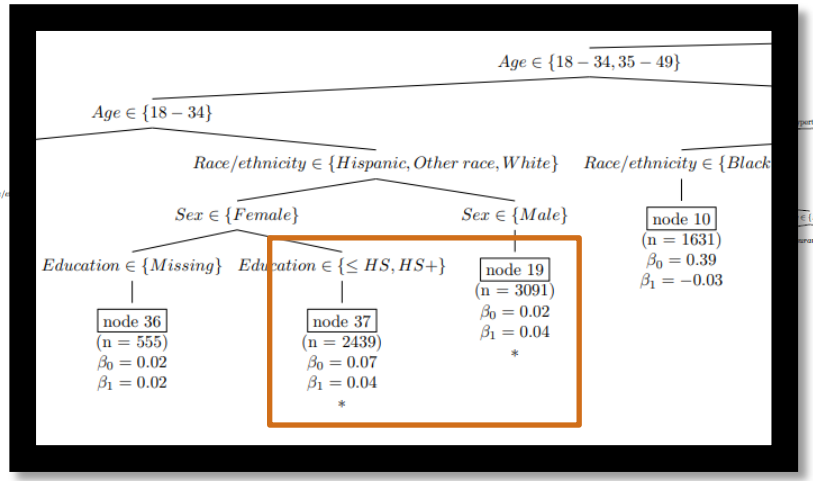
* indicates β_1 significantly different from 0, $p < 0.001$
β_0 and β_1 are shown as proportions

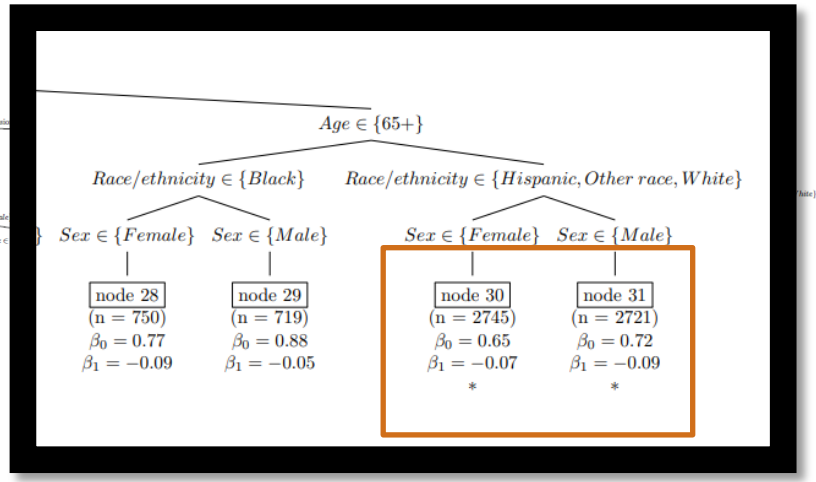
Results

Figure 2. Hypertension Tree Model



Results


Age ∈ {18 − 34, 35 − 49}
Age ∈ {18 − 34}
Race/ethnicity ∈ {Hispanic, Other race, White}
Race/ethnicity ∈ {Black
Sex ∈ {Female}
Sex ∈ {Male}
node 10
(n = 1631)
β0 = 0.39
β1 = −0.03
Education ∈ {Missing}
Education ∈ {≤ HS, HS+}
node 19
(n = 3091)
β0 = 0.02
β1 = 0.04
*
node 36
(n = 555)
β0 = 0.02
β1 = 0.02
node 37
(n = 2439)
β0 = 0.07
β1 = 0.04
*

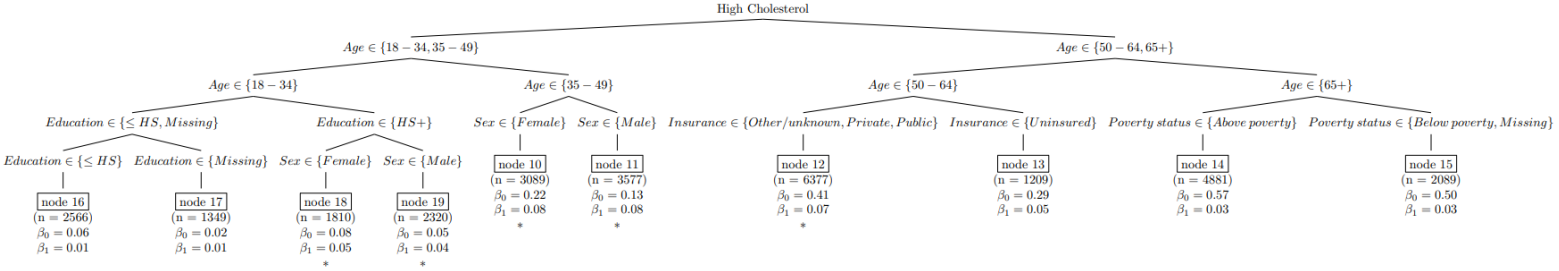

Age ∈ {65+}
Race/ethnicity ∈ {Black}
Race/ethnicity ∈ {Hispanic, Other race, White}
Sex ∈ {Female}
Sex ∈ {Male}
Sex ∈ {Female}
Sex ∈ {Male}
node 28
(n = 750)
β0 = 0.77
β1 = −0.09
node 29
(n = 719)
β0 = 0.88
β1 = −0.05
node 30
(n = 2745)
β0 = 0.65
β1 = −0.07
*
node 31
(n = 2721)
β0 = 0.72
β1 = −0.09
*

Results

- **High Cholesterol**:
  - Initial split: age
  - Consistent overreporting across all subgroups
  - Significant **overreporting** for respondents:
    - Ages 18-34 with greater than a high school education by 5.4% for females (node 18) and 4.0% for males (node 19)
    - Ages 35-49 by 7.6% for females (node 10) and 8.0% for males (node 11)
    - Ages 50-64 with health insurance or an unknown status by 6.8% (node 12)

Results

Figure 3. High Cholesterol Tree Model

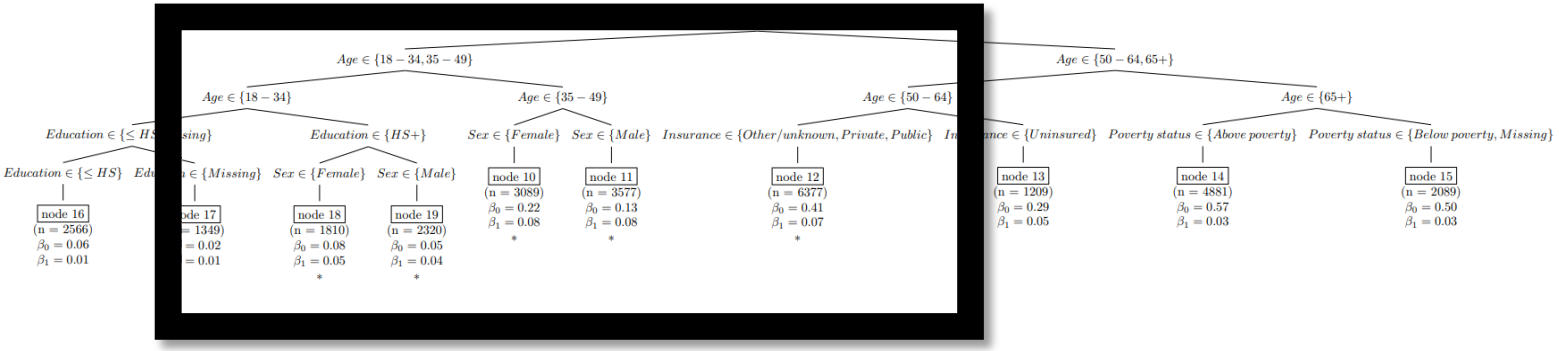


* indicates β_1 significantly different from 0, $p < 0.001$

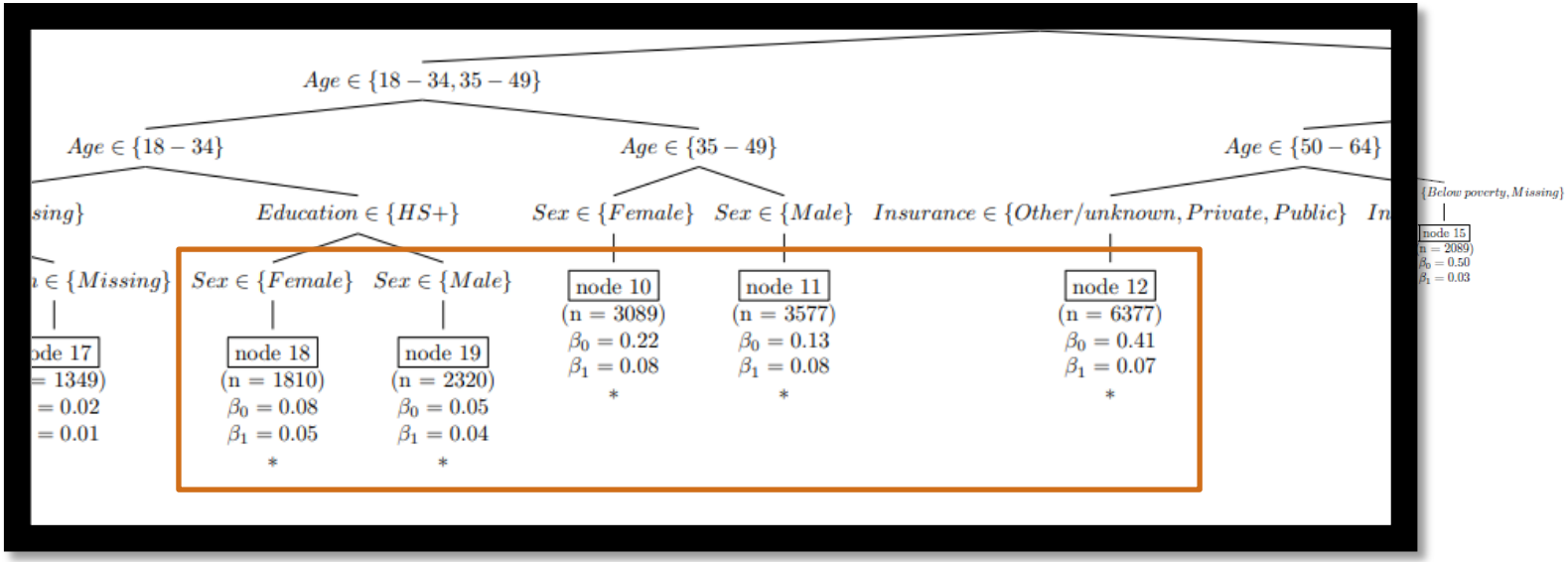
β_0 and β_1 are shown as proportions

Results

Figure 3. High Cholesterol Tree Model



Results

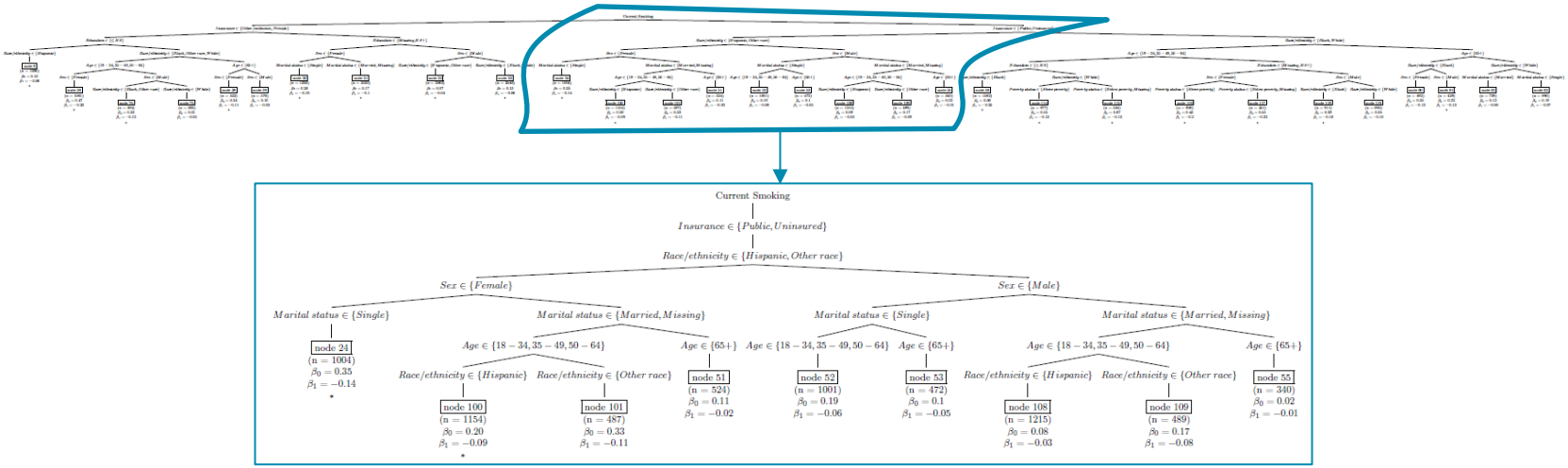

$Age \in \{18-34, 35-49\}$
$Age \in \{18-34\}$
$Age \in \{35-49\}$
$Age \in \{50-64\}$
$sing\}$
$Education \in \{HS+\}$
$Sex \in \{Female\}$
$Sex \in \{Male\}$
$Insurance \in \{Other/unknown, Private, Public\}$
In
$\{Below\ poverty, Missing\}$
$\in \{Missing\}$
$Sex \in \{Female\}$
$Sex \in \{Male\}$
ode 17
= 1349)
= 0.02
= 0.01
node 18
(n = 1810)
$\beta_0 = 0.08$
$\beta_1 = 0.05$
*
node 19
(n = 2320)
$\beta_0 = 0.05$
$\beta_1 = 0.04$
*
node 10
(n = 3089)
$\beta_0 = 0.22$
$\beta_1 = 0.08$
*
node 11
(n = 3577)
$\beta_0 = 0.13$
$\beta_1 = 0.08$
*
node 12
(n = 6377)
$\beta_0 = 0.41$
$\beta_1 = 0.07$
*
node 15
n = 2089)
$\beta_0 = 0.50$
$\beta_1 = 0.03$

Results

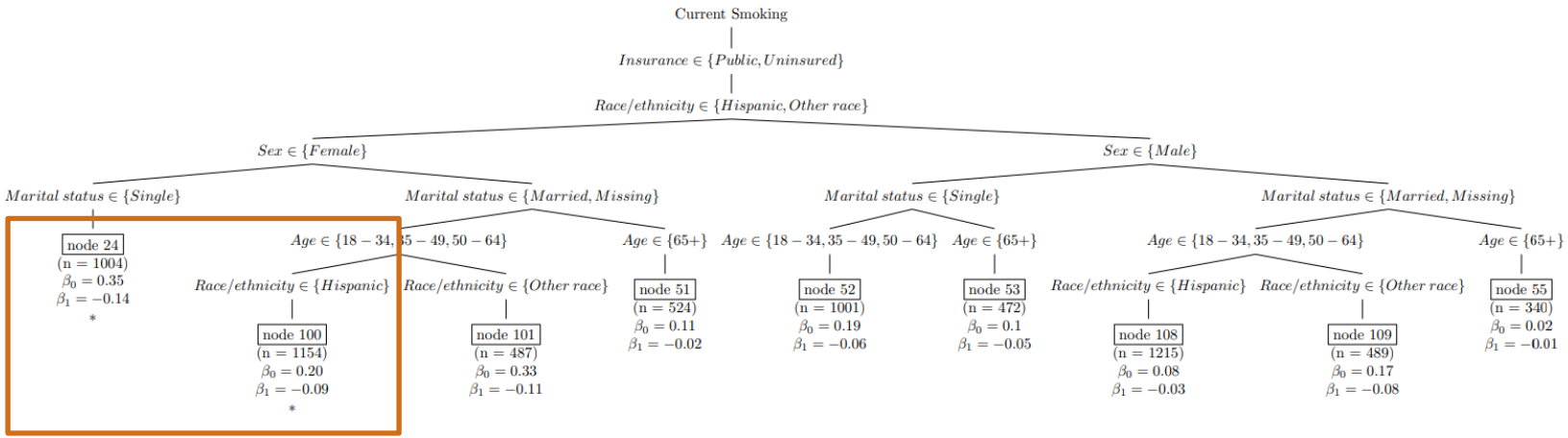
- **Current Smoking**:
  - Initially split based on health insurance type, but *all* variables were modifiers
  - Consistent underreporting across all subgroups, and significant **underreporting** for 17 out of 30 end nodes:
    - Among almost all respondents with private or other/unknown health insurance
    - Among almost all subgroups under 65 years with public health insurance or who were uninsured, and either non-Hispanic Black or non-Hispanic White
  - Percent difference between self-reported vs. measured smoking status indicates underestimation by a relative 15% to 72%

Results

Figure 4a. Current Smoking Tree Model: Public insurance or uninsured, Hispanic or non-Hispanic Other race

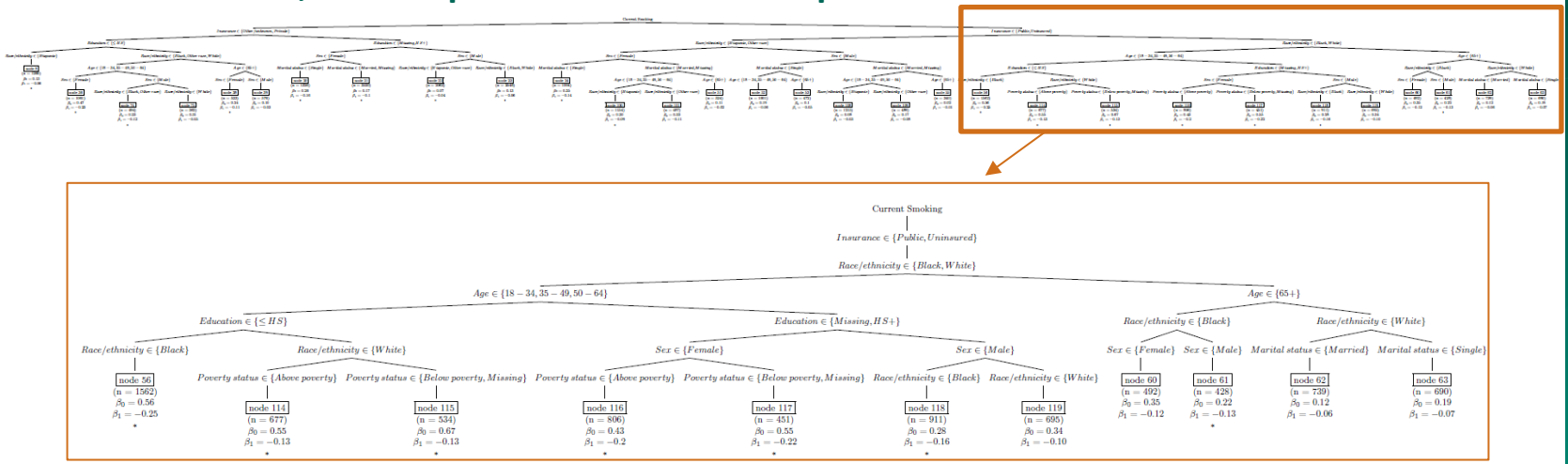


Results

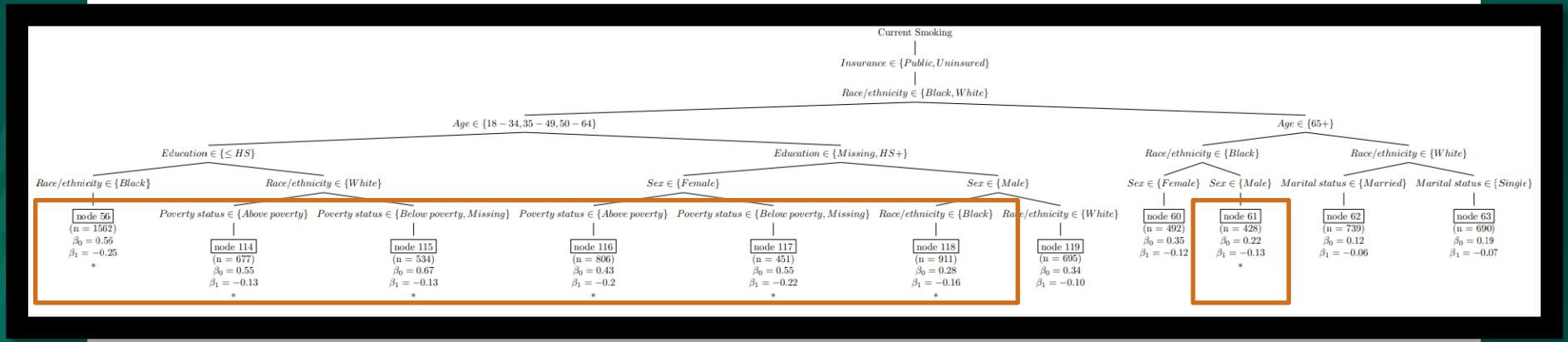

Current Smoking
Insurance ∈ {Public, Uninsured}
Race/ethnicity ∈ {Hispanic, Other race}
Sex ∈ {Female}
Sex ∈ {Male}
Marital status ∈ {Single}
Marital status ∈ {Married, Missing}
Marital status ∈ {Single}
Marital status ∈ {Married, Missing}
node 24
(n = 1004)
β0 = 0.35
β1 = −0.14
*
Age ∈ {18 − 34, 35 − 49, 50 − 64}
Age ∈ {65+}
Age ∈ {18 − 34, 35 − 49, 50 − 64}
Age ∈ {65+}
Age ∈ {18 − 34, 35 − 49, 50 − 64}
Age ∈ {65+}
Race/ethnicity ∈ {Hispanic}
Race/ethnicity ∈ {Other race}
node 51
(n = 524)
β0 = 0.11
β1 = −0.02
node 52
(n = 1001)
β0 = 0.19
β1 = −0.06
node 53
(n = 472)
β0 = 0.1
β1 = −0.05
Race/ethnicity ∈ {Hispanic}
Race/ethnicity ∈ {Other race}
node 55
(n = 340)
β0 = 0.02
β1 = −0.01
node 100
(n = 1154)
β0 = 0.20
β1 = −0.09
*
node 101
(n = 487)
β0 = 0.33
β1 = −0.11
node 108
(n = 1215)
β0 = 0.08
β1 = −0.03
node 109
(n = 489)
β0 = 0.17
β1 = −0.08

Results

Figure 4b. Current Smoking Tree Model: Public health insurance or uninsured, non-Hispanic Black or non-Hispanic White

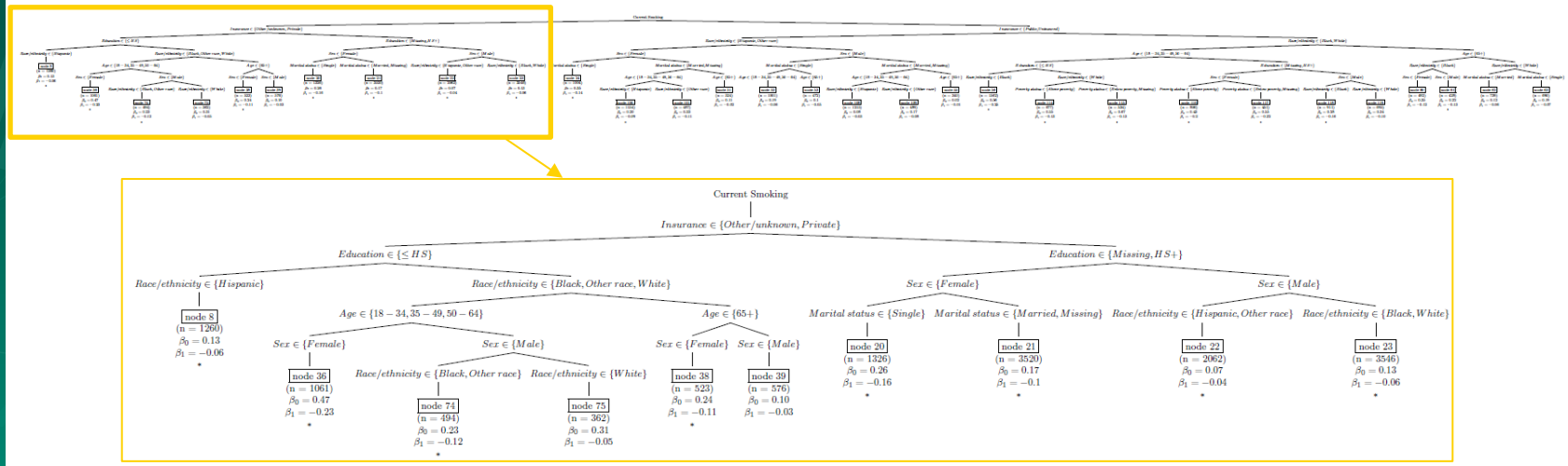


Results

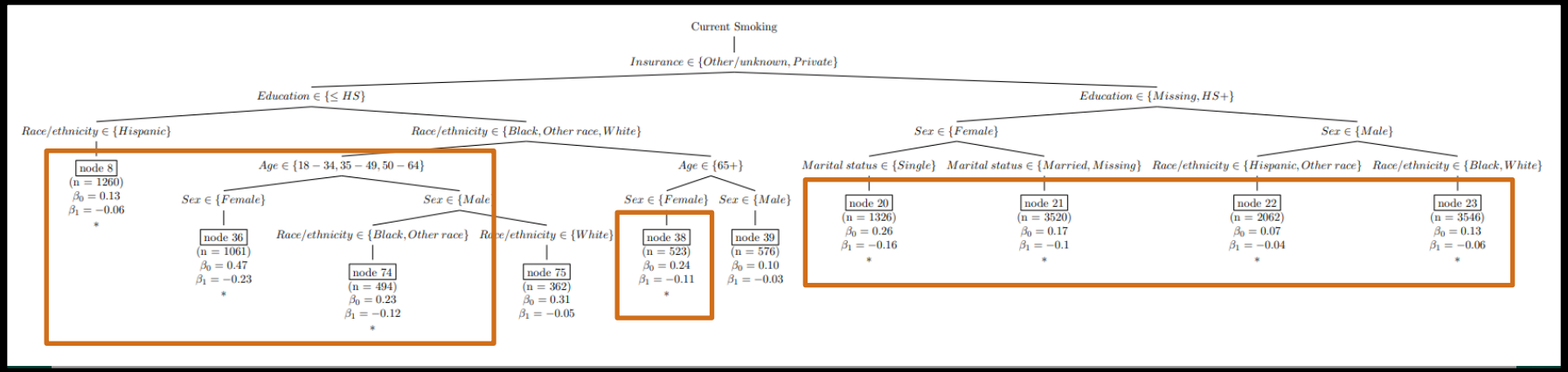

Current Smoking
Insurance ∈ {Public, Uninsured}
Race/ethnicity ∈ {Black, White}
Age ∈ {18 − 34, 35 − 49, 50 − 64}
Age ∈ {65+}
Education ∈ {≤ HS}
Education ∈ {Missing, HS+}
Race/ethnicity ∈ {Black}
Race/ethnicity ∈ {White}
Sex ∈ {Female}
Sex ∈ {Male}
Race/ethnicity ∈ {Black}
Race/ethnicity ∈ {White}
Sex ∈ {Female}
Sex ∈ {Male}
Marital status ∈ {Married}
Marital status ∈ {Single}
Poverty status ∈ {Above poverty}
Poverty status ∈ {Below poverty, Missing}
Poverty status ∈ {Above poverty}
Poverty status ∈ {Below poverty, Missing}
Race/ethnicity ∈ {Black}
Race/ethnicity ∈ {White}
node 56
(n = 1562)
β0 = 0.56
β1 = −0.25
*
node 114
(n = 677)
β0 = 0.55
β1 = −0.13
*
node 115
(n = 534)
β0 = 0.67
β1 = −0.13
*
node 116
(n = 806)
β0 = 0.43
β1 = −0.2
*
node 117
(n = 451)
β0 = 0.55
β1 = −0.22
*
node 118
(n = 911)
β0 = 0.28
β1 = −0.16
*
node 119
(n = 695)
β0 = 0.34
β1 = −0.10
node 60
(n = 492)
β0 = 0.35
β1 = −0.12
node 61
(n = 428)
β0 = 0.22
β1 = −0.13
*
node 62
(n = 739)
β0 = 0.12
β1 = −0.06
node 63
(n = 690)
β0 = 0.19
β1 = −0.07

Results

Figure 4c. Private health insurance or other/unknown health insurance



Results


Current Smoking
Insurance ∈ {Other/unknown, Private}
Education ∈ {≤ HS}
Education ∈ {Missing, HS+}
Race/ethnicity ∈ {Hispanic}
Race/ethnicity ∈ {Black, Other race, White}
Sex ∈ {Female}
Sex ∈ {Male}
node 8
(n = 1260)
β0 = 0.13
β1 = −0.06
Age ∈ {18 − 34, 35 − 49, 50 − 64}
Age ∈ {65+}
Sex ∈ {Female}
Sex ∈ {Male
Sex ∈ {Female}
Sex ∈ {Male}
node 36
(n = 1061)
β0 = 0.47
β1 = −0.23
Race/ethnicity ∈ {Black, Other race}
Race/ethnicity ∈ {White}
node 38
(n = 523)
β0 = 0.24
β1 = −0.11
node 39
(n = 576)
β0 = 0.10
β1 = −0.03
node 74
(n = 494)
β0 = 0.23
β1 = −0.12
node 75
(n = 362)
β0 = 0.31
β1 = −0.05
Marital status ∈ {Single}
Marital status ∈ {Married, Missing}
Race/ethnicity ∈ {Hispanic, Other race}
Race/ethnicity ∈ {Black, White}
node 20
(n = 1326)
β0 = 0.26
β1 = −0.16
node 21
(n = 3520)
β0 = 0.17
β1 = −0.1
node 22
(n = 2062)
β0 = 0.07
β1 = −0.04
node 23
(n = 3546)
β0 = 0.13
β1 = −0.06

Discussion

- The magnitude and direction of measurement error varied by health outcome and subgroup
  - Smaller and simpler trees for diabetes & high cholesterol (less variation)
  - Larger and more complex trees for hypertension & smoking (more variation)
- Measurement error varied by age, sex, race/ethnicity, education level, and health insurance type for most health outcomes
  - Marital status and poverty level were less important
- Regression trees can highlight where we are more likely to under- or over-estimate prevalence when relying on self-reported data

Discussion

- Underreporting of smoking suggests a narrow interpretation of the self-report definition, missing some cases of current tobacco or nicotine use
- Questions that include alternative tobacco/nicotine products could better capture usage, especially for younger populations
  - e.g., NHANES includes additional questions that ask about ever use of cigars, e-cigarettes, and smokeless tobacco
- Sole use of exclusive definitions may result in systematic misreporting across groups

Conclusion

- Quantifying the degree of measurement error inequity and identifying strategies to reduce it is critical for the accurate assessment of subgroup disparities in health outcomes
- Our analytic approach offers detailed picture of how multiple factors may interact and how measurement error differs across intersectional social, demographic, economic, and health-related dimensions
  - Necessary first step in remediating health inequities

Contact Us

Morgan Earp
mearp@cdc.gov

Lauren Rossen
lrossen@cdc.gov

Sarah Forrest
sforrest@cdc.gov

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 www.cdc.gov

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.