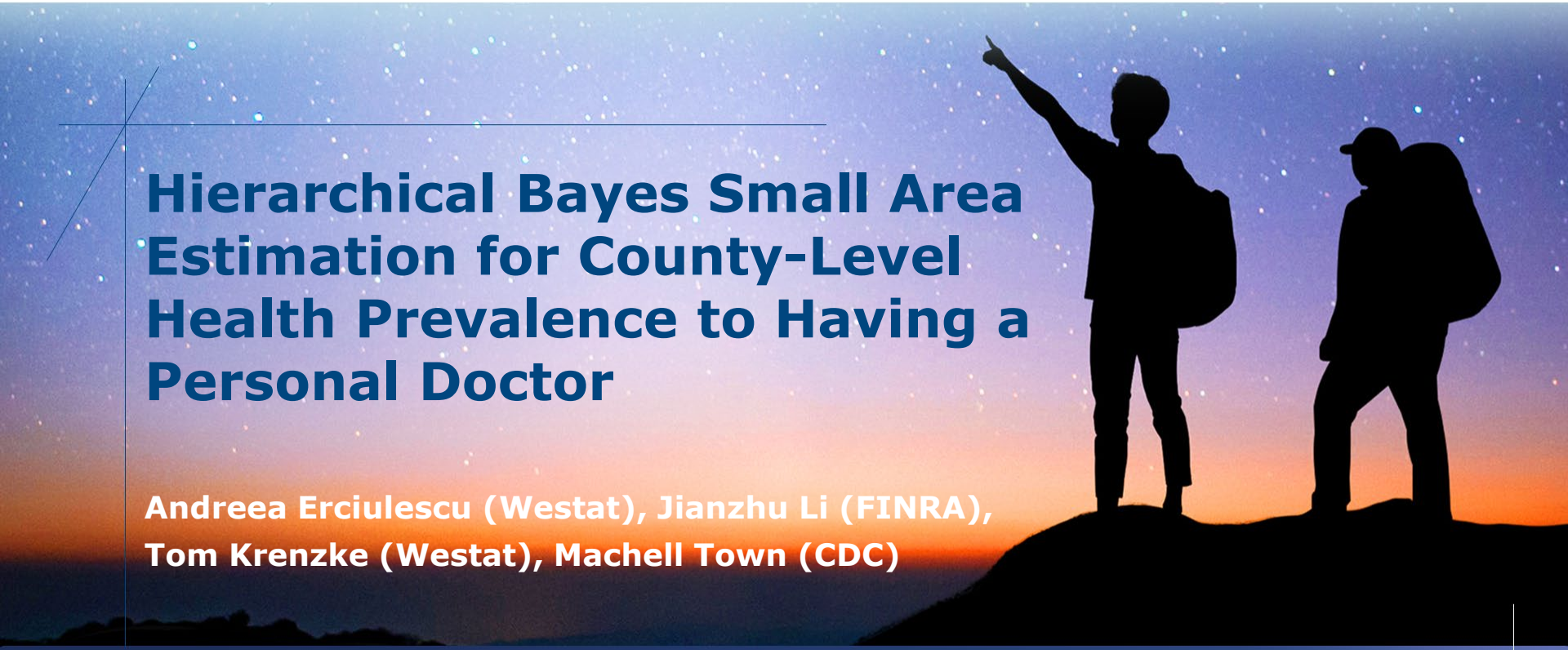


Hierarchical Bayes Small Area Estimation for County-Level Health Prevalence to Having a Personal Doctor

Andreea Erciulescu (Westat), Jianzhu Li (FINRA), Tom Krenzke (Westat), Machell Town (CDC)

WESTAT @ FCSM 2024 – The Relevance, Timeliness, and Integrity of Federal Statistics

Health prevalence estimation | Motivation

- **Agency** | U.S. Centers for Disease Control and Prevention
- **Surveys** | Behavioral Risk Factor Surveillance System (BRFSS)
- **Reference time period** | 2018
- **Outcome** | Prevalence of not having a personal doctor or health care provider
- **Domain** | County

Domain | County

- High uncertainty for small sample sizes → SAE
- 3,142 counties
  - 3,114 counties with sample (n~400,000)
  - 213 counties with sample size of 500 or more
    - Selected Metropolitan/Micropolitan Area Risk Trends (SMART) counties
  - Model-based estimates produced for all 3,142 counties

Select Literature | BRFSS Related

- **Cadwell et al. (2010)** | unit-level model for self-reported diabetes for counties
- **Zhang et al. (2014)** | unit-level model for chronic obstructive pulmonary disease prevalence for tracts and more
- **Perannunzi et al. (2016)** | Zhang model applied to health status and access indicators, and Berkowitz et al. (2018, 2019) – Zhang model applied to colorectal cancer screening prevalence, and mammography screening rates
- **Holt et al. (2019)** | similar to Zhang model applied to at-risk adults with chronic condition
- **Raghunathan et al. (2007), Liu et al. (2019)** | joint county-level model for BRFSS and NHIS

We introduce an area-level model to incorporate survey design effects and broaden the pool of predictor variables beyond BRFSS data

Health Prevalence Estimation

Health Prevalence Estimation | Initial Steps

- Raking adjustment to survey weights
  - Age, race/ethnicity, sex dimensions within SMART counties, and
  - Age, race/ethnicity, sex, county dimensions within state for smaller counties
  - Control totals from 2014-2018 American Community Survey data
- Arcsine square root transformation
  - Prevents negative estimates
  - Variance accounts for design effect due to unequal weights

Health Prevalence Estimation | Challenges

- **Sparse survey data**
  3,114 out of 3,142 counties with survey sample data
- **Rich auxiliary data**
  dozens of county-level potential covariates
- **Restricted range**
  for the outcome observations: [0,1]
- **Design effects**
  Variance from complex sample design

Source	Auxiliary variable
Census Bureau American Community Survey (ACS)	5-year estimates (2013–2017) of socioeconomic demographic, and housing characteristics
Census Bureau Small Area Income and Poverty Estimates (SAIPE)	Small area estimates of selected income and poverty statistics
Census Bureau Small Area Health Insurance Estimates (SAHIE)	Health insurance coverage status by selected economic and demographic characteristics
U.S. Department of Agriculture (USDA) Economic Research Service	Classification of counties into metro and non-metro (OMB subdivided counties into three metro and six non-metro categories)
Bureau of Labor Statistics (BLS) The Local Area Unemployment Statistics (LAUS) program	Monthly and annual employment, unemployment, and labor force data
Bureau of Economic Analysis (BEA) Centers for Disease Control and Prevention Division of Diabetes Translation (DDT)	Estimates of personal income for local areas updated statistics about diabetes
Centers for Medicare & Medicaid Services (CMS) Geographic variation public use file	Utilization and quality of health care services for the Medicare fee-for-service population
The Internal Revenue Service The Statistics of Income (SOI) Data	Income and tax data such as number of tax returns, returns with unemployment compensation, returns with taxable social Security benefit, adjusted gross personal income, personal unemployment compensation amount, and personal taxable social Security benefit amount, etc.
Health Resources and Services Administration Area Health Resources Files (AHRF)	Data on health care professions, health facilities, population characteristics, economics,health professions training, hospital utilization, hospital expenditures, and environment
National Center for Health Statistics (NCHS)	Birth rate and infant mortality rate
Federal Bureau of Investigation (FBI)	Crime rates
National Highway Traffic Safety Administration (NHTSA)	Traffic fatalities
U.S. Energy Information Administration (EIA)	Energy consumption height

Health Prevalence Estimation | Solution

- Area-level univariate linear (on **arsine-square-root** scale) **three-fold** model
- Hierarchical Bayes inference
- **Multi-stage variable selection framework**
  - Most complete and less prone to error information
  - Strong associations between outcome and covariates
  - Weak associations between covariates
  - Predictive covariates: least absolute shrinkage and selection operator (LASSO) and cross-validation

Models

Model Structure

Sampling level: $y_{ijk} \| (\theta_{ijk}, \sigma^2_{ijk}) \sim N(\theta_{ijk}, \sigma^2_{ijk})$

Linking level:
$$\theta_{ijk} \| (\beta, c_{ijk}, s_{jk}, d_k) = x_{ijk}\beta + c_{ijk} + s_{ij} + d_i,$$
$$c_{ijk} \| (\sigma^2_c) \sim N(0, \sigma^2_c),$$
$$s_{ij} \| (\sigma^2_s) \sim N(0, \sigma^2_s),$$
$$d_i \| (\sigma^2_d) \sim N(0, \sigma^2_d),$$

Priors:
$$\beta \sim N(0, 10^4), \text{ component-wise},$$
$$(\sigma_c, \sigma_s, \sigma_d) \sim \text{Cauchy } (0, 5), \text{ component-wise}$$

Model Fit

- Four models evaluated
  - Based on 3,114 counties
    - Full (9 cov) and reduced models (5 cov)
  - Based on 213 counties (SMART counties)
    - Full (9 cov) and reduced models (6 cov)
- 3 chains – each chain has 20,000 samples, with 5,000 burn-in, and thinning every tenth iteration, resulting in 4,500 samples for inference – R STAN

Prediction

- Composite of survey estimates and model-synthetic predictions
  - Model estimates for SMART counties are closer to survey estimates
  - Model estimates for smaller counties are smoothed more than others toward a prediction based on linear relationship between the survey estimates and covariates
- Not-in-sample counties
  - Relies more on the model structure and predictions
    - Contribution of survey data comes through the random effects
- Back transform – for each of the 4,500 samples
- Aggregation – to state, division, and national levels, using population totals
- Cross-validation – two groups

Results

Variable Selection Results

Variable	Source	All counties with sample		SMART counties	
		Full	Reduced	Full	Reduced
Prop. HS diploma	ACS	X	X		
Prop. Hispanic	ACS	X	X	X	X
Prop. Non-Hisp White	ACS	X	X	X	X
Prop. Native Americans	ACS	X			
Prop. Owner occupied	ACS	X	X		
Prop. In different house	ACS	X			
Prop. In different state	ACS	X		X	
Prop. Uninsured	SAHIE	X	X	X	X
Prop. Returns with taxable	SOI	X		X	
Prop. 55-64 years old	ACS			X	X
Birth rate	NCHS			X	X
Per capita energy consumption	EIA			X	
Traffic fatalities per 100 miles	NHTSA			X	X

Other Results

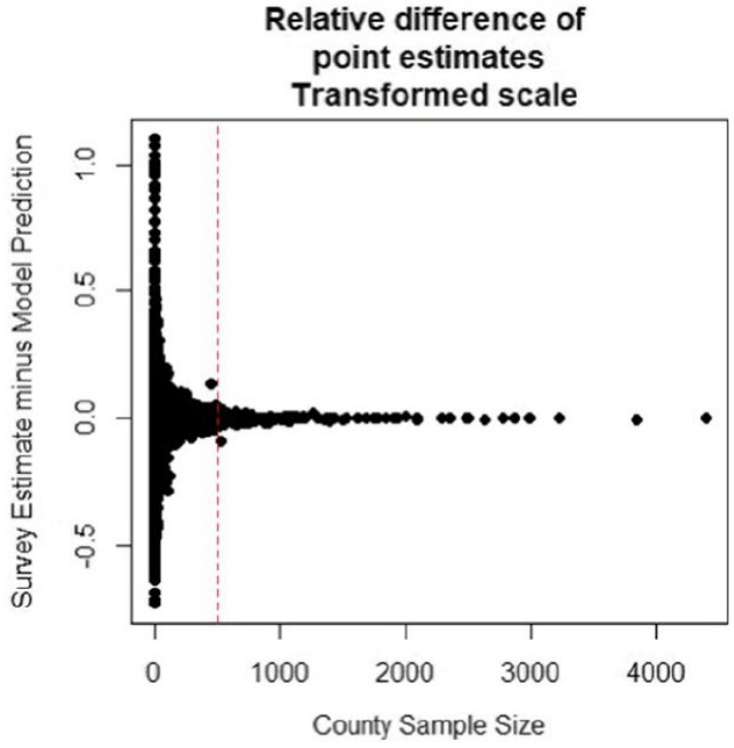
Cross-Validation

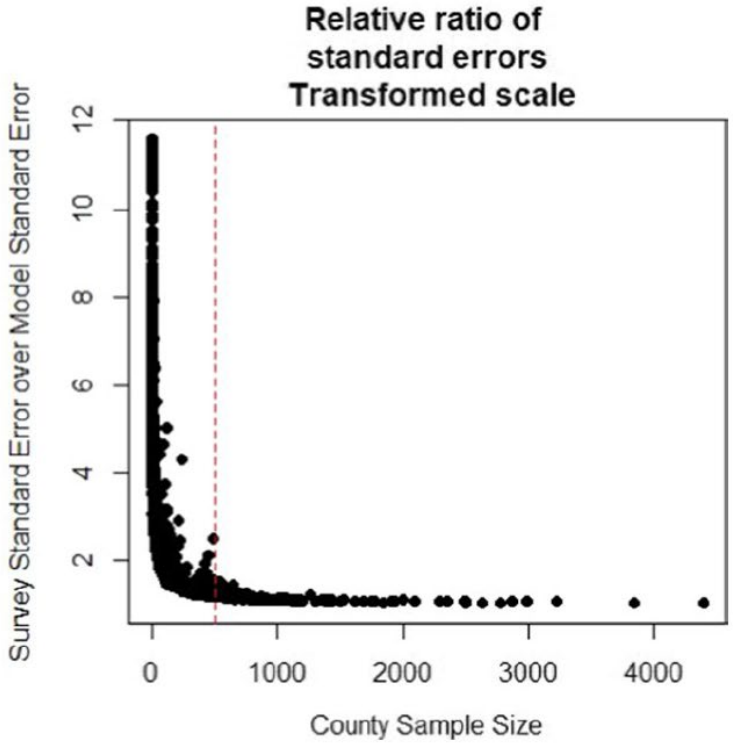
County set	All counties with sample		SMART counties	
	Full	Reduced	Full	Reduced
SMART counties (213)	0.350	0.416	0.409	0.483

WAIC

County set	All counties with sample		SMART counties	
	Full	Reduced	Full	Reduced
SMART counties (213)	-0.7198	-0.7177	-2.0346	-2.0117

External Validation Checks on Final Model



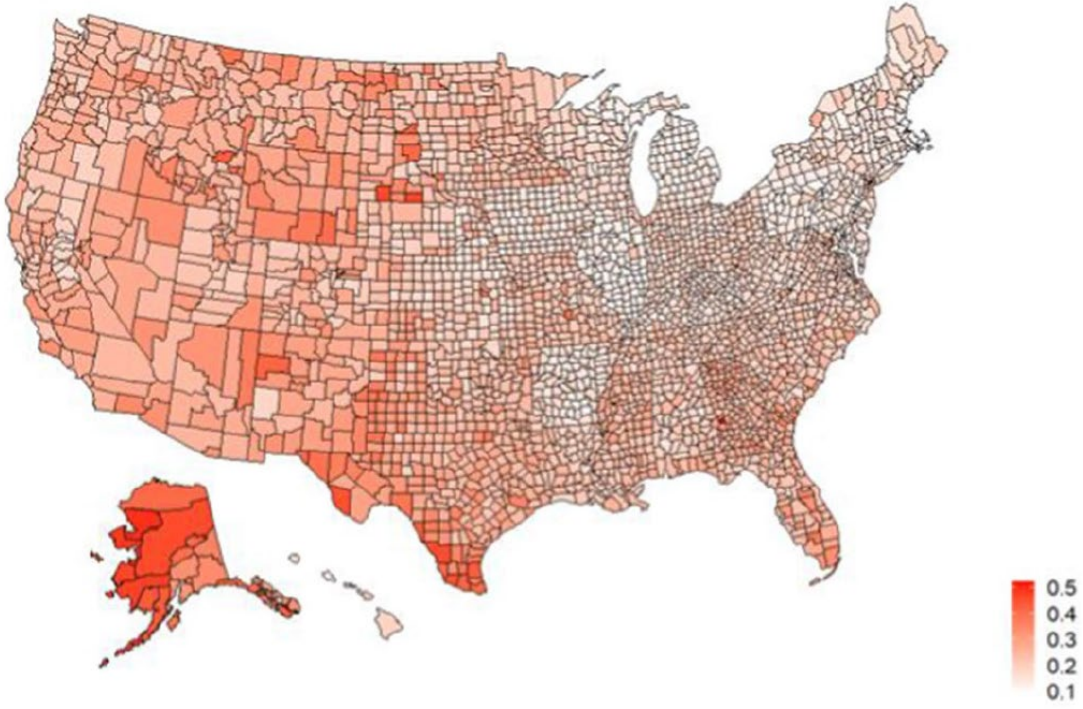


Discussion

Discussion

- Combined BRFSS survey data with data from other sources via area-level models
- Accounted for error in survey data and nested structure of the data
- More accurate point estimates when model is based on 3,114 counties vs data from much larger sample sizes from fewer (213) counties
- Investigated other transformations, but oddities existed
- Considered a hybrid model (sampling level at unit-level) – computationally intensive, some model misspecification
- Future investigations (Bayesian LASSO, multivariate models, measurement error specifications)

Thank you (tomkrenzke@westat.com)



Model-based estimates of not having a personal care provider or doctor