American Journal of Epidemiology
Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health 2014. This work is written by (a) US Government employee(s) and is in the public domain in the US.

Vol. 179, No. 8
DOI: 10.1093/aje/kwu018
Advance Access publication:
March 4, 2014

Practice of Epidemiology

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

Xingyou Zhang*, James B. Holt, Hua Lu, Anne G. Wheaton, Earl S. Ford, Kurt J. Greenlund, and Janet B. Croft

* Correspondence to Dr. Xingyou Zhang, Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, 4770 Buford Highway, Mailstop F78, Atlanta, GA 30341 (e-mail: gyx8@cdc.gov).

Initially submitted August 23, 2013; accepted for publication January 15, 2014.

A variety of small-area statistical models have been developed for health surveys, but none are sufficiently flexible to generate small-area estimates (SAEs) to meet data needs at different geographic levels. We developed a multilevel logistic model with both state- and nested county-level random effects for chronic obstructive pulmonary disease (COPD) using 2011 data from the Behavioral Risk Factor Surveillance System. We applied poststratification with the (decennial) US Census 2010 counts of census-block population to generate census-block-level SAEs of COPD prevalence which could be conveniently aggregated to all other census geographic units, such as census tracts, counties, and congressional districts. The model-based SAEs and direct survey estimates of COPD prevalence were quite consistent at both the county and state levels. The Pearson correlation coefficient was 0.99 at the state level and ranged from 0.88 to 0.95 at the county level. Our extended multilevel regression modeling and poststratification approach could be adapted for other geocoded national health surveys to generate reliable SAEs for population health outcomes at all administrative and legislative geographic levels of interest in a scalable framework.

Behavioral Risk Factor Surveillance System; chronic obstructive pulmonary disease; multilevel regression and poststratification; population health outcomes; small-area estimation

Abbreviations: BRFSS, Behavioral Risk Factor Surveillance System; COPD, chronic obstructive pulmonary disease; MRP, multilevel regression and poststratification; SAEs, small-area estimates.

National health surveys in the United States provide a critical cost-effective way to generate suitable statistics for measuring and monitoring national/state population health, but they do not have statistically sufficient samples to produce direct survey estimates for most counties or subcounty areas. In addition, population health data collection and surveillance systems are largely based on administrative geographic units (city, county, or state), so population health outcome data are not often available for legislative geographic units, such as congressional districts and state legislative districts. Thus, small-area estimation techniques (1), especially the statistical model-based approaches (2), have been extensively applied to national or state health surveys to generate reliable small-area estimates (SAEs) and to meet local data needs for public health program planning and evaluation.

Small-area statistical models, both unit-level and area-level, have been developed for the Behavioral Risk Factor Surveillance System (BRFSS), which was originally designed for reliable state-level survey estimates, for a variety of health outcome estimates at the county level (3–13) and zip-code level (14–17). Unit-level models use individual-level data from the BRFSS as outcomes, and area-level models use aggregated county-level estimates from the BRFSS as outcomes (3, 5, 11). These models have aimed to generate county or zip-code estimates only and have lacked the flexibility to simultaneously generate SAEs of multiple geographic units, such as congressional districts and local

neighborhoods (census tracts). Congdon (14) adapted a flexible Bayesian multilevel logistic model to the BRFSS for zip-code-level estimates, but the model ignored important county context effects (county random effects) on health outcomes (14, 15). Gelman and Little (18) and Park et al. (19) developed the multilevel regression and poststratification (MRP) approach for small-area estimation using national polling data. The MRP approach took both demographic and geographic characteristics into account; both internal split-sample and external validation showed that MRP with national polling data could generate more accurate and reliable state or congressional district estimates than direct survey estimates of public opinion outcomes (20, 21). We expect that the combination of the powerful model inference and prediction of MRP and the flexibility of Congdon's unit-level multilevel model, with a large sample of BRFSS data, could generate accurate and reliable SAEs at various geographic levels.

Thus, our main goal in this paper was to develop a more flexible unit-level multilevel model based on the (single-year) 2011 BRFSS data and apply poststratification with the (decennial) US Census 2010 counts of census-block population to generate census-block-level SAEs of population health outcomes. We used a unit-level, as opposed to an area-level, multilevel model for small-area estimation, because unit-level models are usually a better alternative in terms of model flexibility and avoiding the ecological fallacy (22). We used a single year because the temporal trend of population health outcomes over multiple years could introduce additional bias in the SAEs, and we also chose a single year in order to assess the feasibility of producing reliable annual SAEs from each BRFSS survey year. We chose the census block, the smallest basic unit of census geography (23), because census-block-level SAEs of health outcomes could be easily aggregated to meet data needs for larger census geographic units, such as census tracts, counties, and congressional districts. Since state and local health researchers and practitioners are more familiar with statistical programming and analysis in SAS (SAS Institute, Inc., Cary, North Carolina), we implemented this unit-level multilevel logistical mixed model with both individual-level fixed effects and county-level and state-level random effects using the GLIMMIX procedure in SAS 9.3; and we fitted this mixed model by maximum likelihood with Laplace approximation, which typically exhibits better asymptotic behavior and less small-sample bias than GLMMIX's default pseudolikelihood estimators (24).

We selected chronic obstructive pulmonary disease (COPD) as the individual health outcome because chronic lower respiratory disease (primarily COPD) has emerged as the third leading cause of US death since 2008 (25) and because evidence suggests that there exist state-level variations in COPD prevalence (26), Medicare hospitalizations for COPD (27), and COPD deaths (28).

METHODS

Data sources

We used the following 2 data sources in this paper: the BRFSS (2011 data) and US Census 2010. The 2011 BRFSS survey had a sample size of 497,967 persons aged ≥18 years in the 50 states and the District of Columbia. After removal of records with missing values for age (n = 4,903), race/ethnicity (n = 4,252), residential county (n = 108), and self-reported COPD (n = 3,811), 485,594 records remained. State-level sample sizes ranged from 3,349 (Alaska) to 25,075 (Nebraska), with a median of 8,258 (Indiana). The final sample included data from 3,124 counties (99.4% of 3,143 US counties), with 1,630 counties having a sample size of 50 or greater. Geographic coverage and large sample size make the BRFSS the most popular, and sometimes the only, data source for obtaining county or subcounty SAEs. Publicly available US Census small-area population and socioeconomic data are other important administrative data necessary for small-area estimation.

The BRFSS. The BRFSS compiles data from a state-based random-digit-dialed telephone survey of the noninstitutionalized US adult population aged ≥18 years. The survey is administered annually to households with landlines or cellular telephones by state health departments in collaboration with the Centers for Disease Control and Prevention. The median survey response rate in 2011 for all states and the District of Columbia was 49.7%, and response rates ranged from 33.8% to 64.1% (29).

Individual-level self-reported data were extracted from the 2011 BRFSS survey, including data on individual health outcomes and covariates. The following question on COPD diagnosis was introduced in the 2011 BRFSS core questionnaire: "Has a doctor, nurse, or other health professional ever told that you had chronic obstructive pulmonary disease (COPD), emphysema, or chronic bronchitis?" A binary variable (1 = yes; 0 = no) was based on the response to this COPD question, excluding persons who refused or did not know.

Individual covariates included individual respondents' age (18–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, or ≥80 years), sex (male or female), and race/ethnicity (non-Hispanic white; non-Hispanic black; American Indian or Alaska Native; Asian; Native Hawaiian or other Pacific Islander; other single race; 2 or more races; or Hispanic), as well as sampled subjects' residential counties and states.

Census 2010 and the American Community Survey. Census block-level data corresponding to BRFSS age × sex × race/ethnicity cross-tabulated categories were extracted from Census 2010, resulting in 208 demographic categories for each census block. Five-year estimates from the American Community Survey provide up-to-date sociodemographic data for census tracts and block groups that are needed for incorporating local community contexts, such as poverty. We used census-tract-level and county-level poverty rates (under 150% of the federal poverty level) from the most recent American Community Survey 5-year estimate (2007–2011).

Direct survey estimates

The unadjusted weighted prevalences of COPD and 95% confidence intervals were obtained from the BRFSS for population subgroups defined by selected characteristics (age, sex, and race/ethnicity) using SAS-callable SUDAAN 11.0.0 (Research Triangle Institute, Research Triangle Park,

North Carolina). Data were weighted using the new raking methods (30). The BRFSS sampling strata, final weights, and primary sampling units were used to calculate direct survey estimates of overall COPD prevalence in SUDAAN for the entire United States, for all 50 states and the District of Columbia, and for those counties with at least 50 valid subjects.

Model specification

Our unit-level multilevel logistic model for COPD followed the general format of generalized linear mixed models: $y = X\beta + Z\alpha + \varepsilon$. Specifically, the probability of self-reported COPD (P_{ijkcs}) was assumed to be associated with 3 level-related factors—individual, county, and state—via a logit link:

$$P_{ijkcs}(y_{ijkcs} = 1) = \text{logit}^{-1}(\alpha_i + \beta_j + \gamma_k + x_c'\eta + \mu_c + v_s + e_{ijkcs}). \quad (1)$$

In this 3-level logistic regression model, Y_{ijkcs} is self-reported COPD status (1 = yes; 0 = no) for an individual of age group i (i = 1–13), sex j (j = 1, 2), and race/ethnicity k (k = 1–8) from county c in state s; α_i, β_j, and γ_k are the regression coefficients corresponding to age group i, sex j, and racial/ethnic group k, respectively; and all individual-level data on COPD status, age, sex, race/ethnicity, and their corresponding county and state identifiers are from the BRFSS. x_c is the vector of county-level covariates, and η is the vector of corresponding regression coefficients. County-level factors are usually from data sources other than the BRFSS. For simplicity, we included only county-level poverty status from American Community Survey 2007–2011 in the model, since economic poverty is a robust factor associated with local health disparities (31, 32). μ_c, v_s, and e_{ijkcs} are the county, state, and residual random effects; all 3 random effects are assumed to be independent and normally distributed. We refer to equation 1 as the multilevel prevalence model.

County-level random effects statistically account for county-level correlations among individual observations in model-fitting; epidemiologically, they represent county-level contextual effects on health outcomes. County-level random effects may become insignificant if we explicitly include all relevant county-level risk factors in the model. However, data on many important county-level factors associated with health outcomes are unavailable. A more complex situation is that the same factor could have differential impacts on health outcomes in different counties, and the important factors for a health outcome could be very different between counties. County random effects allow us the flexibility to incorporate county-level contextual effects while not imposing a universal impact of 1 county-level factor on health outcomes of interest. Similar statements also apply to state-level random effects in the model.

We fitted the above multilevel prevalence model using the procedure GLMMIX in SAS. To account for unequal probability sampling of respondents in the BRFSS, the rescaled weights (RW_{ijkcs}) were included in the model estimation.

$$\text{RW}_{ijkcs} = \frac{W_{ijkcs}}{\sum_S W_{ijkcs}P_{ijkcs}} \times N_s, \quad (2)$$

where W_{ijkcs} is the original BRFSS weight and $\Sigma_S W_{ijkcs}$ and N_s are the sum of total original weights and total sample sizes for state s, respectively. We rescaled the weights by state to reflect the reality that BRFSS data represent a group of independent state-based surveys. We rescaled the weights because GLIMMIX's WEIGHT statement treats its weight variable as a frequency weight. Therefore, had we included BRFSS original weights in the GLIMMIX WEIGHT statement, GLIMMIX would have fitted the models with a sample size equivalent to the total sample weights of the BRFSS data set in the analysis (more than 200 million), which would have significantly underestimated the standard errors associated with model parameters. Thus, we rescaled the original weights to ensure that the sum of rescaled weights equaled the sample size of the final BRFSS data set in the analysis. We also ran analyses from this model without BRFSS weighting to see whether the SAEs from the unweighted model had larger bias than those from the weighted one.

Model prediction

Individual-level expected probability of COPD. From the multilevel prevalence model above, we obtained model parameters for 13 age categories, 2 sex categories, and 8 racial/ethnic categories, county-level poverty, and county- and state-level random effects. We defined county-level random effects for those counties without samples (μ_c^i) by spatially smoothing their neighboring counties' random effects (μ_c^j):

$$\mu_c^i = \frac{\sum_{j=1}^{N_j} \mu_c^{ji}}{N_j},$$

where N_j is the number of spatially adjacent counties for county i.

We applied the model parameters from the multilevel prevalence model (equation 1) to census-block-level population counts to construct the multilevel prediction model (equation 3). To obtain the individual expected COPD risk for all age, sex, and racial/ethnic groups in all census blocks for all counties within all states, we take

$$P_{ijkcs}^b(y_{ijkcs} = 1) = \text{logit}^{-1}(\alpha_i + \beta_j + \gamma_k + x_b'\eta + \mu_c + v_s) = \frac{\exp(\alpha_i + \beta_j + \gamma_k + x_b'\eta + \mu_c + v_s)}{1 + \exp(\alpha_i + \beta_j + \gamma_k + x_b'\eta + \mu_c + v_s)}, \quad (3)$$

where P_{ijkcs}^b is the predicted COPD risk for an individual of age group i, sex j, and racial/ethnic group k in census block b within county c and state s. In Census 2010, we know the population count by age, sex, and race/ethnicity in a census block, as well as its corresponding census-tract-level and county-level poverty rates from American Community Survey 2007–2011. Thus, when we construct the multilevel prediction model, we usually use county-level poverty multiplied by the regression coefficient of county-level poverty (η) to make predictions, but we could also use tract-level poverty multiplied by the regression coefficient of county-level

poverty (η) to make predictions. In the multilevel prediction model (equation 3), all of the individual-level age, sex, and race/ethnicity data and their county and state identifiers were from Census 2010. In order to account for the impact of local community poverty on COPD, we replaced the county-level poverty rate (x_c) with the census-tract-level poverty rate (x_b) in the multilevel prediction model (equation 3). Here we assumed that the impact of county-level poverty (measured by the regression coefficient of county-level poverty) could be linearly transformed to the census tract level. We used continuous poverty rates to reflect the larger variation of poverty rates observed at the census tract level and to better reflect their local impact on COPD outcome. Thus, an individual's predicted COPD risk was adjusted for individual age, sex, and race/ethnicity and further adjusted for local community (census tract) poverty status and county- and state-level contextual effects.

Census-block-level COPD prevalence via poststratification. Census-block-level COPD prevalence (P^b_{cs}) (equation 4) was obtained by summing the predicted individual COPD risks over 208 demographic categories in a census block weighted by the categories' corresponding population sizes (Pop^b_{ijkcs}) in that census block (b):

$$P^b_{cs} = \frac{\sum_i \sum_j \sum_k (P^b_{ijkcs} \times \mathrm{Pop}^b_{ijkcs})}{\sum_i \sum_j \sum_k \mathrm{Pop}^b_{ijkcs}} = \frac{\sum_i \sum_j \sum_k (P^b_{ijkcs} \times \mathrm{Pop}^b_{ijkcs})}{\mathrm{Pop}^b_{cs}}, \quad (4)$$

where Pop^b_{ijkcs} is the population size (count) at the census block level for a person of age i, sex j, and race/ethnicity k in census block b, county c, and state s and Pop^b_{cs} is the total population in census block b, county c, and state s. The census-block-level SAEs could then be conveniently aggregated to obtain the SAEs (P_g) for any larger units of census geography as follows:

$$P_g = \frac{\sum_{b=1}^{N} (P^b_{cs} \times \mathrm{Pop}^b_{cs})}{\sum_{b=1}^{N} \mathrm{Pop}^b_{cs}}, \quad (5)$$

where N is the number of census blocks for the target geographic units (g), such as census tracts, zip codes, counties, and congressional districts. Thus, by predicting the individual-level expected risk of COPD for populations within the lowest geographic unit (census block), we could obtain SAEs of COPD prevalence for any geographic units of interest.

We then randomly drew 1,000 samples of the model parameters from their estimated conditional distributions and generated a sample of 1,000 SAEs for each census-block-level COPD prevalence by age, sex, and racial/ethnic group. We empirically constructed point-prediction mean values and 95% confidence intervals and standard errors from this sample of 1,000 for the SAEs for census blocks and any other units of census geography.

Comparing model-based SAEs and direct survey estimates

It is important to evaluate both the internal and external validity of our model-based census-block-level SAEs. We lacked comparable existing subcounty-level data with which to assess the reliability of SAEs of COPD prevalence. Here we borrow a key concept from benchmarking small-area estimation: SAEs, when aggregated to a higher geographic level, should be consistent with direct estimates from the original survey and should have reliable inferential accuracy (33). Substantial differences between aggregated model-based SAEs and the corresponding direct survey estimates for a desirable large geographic area by original survey design would suggest the misspecification of small-area models. Because the BRFSS was designed to generate reliable state-level estimates, consistency of state-level aggregated model-based SAEs and direct survey estimates should be expected if our small-area model is valid. The BRFSS has a large set of counties with a large sample size of 50 or more. The agreement between county-level model-based SAEs and direct estimates for these counties should also be consistent. Therefore, we compared the model-based SAEs with the corresponding direct estimates for all states and for the counties with a sample size of 50 or more. In addition to basic descriptive statistics, we used the correlation coefficients and mean squared errors and mean absolute differences to evaluate the consistency between model-based SAEs (m_i) and direct survey estimates (s_i). The mean squared error is defined as $1/N \sum_{i=1}^{N} (m_i - s_i)^2$ and the mean absolute difference as $1/N \sum_{i=1}^{N} |m_i - s_i|$, where N is the number of counties or states in the comparison.

RESULTS

BRFSS direct survey estimates and model results

The unadjusted observed prevalence of COPD in the BRFSS sample was 6.36% (Table 1). The prevalence of direct survey estimates differed between groups defined by sex, age, and race/ethnicity. Men had a lower COPD prevalence than women; the prevalence increased among successive age groups; and American Indian/Alaska Native and multiple-race groups had a higher prevalence than other racial/ethnic groups. These differences were also observed in the multilevel model that resulted in SAEs (Table 2). In addition, county poverty status also had a significant influence on model-based estimates of COPD: Persons living in a county with higher poverty rates experienced significantly higher COPD (Table 2). As expected, state- and county-level random effects confirmed the significant impact of both state- and county-level contextual environments (Table 2).

Comparison of model-based SAEs and direct survey estimates

We compared model-based SAEs and BRFSS direct survey estimates at both the county and state levels (Table 3). For 1,630 counties with a sample size of 50 or more (more than half of 3,143 US counties), the Pearson correlation

Table 1. Unadjusted Weighted Prevalence of Chronic Obstructive Pulmonary Disease Among Adults Aged ≥18 Years, by Sex, Age, and Race/Ethnicity, United States, 2011[a]

Characteristic	No. of Respondents	COPD Prevalence	
		%	95% CI
Total	485,594	6.36	6.23, 6.49
Sex			
Male	190,577	5.39	5.21, 5.58
Female	295,017	7.27	7.10, 7.46
Age group, years			
18–24	21,870	2.65	2.29, 3.07
25–29	21,005	2.98	2.56, 3.46
30–34	26,555	2.78	2.46, 3.14
35–39	28,851	3.01	2.66, 3.39
40–44	34,008	4.63	4.21, 5.10
45–49	39,712	5.62	5.22, 6.05
50–54	49,293	7.54	7.11, 8.00
55–59	53,511	8.84	8.39, 9.30
60–64	55,283	9.75	9.28, 10.24
65–69	46,254	12.13	11.56, 12.72
70–74	37,937	12.53	11.89, 13.20
75–79	30,577	12.86	12.18, 13.57
≥80	40,738	10.85	10.32, 11.41
Race/ethnicity			
White, non-Hispanic	386,195	7.11	6.96, 7.26
Black, non-Hispanic	40,143	6.00	5.58, 6.45
American Indian/ Alaska Native	6,919	10.56	9.19, 12.10
Asian	8,655	1.72	1.32, 2.25
Native Hawaiian/ Pacific Islander	910	5.81	3.47, 9.56
Other single race	2,869	5.95	4.10, 8.55
≥2 races	8,665	10.99	9.60, 12.55
Hispanic	31,238	3.50	3.16, 3.87

Abbreviations: CI, confidence interval; COPD, chronic obstructive pulmonary disease.

[a] Data were obtained from the Behavioral Risk Factor Surveillance System.

Table 2. Regression Coefficients for Fixed Effects and Variance Components in the Unit-Level Multilevel Logistic Model of Chronic Obstructive Pulmonary Disease Risk, United States, 2011[a]

Effect and Subgroup	Estimate (β)	Standard Error	*P* Value
Intercept	−3.92	0.055	<0.0001
Sex			
Male	−0.24	0.012	<0.0001
Female	0	Referent	
Age group, years			
18–24	0	Referent	
25–29	0.16	0.040	<0.0001
30–34	0.11	0.039	0.0049
35–39	0.27	0.040	<0.0001
40–44	0.64	0.035	<0.0001
45–49	0.86	0.034	<0.0001
50–54	1.12	0.031	<0.0001
55–59	1.33	0.032	<0.0001
60–64	1.44	0.032	<0.0001
65–69	1.64	0.033	<0.0001
70–74	1.69	0.034	<0.0001
75–79	1.71	0.034	<0.0001
≥80	1.55	0.034	<0.0001
Race/ethnicity			
White, non-Hispanic	0	Referent	
Black, non-Hispanic	−0.07	0.023	0.0038
American Indian/ Alaska Native	0.54	0.044	<0.0001
Asian	−0.93	0.064	<0.0001
Native Hawaiian/Pacific Islander	0.23	0.137	0.0870
Other single race	0.25	0.079	0.0016
≥2 races	0.69	0.042	<0.0001
Hispanic	−0.26	0.027	<0.0001
County poverty, %	0.02	0.002	<0.0001
Variance components			
State level	0.04	0.010	<0.0001
County level	0.12	0.009	<0.0001

[a] Data were obtained from the Behavioral Risk Factor Surveillance System.

coefficient for the correlation between their model-based SAEs and direct survey estimates was 0.878; it increased to 0.912 when weighted by the county BRFSS sample size (Table 3). For 563 counties with a sample size of at least 50 and direct survey estimates with a coefficient of variation no greater than 0.3, the Pearson correlation coefficient was 0.940, and it increased to 0.959 when weighted by the county BRFSS sample size (Table 3). The Pearson correlation coefficient for correlation between state-level model-based SAEs and direct survey estimates was 0.997 (Table 3). Spearman rank correlation coefficients and concordance correlation coefficients for the correlation between model-based SAEs and direct survey estimates yielded similar patterns (data not shown).

Table 4 presents the basic summary statistics for these model-based SAEs and direct estimates. At the state level, both model-based SAEs and direct estimates had roughly equivalent minimum, median, and maximum values and interquartile ranges (Table 4). Compared with direct survey estimates, the corresponding county-level model-based SAEs had much smaller variations in COPD prevalence; the mean squared errors and mean absolute differences of model-based SAEs (direct survey as benchmarks) were 5.00% and 1.51%, respectively, for 1,630 counties with a sample of at least 50; they decreased further to 2.87% and 0.89% for

Table 3. Pearson Correlation Coefficients for the Correlation Between Model-based Small-Area Estimates and Direct Estimates of Chronic Obstructive Pulmonary Disease Prevalence, United States, 2011[a]

Geographic Level and No. of Units	Pearson Correlation Coefficient		Sample Limits	
	Pearson I[b]	Pearson II[c]	Sample Size (*n*)	Coefficient of Variation[d]
County				
1,630	0.878	0.704	≥50	
1,630	0.912	0.779[e]	≥50	
563	0.940	0.853	≥50	≤0.3
563	0.959	0.877[e]	≥50	≤0.3
State				
51	0.997	0.971		
51	0.997	0.975[e]		

Abbreviation: BRFSS, Behavioral Risk Factor Surveillance System.

[a] Data were obtained from the BRFSS.

[b] Correlation between the small-area estimates based on the multilevel logistic model using BRFSS rescaled final survey weights and BRFSS direct survey estimates.

[c] Correlation between the small-area estimates based on the multilevel logistic model without use of BRFSS final survey weights and BRFSS direct survey estimates.

[d] Coefficient of variation for BRFSS direct survey estimates, equivalent to the ratio of the standard error to the mean estimated prevalence of chronic obstructive pulmonary disease for a county or a state.

[e] Weighted by BRFSS county or state sample sizes.

those counties with a sample size of at least 50 and a direct survey estimate coefficient of variation of no more than 0.30. They were 0.02% and 0.11% at the state level.

Geographic variations in COPD prevalence

Model-based COPD prevalence varied widely across different geographic units. The model-based national estimate was 6.33%, which is consistent with the crude COPD prevalence in Table 1 and prevalences reported earlier (26). The overall ranges were smaller at more aggregated levels of census geography: 0.41%–61.01% for census blocks, 1.02%–33.20% for census tracts, 2.31%–26.32% for counties, 2.85%–13.43% for congressional districts, and 4.13%–9.93% for states (Table 4).

The maps of COPD prevalence depicted more detailed geographic variations in COPD prevalence across the United States at different geographic levels (Figures 1–3). Figure 1 demonstrates the county-level model-based prevalence, which shows variation in clustering of high COPD prevalence within states. Figure 2 depicts the model-based COPD prevalence by congressional district; this estimation shows geographic clustering of high and low levels that almost entirely covers some states. Estimation at the census-tract level in Figure 3 provides greater details about clustering at more local levels.

DISCUSSION

We generated census-block-level estimates of COPD prevalence via a unit-level multilevel logistic model with the use of BRFSS and Census 2010 population data. We modified the original MRP and enhanced its inference power by combining a large health survey with the best available small-area population data (14, 18). The high correlations between county- and state-level model-based estimates of COPD prevalence and their corresponding direct survey estimates support the accuracy of our SAEs. Our multilevel approach to small-area estimation with BRFSS data has great flexibility for producing SAEs with a variety of geographic units, from local neighborhoods to congressional districts. These SAEs could be very useful in a variety of contexts and could meet the diverse small-area health data needs of local policy-makers, program planners, and communities.

Our unit-level multilevel logistic model included both state- and nested county-level random effects to take both statewide and countywide contextual effects into account. In previous studies with BRFSS data, county random effects were ignored or dropped, mainly for reasons of simplicity (10, 14, 15). In our study, significant county-level random effects still existed after adjustment for individual-level demographic factors, as well as county-level poverty status. Thus, distinct county-level geographic contextual effects on COPD prevalence seem not to be explainable by local demography alone. MRP ignores the individual survey weights in the model estimation and assumes that later poststratification using small-area populations could overcome the bias from ignoring survey weights (18, 19, 21). However, the Pearson correlation coefficients for correlation between the SAEs based on the model without rescaled BRFSS survey weights and direct survey estimates became smaller at both the county and state levels (Table 3); thus, the unit-level multilevel model without BRFSS weighting may introduce some bias into SAEs.

Several limitations should be noted. First, our model did not directly address the BRFSS stratification, and state- and county-level random effects did not fully incorporate the BRFSS strata into model-fitting. Second, our model did not account for spatial correlations between counties or states, which could be conveniently handled in a full Bayesian approach (14, 18). Third, the transformation of contextual factors between county and census tract may be more complex than we assumed in our model. Further research on this cross-level inference (bias and uncertainty) is needed.

We believe that the methodology proposed in this paper can provide a useful tool for public health practitioners to create SAEs using BRFSS data. We employed a relatively parsimonious model for COPD. The unit-level multilevel model could be easily modified for any other BRFSS health outcome, and its performance could be further improved by introducing more geodemographic factors relevant to specific population health outcomes of interest. Finally, our extended multilevel regression modeling and poststratification approach could be adapted for other geocoded national health surveys to generate reliable SAEs for population health outcomes at all administrative and legislative geographic levels of interest in a scalable framework.

Table 4. Summary Statistics for Model-based Small-Area Estimates and Direct Estimates of Chronic Obstructive Pulmonary Disease Prevalence at Various Geographic Levels, United States, 2011[a]

Geographic Level and Method	No. of Units	Estimated COPD Prevalence, %								
		Minimum	First Quartile	Median	Third Quartile	Maximum	Mean	IQR	MSE	MAD
State										
Model[b]	51	4.13	5.32	6.11	7.70	9.93	6.40	2.38	0.02	0.11
Survey[c]	51	4.00	5.15	6.08	7.62	9.86	6.36	2.47		
County										
Model[b]	3,143	2.31	6.03	7.53	9.26	26.32	7.82	3.23		
Model[b,d]	1,630	2.31	5.56	6.98	8.81	18.79	7.36	3.25	5.00	1.51
Survey[c,d]	1,630	0.00	4.45	6.51	9.21	34.55	7.29	4.76		
Model[b,e]	563	2.31	5.49	6.93	8.79	18.79	7.40	3.30	2.87	0.89
Survey[c,e]	563	0.00	5.23	6.88	9.40	34.55	7.82	4.16		
Congressional district										
Model[b]	436	2.85	4.81	6.11	7.56	13.43	6.28	2.75		
Census tract										
Model[b]	72,531	1.02	4.48	6.04	8.01	33.20	6.53	3.53		
Census block										
Model[b]	6,206,505	0.41	4.79	6.69	9.16	61.01	7.37	4.37		

Abbreviations: BRFSS, Behavioral Risk Factor Surveillance System; COPD, chronic obstructive pulmonary disease; IQR, interquartile range; MAD, mean absolute difference; MSE, mean squared error.

[a] Data were obtained from the BRFSS.

[b] Model-based small-area estimates.

[c] BRFSS direct survey estimates.

[d] Counties with at least 50 BRFSS respondents.

[e] Counties with at least 50 BRFSS respondents and coefficients of variation no greater than 0.3 for BRFSS direct survey estimates.

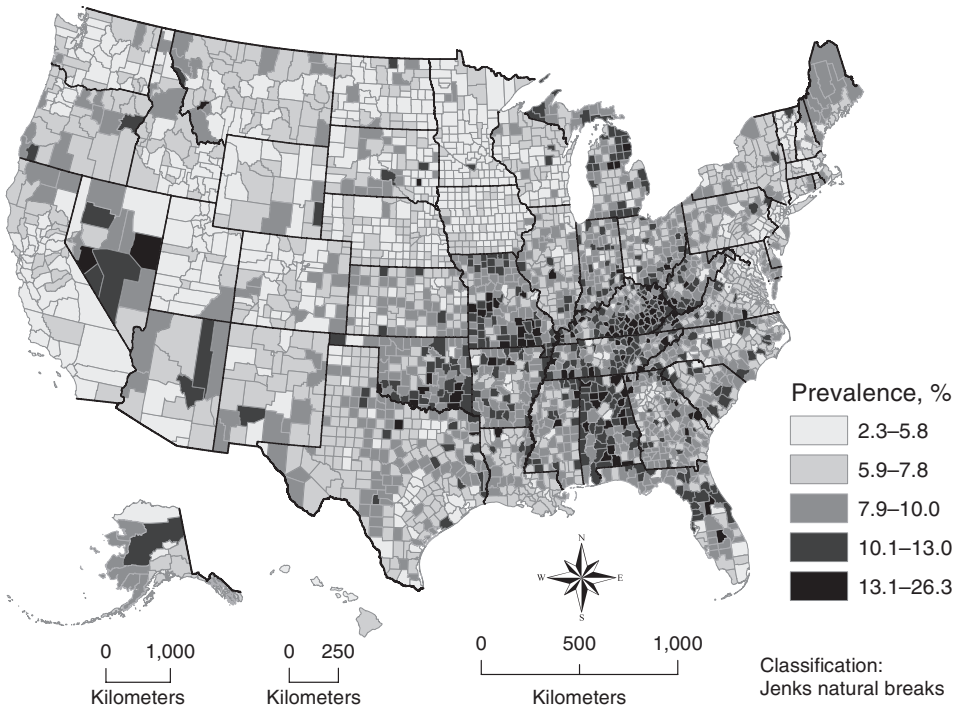

Figure 1. Model-based prevalence of chronic obstructive pulmonary disease, by county, United States, 2011.

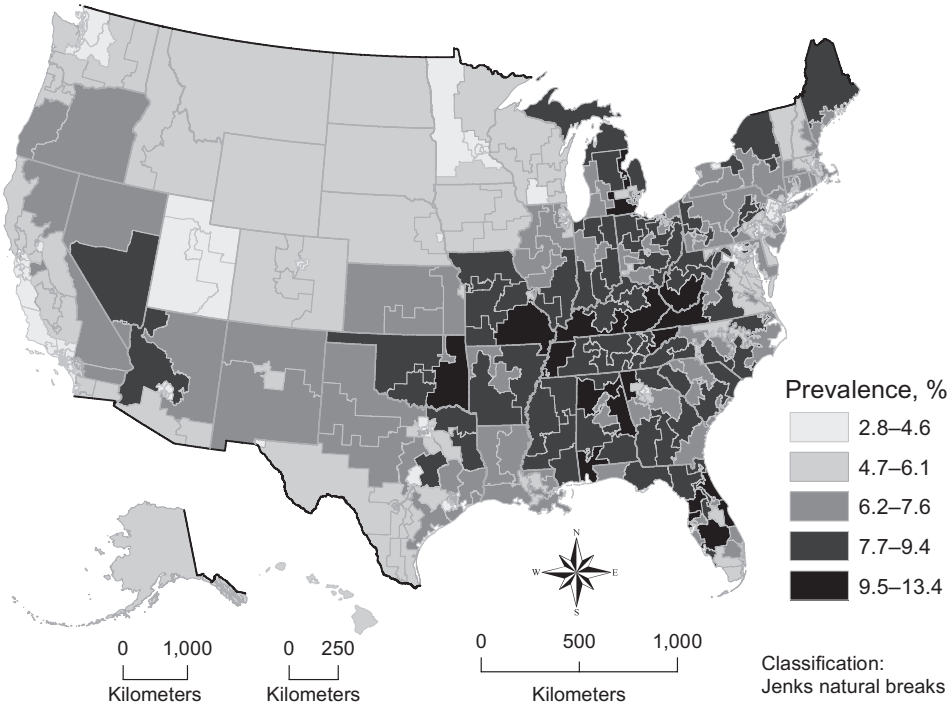

Figure 2. Model-based prevalence of chronic obstructive pulmonary disease, by congressional district, United States, 2011.

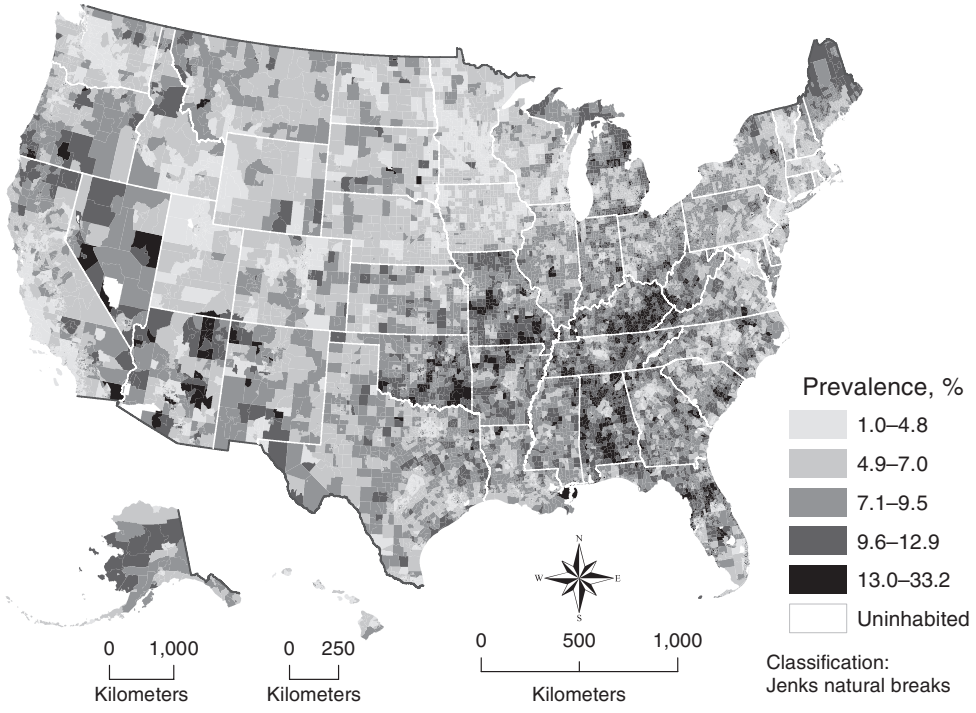

Figure 3. Model-based prevalence of chronic obstructive pulmonary disease, by census tract, United States, 2011. Blank white areas are non-populated areas.

ACKNOWLEDGMENTS

Author affiliations: Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, Georgia (Xingyou Zhang, James B. Holt, Hua Lu, Anne G. Wheaton, Earl S. Ford, Kurt J. Greenlund, and Janet B. Croft).

We thank Dr. Carol Gotway Crawford of the Division of Population Health, National Center for Chronic Disease Prevention and Health Promotion, for her insightful comments.

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

Conflict of interest: none declared.

REFERENCES

1. Rao JNK. *Small Area Estimation*. New York, NY: John Wiley & Sons, Inc.; 2003.
2. Datta GS. Model-based approach to small area estimation. In: Pfeffermann D, Rao CR, eds. *Handbook of Statistics 29. Vol. 29B. Sample Surveys: Inference and Analysis*. Amsterdam, the Netherlands: North Holland; 2009:251–288.
3. Xie D, Raghunathan TE, Lepkowski JM. Estimation of the proportion of overweight individuals in small areas—a robust extension of the Fay-Herriot model. *Stat Med*. 2007;26(13): 2699–2715.
4. Zhang Z, Zhang L, Penman A, et al. Using small-area estimation method to calculate county-level prevalence of obesity in Mississippi, 2007–2009. *Prev Chronic Dis*. 2011; 8(4):A85.
5. Cadwell BL, Thompson TJ, Boyle JP, et al. Bayesian small area estimates of diabetes prevalence by U.S. county, 2005. *J Data Sci*. 2010;8(1):173–188.
6. Earnest A, Beard JR, Morgan G, et al. Small area estimation of sparse disease counts using shared component models—application to birth defect registry data in New South Wales, Australia. *Health Place*. 2010;16(4):684–693.
7. Eberth JM, Hossain MM, Tiro JA, et al. Human papillomavirus vaccine coverage among females aged 11 to 17 in Texas counties: an application of multilevel, small area estimation. *Womens Health Issues*. 2013;23(2):e131–e141.
8. Goodman MS. Comparison of small-area analysis techniques for estimating prevalence by race. *Prev Chronic Dis*. 2010;7(2): A33.
9. Jia H, Link M, Holt J, et al. Monitoring county-level vaccination coverage during the 2004–2005 influenza season. *Am J Prev Med*. 2006;31(4):275–280.
10. Jia H, Muennig P, Borawski E. Comparison of small-area analysis techniques for estimating county-level outcomes. *Am J Prev Med*. 2004;26(5):453–460.
11. Olives C, Myerson R, Mokdad AH, et al. Prevalence, awareness, treatment, and control of hypertension in United States counties, 2001–2009. *PLoS One*. 2013;8(4):e60308.
12. Schneider KL, Lapane KL, Clark MA, et al. Using small-area estimation to describe county-level disparities in mammography. *Prev Chronic Dis*. 2009;6(4):A125.
13. Srebotnjak T, Mokdad AH, Murray CJ. A novel framework for validating and applying standardized small area measurement strategies. *Popul Health Metr*. 2010;8:26.
14. Congdon P. A multilevel model for cardiovascular disease prevalence in the US and its application to micro area prevalence estimates. *Int J Health Geogr*. 2009;8:6.
15. Congdon P, Lloyd P. Estimating small area diabetes prevalence in the US using the Behavioral Risk Factor Surveillance System. *J Data Sci*. 2010;8(2):235–252.
16. Li W, Kelsey JL, Zhang Z, et al. Small-area estimation and prioritizing communities for obesity control in Massachusetts. *Am J Public Health*. 2009;99(3):511–519.
17. Li W, Land T, Zhang Z, et al. Small-area estimation and prioritizing communities for tobacco control efforts in Massachusetts. *Am J Public Health*. 2009;99(3):470–479.
18. Gelman A, Little TC. Poststratification into many categories using hierarchical logistic regression. *Surv Methodol*. 1997; 23(2):127–135.
19. Park DK, Gelman A, Bafumi J. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Polit Anal*. 2004;12(4):375–385.
20. Lax JR, Phillips JH. How should we estimate public opinion in the states? *Am J Polit Sci*. 2009;53(1):107–121.
21. Warshaw C, Rodden J. How should we measure district-level public opinion on individual issues? *J Polit*. 2012;74(1): 203–219.
22. Marhuenda Y, Molina I, Morales D. Small area estimation with spatio-temporal Fay-Herriot models. *Comput Stat Data Anal*. 2013;58(C):308–325.
23. Bureau of the Census, US Department of Commerce. *Standard Hierarchy of Census Geographic Entities*. Washington, DC: US Census Bureau; 2010. (http://www.census.gov/geo/reference/pdfs/geodiagram.pdf). (Accessed April 22, 2013).
24. Shun Z. Another look at the salamander mating data: a modified Laplace approximation approach. *J Am Stat Assoc*. 1997; 92(437):341–349.
25. Miniño AM. *Death in the United States, 2009*. (NCHS data brief, no. 64). Hyattsville, MD: National Center for Health Statistics; 2011.
26. Centers for Disease Control and Prevention. Chronic obstructive pulmonary disease among adults—United States, 2011. *MMWR Morb Mortal Wkly Rep*. 2012;61(46):938–943.
27. Holt JB, Zhang X, Presley-Cantrell L, et al. Geographic disparities in chronic obstructive pulmonary disease (COPD) hospitalization among Medicare beneficiaries in the United States. *Int J Chron Obstruct Pulmon Dis*. 2011;6:321–328.
28. Brown DW, Croft JB, Greenlund KJ, et al. Deaths from chronic obstructive pulmonary disease—United States, 2000–2005. *MMWR Morb Mortal Wkly Rep*. 2008;57(45):1229–1232.
29. Office of Surveillance, Epidemiology, and Laboratory Services, Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System: 2011 Summary Data Quality Report*. Atlanta, GA: Centers for Disease Control and Prevention; 2013.
30. Pierannunzi C, Town M, Garvin W, et al. Methodologic changes in the Behavioral Risk Factor Surveillance System in 2011 and potential effects on prevalence estimates. *MMWR Morb Mortal Wkly Rep*. 2012;61(22):410–413.
31. Krieger N, Chen JT, Waterman PD, et al. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. *Am J Epidemiol*. 2002;156(5):471–482.
32. Thomas AJ, Eberly LE, Davey Smith G, et al. ZIP-code-based versus tract-based income measures as long-term risk-adjusted mortality predictors. *Am J Epidemiol*. 2006;164(6):586–590.
33. Bell WR, Datta GS, Ghosh M. Benchmarking small area estimators. *Biometrika*. 2013;100(1):189–202.