A. Small Area Estimates of County-Level Obesity Prevalence

Inputs

National Survey of Children's Health (NSCH) individual-level data from the 2007 and 2011-2012 survey waves were pooled to increase total sample size. Because post-stratification (see below) requires the cross-classification of individuals in age x sex x race strata, we elected to coarsely categorize NSCH into two age strata (10-14 and 15-17), and two race/ethnic groups (white and non-white). County Federal Information Processing Standard (FIPS) codes from the restricted access data, available in collaboration with the Research Data Center of the Centers for Disease Control and Prevention, were used to identify multiple respondents from the same county and to merge individual-level respondents with county-level variables.

For post-stratification, the population at risk for obesity was defined as the number of adolescents in each county in the eight age *x* sex *x* race strata. These denominator counts were obtained from the 2008-2012 American Community Survey (ACS) of the U.S. Census Bureau. County-level poverty rate was also abstracted from the 2008-2010 ACS.

Because we extended the Zhang approach by incorporating information from geographically contiguous counties in estimation (see below), we restricted to 3,109 counties in the contiguous U.S. plus the District of Columbia. This improves the estimation of sparse-count counties through spatial interpolation based on geographic contiguity.

Multilevel Model

To recover sources of local variation in obesity prevalence we fit a 3-level mixed effects model.[1,2] Individuals are nested within counties which are nested within states. We estimated fixed effects for age, sex, race, survey year, county poverty, and Census region, as well as county- and state-level random intercepts, and state-level random slopes for age, sex, and race. The following model was fit to the NSCH data using survey weights re-scaled to total sample size within each state to account for selection probabilities[3]:

Level 1 (NSCH respondent):

$$Pr\left(y_{ijkycs} = 1\right) = logit^{-1}\left(\beta_{0cs} + \beta_{1s}age_{ics} + \beta_{2s}sex_{jcs} + \beta_{3s}race_{kcs} + \ \beta_4 year_y\right)$$

Level 2 (County):

$$\beta_{0cs} = \gamma_{0s} + \gamma_1 cty\_pov_{cs} + u.county_{cs}$$

Level 3 (State):

$$\gamma_{0s} = \delta_0 + \delta_{reg} region_s + u.state_s$$

$$\beta_{1s} = \delta_{age} + age.slope_s$$

$$\beta_{2s} = \delta_{sex} + sex.slope_s$$

$$\beta_{3s} = \delta_{race} + race.slope_s$$

At level 1, we are interested in the probability that a child in age group *i*, sex group *j*, race group *k*, surveyed in year *y*, and residing in county *c*, and state *s*, is obese as defined as a BMI for age and sex greater than the 95th percentile. We estimate this probability as a function of a county-specific intercept, β_{0cs}, a fixed effect for year, β_4, and β_{1s} – β_{3s} capturing the state-specific effect of age, sex, and race (see below).

At the county level, the intercept β_{0cs} is decomposed into a state-level intercept, γ_s, a fixed effect for county poverty rate, and the county random effect, $u.county_{cs}$. The county random effect represents county-level variation around the state average conditional on other covariates.

At the state level, the state intercepts, γ_s are further decomposed into a global (national) intercept, δ_0, plus three fixed effects for census regions (referent, Northeast region), δ_{reg}, and a state-level random intercept, $u.state_s$. In addition to the random intercepts, random slopes for age, sex, and race are estimated at the state level. In each case there is a fixed effects component representing the average effect of each factor in the full national sample, as well as a random effects component representing the state-specific deviation in the importance of each variable for obesity prevalence.

All variance components (e.g., $u.county_{cs}$, $u.state_s$, $age.slope_s$, $sex.slope_s$, $race.slope_s$) are assumed to arise from a multivariate normal distribution with mean zero and covariance matrix omega as estimated in the *glmer()* function in R package **lme4**.

Spatial Interpolation

In some counties there were no NSCH respondents, and thus no county-specific random intercept was estimated. Building on Tobler's First Law of Geography,[4] we borrow statistical information from geographically proximate counties (e.g., first-order Queen contiguity) to impute missing values under the assumption that neighboring counties are more alike than non-neighboring counties on average. Specifically we imputed missing county random intercepts with the average of geographically contiguous county random intercept values. As a sensitivity analysis, we conducted internal validation after dropping these imputed values with little change in results (results not shown).

Post-Stratification

The predicted prevalence of obesity for every strata of age x sex x race x county/state can then be calculated directly from model coefficients. To summarize highly stratified prevalence in a meaningful way, we apply them to the Census-derived population at risk in each strata to estimate an overall county specific marginal prevalence which takes into account local demographic structure and geographic variation. The prevalence of obesity in each strata is calculated in this manner:

$$Pr(y_{ijkcs} = 1) = \frac{exp(\beta_{0cs} + \beta_{i1s} + \beta_{j2s} + \beta_{k3s})}{1 + exp(\beta_{0cs} + \beta_{i1s} + \beta_{j2s} + \beta_{k3s})}$$

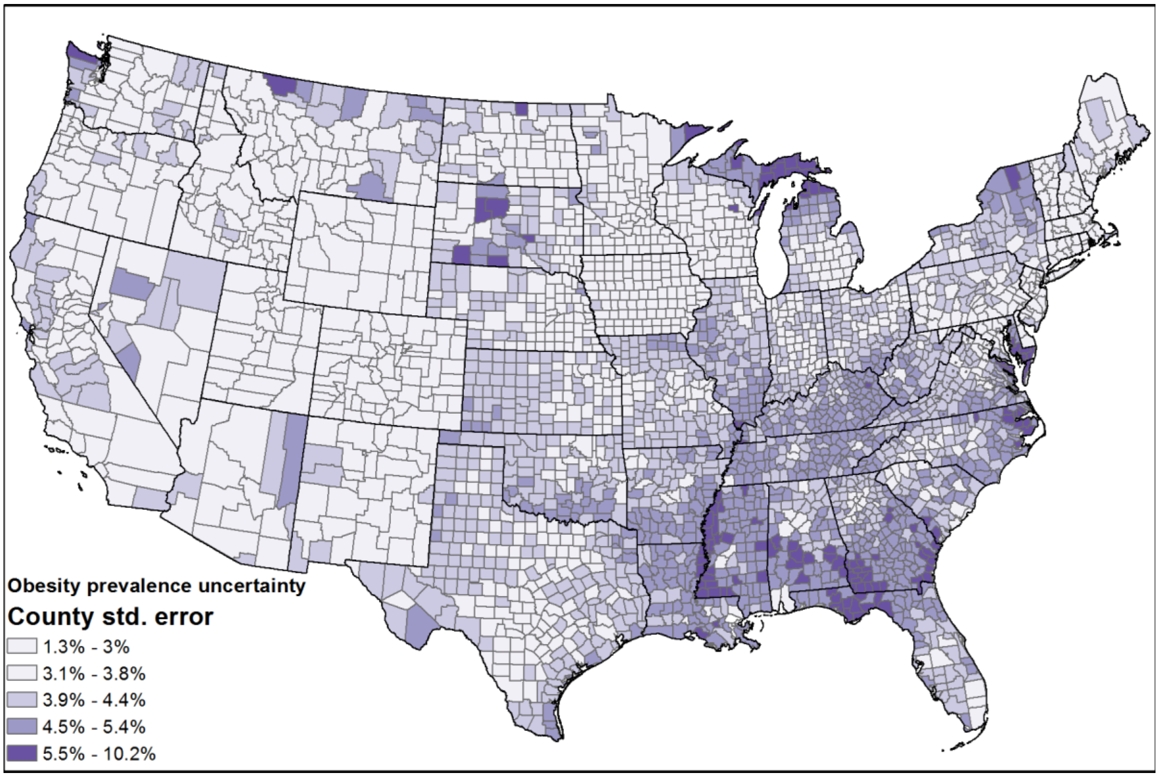
Stratum-specific prevalence is converted to county-specific prevalence in this manner by multiplying the stratum-specific prevalence by the stratum-specific population size (Pop_{ijkcs}):

$$Prev_{cs} = \frac{\sum_i \sum_j \sum_k (Pr_{ijkcs} \times Pop_{ijkcs})}{\sum_i \sum_j \sum_k Pop_{ijkcs}}$$

Estimating Obesity Prevalence Model Uncertainty

Following previous work,[5] we approximated modeled county-level obesity prevalence uncertainty by taking 1,000 draws from the distributions described by the model coefficients and associated SEs, and in each case recalculated the post-stratification prevalence to create a distribution of uncertainty summarized as SEs. We summarize the spatial distribution of these approximated SEs in Appendix Figure 1.

Appendix Figure 1. Approximated SEs from model-estimated obesity prevalence of children 10-17 years old in U.S. Counties, National Survey of Children's Health 2007 and 2011-2012.



Measuring Geographic Disparities

Quantification of the relative disparity or disproportionality among unordered groups such as geographic units can be summarized with measures including the Theil's Index, Mean Log Deviation, and Index of Disparity.[6,7] We carried out calculations on residuals from an empty model (intercept-only: baseline observed geographic disparity) and a fully adjusted model in order to characterize the proportion of geographic disparity which is 'explained' or accounted for by measured variables. Formulas and results for each of the three methods are summarized below.

Appendix Table 1. Measure of Geographic Disparity

Formula[6,7]	**Measure in empty and adjusted model**	**% reduction in adjusted vs empty model**
Theil's Index $TI = \sum_{c} r_c \times \ln(r_c)$	Empty: 346.7 Adjusted: 182.6	47%
Mean Log Deviation $MLD = \sum_{c} -\ln(r_c)$	Empty: 361.8 Adjusted: 200.9	44%
Index of Disparity $ID = {\left(\sum_c \lvert \varepsilon_c - \bar{\varepsilon} \rvert / n\right)} \Big/ {\bar{\varepsilon}} \times 100$	Empty: 37.8 Adjusted: 25.8	32%

$r_c = {\varepsilon_c}/{\bar{\varepsilon}}$

ε_c=county specific model residual: $\varepsilon_c = y_c - \hat{y}$ where y_c is the observed prevalence and y-hat is the model-predicted prevalence

$\bar{\varepsilon} = \frac{\sum_n \varepsilon_c}{n}$ where n is the number of counties

Measuring Geographic Clustering

Spatial autocorrelation is the degree to which like values (highs or lows) are located near one another. The Moran's I is a commonly used measure of global spatial autocorrelation akin to the Pearson correlation coefficient except that instead of correlating two separate variables, it correlates obesity prevalence in each county with the average obesity prevalence in all geographic adjacent counties. The Moran's I statistic ranges from -1 to +1, although in practice Moran's I is typically zero (no spatial clustering or autocorrelation) or positive which is suggestive of spatial autocorrelation or clustering. We calculated the Moran's I separately on the model residuals from the empty model (intercept only) and from the adjusted model, in order to understand the crude observed degree of clustering of obesity as compared to the residual clustering above and beyond that explained by included variables. The Moran's I from the empty model is:

$$I = \frac{\sum_{c=1}^{N}\sum_{c\prime=1}^{N} w_{cc\prime}\,(\varepsilon_c - \bar{\varepsilon})(\varepsilon_{c\prime} - \bar{\varepsilon})}{\sum_{c=1}^{N}\sum_{c\prime=1}^{N} w_{cc\prime}} \left(\frac{1}{\frac{1}{N}\sum_{c=1}^{N}(\varepsilon_c - \bar{\varepsilon})^2}\right)$$

Where I=Moran's I statistic, N indexes counties; $w_{cc'}$ is an n x n Queen contiguity matrix representing spatial relationships among index counties (*c*) and other counties (*c'*); ε_c is the model residual (observed adolescent obesity prevalence minus model predicted prevalence) in the c_{th} county, and $\bar{\varepsilon}$ is the mean of all county-specific residuals in the geographically contiguous counties indicated in the $w_{cc'}$ weights matrix.

B. Hierarchical Bayesian Regression Approach

The Bayesian framework can help address multiple comparisons and multi-collinearity in part by a change in inferential focus.

Through the use of parameter shrinkage, the Bayesian regression framework is well adapted to address concerns with multi-collinearity and concerns for multiple comparisons.[8] We replicated the multivariable adjusted regression model in a hierarchical Bayesian framework by placing a common uninformative prior on the betas for county and state variables (β_c and β_s, respectively).

$$y_c \sim N\left(\hat{y}_c, \sigma_y^2\right)$$

$$\hat{y}_c = \alpha_0 + \alpha_s + \boldsymbol{\beta_c} + \boldsymbol{\beta_s}$$

$$\boldsymbol{\beta_c} \sim N(0, \sigma_c^2)$$

$$\boldsymbol{\beta_s} \sim N(0, \sigma_s^2)$$

$$a_0, a_j \sim N(0, 100)$$

$$\sigma_y^2, \sigma_c^2, \sigma_s^2 \sim uniform(0,100)$$

Where y_c, the county-level obesity prevalence is assumed to arise from a normal distribution with mean y-hat and variance sigma squared-y. Y-hat in turn is a function of a global intercept, α_0, a random intercept for state *s*, and the linear combination of betas from vector β_c and β_s, representing the effects of each covariate measured at the county and state level, respectively. Uniform priors were placed on each of the variance parameters, and normal priors were placed separately on the county and state level predictors. Inference is made from the median and 95% Bayesian credible interval of each parameter.

References

1. Zhang X, Holt JB, Yun S, Lu H, Greenlund KJ, Croft JB. Validation of Multilevel Regression and Poststratification Methodology for Small Area Estimation of Health Indicators From the Behavioral Risk Factor Surveillance System. *Am J Epidemiol*. 2015;182(2):127-137. http://dx.doi.org/10.1093/aje/kwv002.
2. Zhang X, Holt JB, Lu H, et al. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the behavioral risk factor surveillance system. *Am J Epidemiol*. 2014;179(8):1025-1033. http://dx.doi.org/10.1093/aje/kwu018.
3. Carle AC. Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Med Res Methodol*. 2009;9(1):49. http://dx.doi.org/10.1186/1471-2288-9-49.
4. Tobler WR. Smooth pycnophylactic interpolation for geographical regions. *J Am Stat Assoc*. 1979;74(367):519-530. http://dx.doi.org/10.1080/01621459.1979.10481647.

5. Zhang X, Onufrak S, Holt JB, Croft JB. A multilevel approach to estimating small area childhood obesity prevalence at the census block-group level. *Prev Chronic Dis*. 2013;10(8):E68. http://dx.doi.org/10.5888/pcd10.120252.

6. Harper S, Lynch J, Health P, Hall P. *Methods for Measuring Cancer Disparities : Using Data Relevant to Healthy People 2010 Cancer-Related Objectives*. Ann Arbor, MI; 2010.

7. Pearcy JN, Keppel KG. A summary measure of health disparity. *Public Health Rep*. 2002;117(3):273-280. http://dx.doi.org/10.1016/S0033-3549(04)50161-9.

8. Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press; 2007.