Journal of Data Science **8**(2010), 173-188

Bayesian Small Area Estimates of Diabetes Prevalence by U.S. County, 2005

Betsy L. Cadwell, Theodore J. Thompson, James P. Boyle
and Lawrence E. Barker
National Center for Chronic Disease Prevention and Health Promotion

Abstract: County specific estimates promote understanding of national and state patterns of the diabetes burden and can help better target diabetes programs. Using Bayesian multilevel models, the authors estimated the prevalence of self reported diagnosed diabetes for adults aged 20 years or older for each of the United States' 3,141 counties/county equivalents. These estimates provide the first comprehensive county level estimates of diabetes for the U.S. and provide opportunities for the practical targeting of interventions and new lines of investigations into area level risk factors for diabetes. The ranks' posterior distribution was used to identify counties with extreme diabetes burden. Counties with high (low) diabetes burden were identified as those for which at least 95% of the posterior distribution for the rank was above (below) the median. In 2005, 428 (480) counties had high (low) diabetes burden. Design-based estimates could be obtained for 232 large population counties; model-based estimates compared favorably with these design-based estimates.

Key words: Bayesian analysis, BRFSS, census data, diabetes mellitus, multilevel model, small area estimation.

1. Introduction

In the United States, diabetes is a common chronic disease with enormous social and economic cost (American Diabetes Association, 2008). Behavioral Risk Factors Surveillance System (BRFSS) based estimates of diagnosed diabetes prevalence have increased in every state between 2002 and 2006 (Centers for Disease Control and Prevention, 2008a). The national prevalence of diagnosed diabetes was 5.5% in 2005 (Centers for Disease Control and Prevention, 2008b) and is projected to be 7.2% by 2050 (Boyle *et al.*, 2001). A child born in 2000 in the U.S. has more than a 1 in 3 risk of developing diabetes sometime in his or her life (Narayan *et al.*, 2003).

Efforts to control diabetes are generally conducted locally, often at the level of the county (a unit of local government in the U.S.) or county equivalent (hereafter 'county'). Concentrating efforts on high burden counties can be an effective strategy for preventing future cases of diabetes and complications for those who already have diabetes. Similarly, identifying low burden counties might show where prevention efforts have succeeded. However, estimates of diabetes prevalence have historically been available primarily for states. Existing surveys such as the BRFSS are not designed to produce direct county-level estimates. Although the BRFSS can provide design-based direct estimates for a few counties, it can not for most; the BRFSS sample size for many counties is small, even zero.

Several approaches for producing small area estimates from large complex sample surveys have been proposed; these include demographic methods, synthetic estimators, composite estimators, and model-based estimates (Rao, 2003). Demographic methods, which do not involve sampling, use administrative data to adjust Census estimates for population changes in post-censual years (Purcell and Kish, 1979). Synthetic estimators apply an estimate from a large area to small areas. These have the appeal of simplicity. However, these estimates assume that small and large areas have the same characteristics and are usually not sufficiently variable among locations for plausibility (Sarndal, 1984). Composite estimators are a weighted average of direct and synthetic estimators but differ little from synthetic estimators for areas with a small sampling fraction (Falorsi *et al.*, 1994). More recently, model-based approaches with random effects have been used to overcome the limitations of the approaches mentioned above (Lawson *et al.*, 2000; MacNab, 2003; Malec *et al.*, 1997; Malec *et al.*, 1999; Raghunathan *et al.*, 2007; Xie *et al.*, 2007).

Model-based approaches can be either area or unit- level (Ghosh and Rao, 1994). Area-level models use direct estimates and their estimated variances as data in a multilevel model with area-level auxiliary data as random effects. The popular Fay-Herriott model is an example (Fay and Herriott, 1979). Unit-level models relate unit values of the study variable to unit-specific auxiliary data and include area-level auxiliary data as random effects. Unlike area-level models, unit-level models require the sample design be taken into account and require unit-specific auxiliary data for every person in the small area.

Here, we develop a Bayesian unit-level model for the 2005 prevalence of self-reported diabetes in each of the United States' 3141 counties. For each county, we generate the posterior distribution for the prevalence and use its mean as the estimated prevalence for the county. We define a 95% interval for the estimate as the interval between the 2.5th and 97.5th percentile of the posterior distribution. We refer to these as confidence intervals although technically they are central posterior intervals. To answer questions, commonly arising in public health, concerning county ranks, we also produce the posterior distribution for each county's rank.

The general framework for our model-based approach follows the recommendation of Gelman *et al.* (2004, chap. 21) and treats available survey data as observed data collected from a larger set of complete data. We develop a probability model for the unobserved data using a method similar to the two-level binomial model proposed by Malec *et al.* (1997).

2. Methods

2.1 Data

The BRFSS is an ongoing, state-based, random-digit-dialing telephone survey of non-institutionalized adults aged 18 years or older in all 50 states, the District of Columbia and Puerto Rico. It provides the most widely-used state-level estimates of health status, including prevalence of diagnosed diabetes. Respondents are classified as having diagnosed diabetes if they responded "yes" to "Have you ever been told by a doctor that you have diabetes?" Those who responded "yes," but told only during pregnancy (gestational diabetes), or those who responded "no" were considered not to have diabetes. Respondents who did not know or refused to answer the question were considered to have missing diabetes status. County of residence for survey respondents is available through the internal BRFSS files of the Centers for Disease Control and Prevention (CDC). For respondents with missing county of residence, CDC uses the county most likely associated with the respondent's telephone number.

We combined data from the 2004, 2005 and 2006 BRFSS surveys. For each year and each of the 3141 counties, excluding those in Puerto Rico, sampled persons were cross-classified by age group (20–44, 45–64, and ≥ 65 years), gender (male and female), and race/ethnicity (non-Hispanic white and other); county level sample size did not support a finer age or race/ethnicity classification. This cross-classification resulted in 12 classes per county per year. The number of people sampled in each class that have diabetes can be determined. Specifically, let n_{ijk} = the number of sampled people in county i, class $j = 1, \ldots, 12$, year $k = 1, 2, 3$ and y_{ijk} = the number of sampled people with diagnosed diabetes in county i, class j, year k. In some years, some counties have n_{ijk}'s $= 0$. For these, the corresponding y_{ijk}'s $= 0$.

The U.S. Census Bureau publishes population estimates by demographic characteristics (unit-level auxiliary information) for all counties; the Census provides no information on diabetes status. We used the 2005 Census county projections to obtain estimates for the number of persons in each age, sex and race group used to cross-classify the BRFSS data. Let, N_{ij} = the estimated number of people in county i, class $j = 1, \ldots, 12$, in 2005. Variability in Census projections was ignored.

2.2 Diabetes prevalence estimator

Estimating diabetes status for all adults 20 years of age or older in 2005 (roughly 200 million people) will allow estimation of 2005 diabetes prevalence in all U.S. counties. However, the number of people Y_{ij}, in county i, class j in 2005 with diagnosed diabetes is unobserved. For county i, the unobserved prevalence of diagnosed diabetes was:

$$p_i = \frac{\sum_{j=1}^{12} Y_{ij}}{\sum_{j=1}^{12} N_{ij}} \tag{2.1}$$

In some cases, N_{ij}'s $= 0$, with corresponding Y_{ij}'s $= 0$.

We combine the information from the Census and the BRFSS to predict diabetes status for all the people in the 2005 Census who did not participate in the 2005 BRFSS. Setting $Z_{ij} = (Y_{ij} - y_{ij2})$ = the total number of people in 2005 with diabetes in county i, class j who were not in the sample, the county prevalences in equation (2.1) can be expressed as

$$p_i = \frac{\sum_{j=1}^{12} (Z_{ij} + y_{ij2})}{\sum_{j=1}^{12} N_{ij}} \tag{2.2}$$

Eighty-eight sampled persons (20 with diabetes) from 20 counties belonged to a class where $N_{ij} < n_{ij2}$. In these cases, we set $y_{ij2} = 0$.

2.3 Probability model

We applied Bayesian multilevel models to each of the four Census regions (Northeast, South, Midwest and West). These models relate partially observed quantities to other variables of interest. In particular,

$$y_{ij\cdot} \sim \text{Poisson}(\lambda_{ij}); \quad i = 1, \ldots, c \;\text{ and }\; j = 1, \ldots, 12$$

where c is the number of counties in a given region and $\lambda_{ij} = n_{ij\cdot} p_{ij}$ With p_{ij} = the prevalence of diagnosed diabetes in county i, class j, $y_{ij\cdot} = \sum_{k=1}^{3} y_{ijk}$ and $n_{ij\cdot} = \sum_{k=1}^{3} n_{ijk}$. Now, $\log(\lambda_{ij}) = \log(n_{ij\cdot}) + \log(p_{ij}) = \log(n_{ij\cdot}) + \beta_{ij}$ and the full probability model is specified with the prior distributions described below. Let $s[i]$ denote the state s that contains county i. Defining the matrix R as having diagonal elements equal to estimates of the variances of the β_{ij}'s (defined below) and off diagonals equal to 0.001 (Gelfand *et al.*, 1990), we assign the prior distributions:

$$
\begin{aligned}
\beta_i &= (\beta_{i1}, \ldots, \beta_{i12})' \sim \mathrm{MVN}(\mu_{s[i],\Sigma}) \\
\mu_s &= (\mu_{s1}, \ldots, \mu_{s12})' \sim \mathrm{MVN}(\nu, \mathrm{diag}\{\psi_1, \ldots, \psi_{12}\}) \\
\Sigma^{-1} &\sim \mathrm{Wishart}((12R)^{-1}, 12) \\
\nu_j &\sim N(0, 10^4); \quad j = 1, \ldots, 12 \\
1/\psi_j &\sim \mathrm{Gamma}(1, 1)
\end{aligned}
\tag{2.3}
$$

The probability specification in the above model implies a proper joint posterior distribution for $(\beta, \mu, \Sigma, \nu, \psi | y, n)$, because all prior are proper. We used an informative prior for ψ_j to calibrate the models to state level observations. Posterior predictive distributions of diabetes prevalence among sampled persons by state and class were compared to observed values. The prior specification indicates little a priori information about parameters other than ψ_j and hence has little impact on the estimates.

2.4 Estimates of diabetes prevalence

Our prevalence estimates of diagnosed diabetes in each county are the means of the posterior predictive distributions of the p_i's:

$$
\begin{aligned}
\hat{p}_i &= E(p_i|y, n, N) = E\left(\frac{\sum_j (Z_{ij} + y_{ij2})}{\sum_j N_{ij}} |y, n, N\right) \\
&= \frac{\sum_j [E(Z_{ij}|y, n, N) + y_{ij2}]}{\sum_j N_{ij}}
\end{aligned}
\tag{2.4}
$$

From the probability model, the posterior predictive distributions of the Z_{ij}'s are

$$(Z_{ij}|y, n, N) \sim \mathrm{Poisson}(\gamma_{ij}); \quad \gamma_{ij} = (N_{ij} - n_{ij2}) \exp(\beta_{ij})$$

and the posterior predictive distributions of the Z_{ij}'s, and thus the p_i's, are determined by the posterior distributions $(\beta_{ij}|y, n)$. County level diabetes prevalence estimators can identify counties with extreme (high or low) diabetes burden relative to other counties. We identified extreme counties by generating the posterior predictive distribution of the ranks for each county (Spiegelhalter, *et al.*, 2004, chap. 7.4). We generated the posterior by combining each draw across counties and assigning ranks $(r_i : \ i = 1, \ldots, 3141)$ to p_i's in the descending order; because of the posterior's continuity, there were no ties and therefore no ambiguity in ranks. A county was defined as having low diabetes prevalence if $Pr(r_i < 1571) \geq 0.95$ and as having high diabetes prevalence if $Pr(r_i > 1571) \geq 0.95$, where 1,571 is the median rank.

All posterior distributions were simulated in WinBUGS (Lunn *et al.*, 2000). The 2.5th and 97.5th percentiles of the posterior distributions of the p_i's provided the 95% confidence intervals for the county prevalence of diagnosed diabetes. We used a burn-in of 1000 and then monitored a single chain for 10000 iterations.

2.5 Model checking

We implemented posterior predictive checks to examine the consistency of the model with the data (Gelman and Hill, 2007, chap 8.3). In posterior predictive checking, the entire data set is replicated for each posterior draw of the model parameters. A test quantity that reflects relevant aspects of the model is calculated for each replicate. Each value of the test quantity is subtracted from the observed data value. Using a boxplot, we plot the distribution of the subtracted values to compare the distribution of the test quantity with its observed value. The following test quantities were selected: percent of sample with diabetes by county, by state, by class by state and by race/ethnicity by state.

The BRFSS provides direct estimates of diabetes prevalence for 232 large population counties (Centers for Disease Control and Prevention, 2007). These estimates allow evaluation of model-based estimates against design-based direct estimates, which have long been the standard. For comparison, we generated random draws for each of the 232 counties by using a Normal distribution with mean and standard deviation equal to the direct estimate of the mean and its standard error for persons 20 years of age or older. For each county, the distribution of draws from the direct estimate was compared with the corresponding posterior predictive distribution. Specifically, we subtracted each draw from the posterior of the model-based estimate from the corresponding draw for the direct estimate. For each county, the percent of draws ≤ 0 were calculated. If this percentage was $> 97.5\%(< 2.5\%)$, the model-based estimates were identified as overestimates (underestimates) of the design-based estimates.

3. Results

3.1 Description of data

Table I provides details on the number of states and counties and the number of adults 20 years of age or older in the 2005 Census, 2005 BRFSS and 2004 – 2006 BRFSS by census region. The number of respondents in the BRFSS who self-reported diabetes is also given for each Census region. The BRFSS provides diabetes status for roughly 0.16% of the 2005 U.S. population 20 years of age or older. For counties with small sample sizes, direct estimation of diabetes prevalence is impractical; three of the four census regions contain at least one county with no BRFSS sample. For the 2909 counties for which direct estimates are impractical, small area estimation techniques can provide estimates. Our model uses the 964007 observations from the 2004 – 2006 BRFSS to predict diabetes status for the proportion of the 2005 population that was not sampled in the 2005 BRFSS.

Table 1: Descriptive statistics of the 2005 census and 2005 BRFSS by region.

	Census (2005)	BRFSS (2005)	BRFSS (2004-2006)
Northeast Census Region			
States	9	9	9
Counties	217	217	217
Adults 20+	40394742	67178	195443
Adults 20+ with Diabetes		5971	17167
Midwest Census Region			
States	12	12	12
Counties	1055	1049	1055
Adults 20+	47955432	74179	209766
Adults 20+ with Diabetes		6930	19170
South Census Region			
States	17	17	17
Counties	1423	1405	1422
Adults 20+	77790003	109726	319094
Adults 20+ with Diabetes		11812	34539
West Census Region			
States	13	13	13
Counties	446	444	444
Sample	48530978	88317	239704
Adults 20+ with Diabetes		7115	19607
Total			
States	51	51	51
Counties	3141	3115	3138
Sample	214671155	339400	964007
Adults 20+ with Diabetes		31828	90483

3.2 Estimates and maps

Estimates for all counties appear on the CDC Web site (2008b). Prevalence estimates of diabetes were 2.98% – 14.83% (mean, 8.63%). The average width for 95% confidence intervals was 4.35% (range, 0.91% – 20.84%).

We classified 908 (29%) of the counties as having either a high or low diabetes burden. Although ranks were estimated for the remaining counties, not all counties were classified. Some may truly have ranks near the median. Others may have a true rank far from the median, but a confidence interval for the rank was too wide to meet our criterion for extreme prevalence.

428 counties met our criterion for high diabetes burden. These counties were primarily located in northern Maine, West Virginia, coastal North Carolina and

South Carolina, middle Georgia, southern Alabama and Mississippi. In addition, the eastern part of Oklahoma, parts of South Dakota and a few counties in Montana and New Mexico also met the criterion for high diabetes burden (Figure 1). The distribution of point estimates for diabetes prevalence in these 428 counties was right skewed and unimodal; the median was 11.29% (range, 9.34% – 14.83%). The confidence interval width for these counties ranged from 1.75% to 9.10%. The percentage of counties categorized with a high diabetes burden varied by region: Northeast, 0.28%; Midwest, 1.99%; South, 27.76%; and West, 1.35%.

Areas with low diabetes burden were the Rocky Mountain states, eastern Minnesota, the coastal counties of Alaska, some islands of Hawaii, and some northeastern states (Figure 1). The distribution of point estimates for the 480 low prevalence counties was left skewed and unimodal; the median was 6.30% (range, 2.98% – 7.82%). The confidence interval width for the prevalence estimates ranged from 0.91% to 5.39%. The percentage of counties with a low diabetes burden varied considerably: Northeast, 31.80%; Midwest, 13.65%; South, 2.81%; and West, 50.90%.

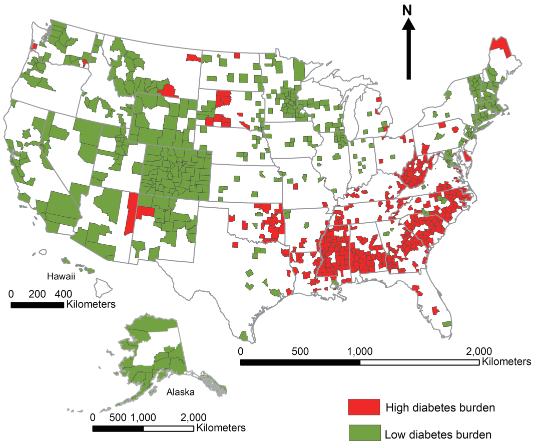

Figure 1: 2005 U.S. counties with extreme diabetes burden in adults aged 20 years or older. Counties with high (low) diabetes burden are those for which at least 95% of the posterior distribution for the rank was above (below) the median. Counties in white are unclassified

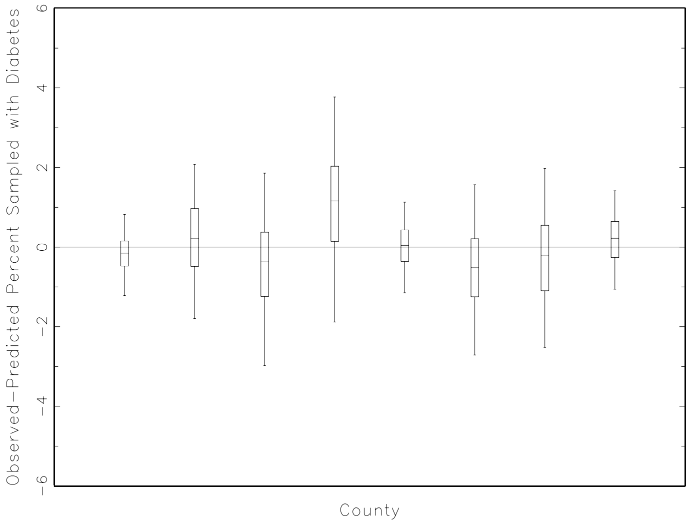

Figure 2: Posterior distribution of the difference between the observed and predicted percent sampled with diabetes for counties in Alaska, 2005.

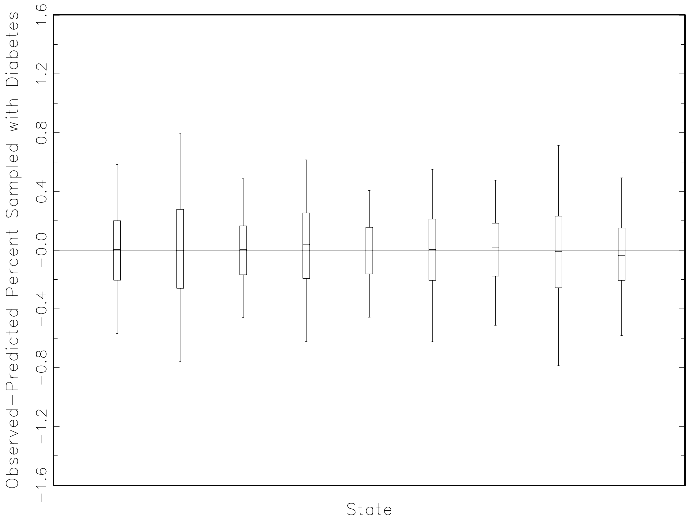

Figure 3: Posterior distribution of the difference between the observed and predicted percent sampled with diabetes for states in the northeast region, 2005.

3.3 Model checking

Figure 2 displays boxplots for the distribution of the test quantity, percent with diabetes, for Alaska. Alaska was chosen because the number of counties was small enough for the results to be readily interpreted visually. Each boxplot represents one of the 27 counties in Alaska. Boxplots which overlap 0 indicate lack of evidence for a difference between the observed percentage of the sample and the predictive distribution. For all states (data other than Alaska not shown), the boxplots indicated consistency between the model and the data.

The distribution of the test quantity, percent with diabetes by state, for the Northeast region, an arbitrary choice, is displayed in the boxplot in Figure 3. Each boxplot represents one of the nine states in the Northeast. The boxplots indicate consistency between the model and the data. Boxplots for all other regions led to similar conclusions (data not shown). Boxplots for the other test quantities, percent with diabetes by class and state and percent with diabetes by race and state, also indicate data/model consistency (data not shown).

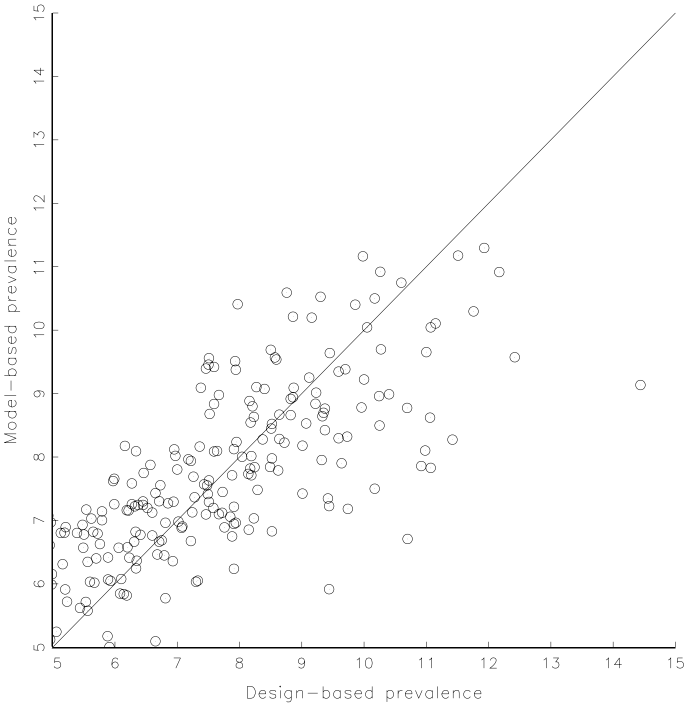

Figure 4: Scatterplot of design-based BRFSS diabetes prevalence estimates versus model-based diabetes prevalence estimates for U.S. counties with large sample size, 2005.

A scatterplot of the design-based estimates versus the model-based estimates for the counties for which design-based BRFSS estimates were available is displayed in Figure 4. Comparisons of our modeled estimates with the design-based estimates indicated that both have considerable variability, but the model-based estimates have less. Once this variability was accounted for, the estimates largely agreed. Five (2.2%) of the 232 counties did not show agreement. Three counties had higher and two counties had lower model-based estimates than direct estimates in more than 97.5% of the draws. We would expect 5% to lack agreement due to sampling variability. Therefore, we find no evidence that model-based and direct estimates were inconsistent.

4. Discussion

We produced estimates of diabetes prevalence for all 3141 counties in the United States. These estimates are currently available on CDC's web site (2008b). Our model is an expansion of the model proposed by Malec *et al.* (1997). The multi-level modeling approach allows the "effect" of age, race/ethnicity and sex on prevalence to vary by county. We view this as a major strength compared to synthetic estimators. Our model-based approach relies on combining multiple years of data. We investigated using a single year of data but chose to combine three years. Combining years increases precision but adds little bias, because the increase in diabetes prevalence is thought to be roughly linear; negative bias from using 2004 data for 2005 and positive bias from using 2006 data for 2005 roughly cancel. We also fit independent models in each of four Census regions, because running a single U.S. model was resource and time intensive using available software. A single model required from two to nine hours to run using WinBUGS (Lunn *et al.*, 2000). Our approach limited the borrowing of information to counties within the region. Our models have a random effect for state, allowing counties within the same state to be more similar than counties across state boundaries. Since we are using a model-based approach we performed extensive checking to validate the models.

Reduction of the diabetes burden through prevention programs is among the goals for *Healthy People 2010*, which defines the U.S.'s public health goals (U.S. Department of Health and Human Services, 2000). The objectives for *Healthy People 2010* include educating people with diabetes, preventing diabetes and increasing the percent of adults with diabetes who are diagnosed. County prevalence estimates of diabetes can be used, alongside other information, to identify areas of the country where prevention programs might have the greatest impact toward reducing the diabetes burden.

Having an estimate and standard deviation for the diabetes prevalence for every county in the U.S. is useful. However, knowing how a county's estimate compares to other counties is imperative to the success of prevention programs targeting specific areas. Using only the point estimate to identify counties with extreme burden might be appealing to many, but is a flawed approach. Very small changes in point estimates can lead to large changes in a county's rank. One approach for comparing county estimates is to determine if the 95% confidence interval for the estimate overlaps some standard value above or below which the burden would be considered extreme. Intervals that do not overlap this standard are considered extreme. With this approach, we can expect 5% of the counties to be extreme even if prevalence for all counties were at the standard level. We estimated the distribution of the rank for each county and are therefore highly confident about the counties identified with an extreme burden, although considerable uncertainty remains about some counties not identified as extreme. While we focused on estimates where 95% of the posterior distribution for the rank was above or below the median, other definitions could be readily applied. For example, we could have identified average counties for which 95% of the rank's posterior distribution was completely between the first and third quartiles of the ranks.

Counties with high diabetes burden may benefit more from targeted prevention programs than from a "one size fits all" national or state program. For example, CDC's state specific estimates indicate that the prevalence of diagnosed diabetes in Alabama and North Carolina is $> 8\%$ (Centers for Disease Control and Prevention, 2007). Our map shows areas within these states that are likely contributing to this high prevalence: the southern part of the Alabama and the coastal counties of North Carolina.

County level estimates allow for exploration of ecological relationships between diabetes and county level explanatory covariates. While inherently limited, ecologic analyses can offer insight. For example, the Social Science Data Analysis Network provides a map of the percentage of a county's residents in poverty based on 2000 U.S. Census data (University of Michigan, 2008).Visual comparison with our map indicates the counties with high diabetes burden (southern Alabama, middle Georgia, Mississippi, and West Virginia) coincide with counties with a large percent of residents in poverty. Pickle and Su provide a map of the proportion of residents within a county who are at risk of obesity (Pickle and Su, 2002). There is concordance between counties with the lowest diabetes burden and those with the lowest risk of obesity — the New England states of Connecticut, Massachusetts, New Hampshire, and Vermont; Colorado; and many counties in the West. CDC provides county estimates of the stroke death rates for adults aged 35 years or older (Centers for Disease Control and Prevention, 2008c). High stroke death rates are seen in the southeast part of the country, where the diabetes burden is great. Apache County (Arizona), Big Horn and Roosevelt Counties (Montana), McKinley County (New Mexico), and the South Dakota counties of Carson, Dewey, Jackson, Mellette, Shannon, Todd, and Ziebach, all with high diabetes burden, have a large percent of land area designated as Indian reservations and/or a large population of American Indians (National NAGPRA, 2008; U.S. Census Bureau, 2001). Diabetes prevalence among American Indians is known to be high.

Visual exploration of maps is a useful tool for generating hypotheses. By coupling point estimates of diabetes prevalence with measures of variability, models could be built to formally investigate ecological relationships between county level covariates and diabetes burden. In particular, one could examine the impact of county level explanatory factors after controlling for characteristics known to be associated with high or low diabetes prevalence.

Our estimates should be considered in light of the limitations of the method applied. Extreme diabetes burden was determined by the posterior distribution for the ranks. These ranks are based on unadjusted (crude) estimates. For example, a county could have an extreme diabetes burden and a larger than average elderly population; diabetes prevalence tends to increase with age. In this case, a high prevalence might not be higher than that accounted for by the county's age distribution. Our model only accounts for diagnosed diabetes. Although the fraction of all undiagnosed cases on a national level is well understood, little is known about the geographic variability of undiagnosed diabetes.

Areas for future research are the addition of spatial effects and county level covariates. Currently, we do not consider an explicit spatial effect that accounts for similarity of adjacent counties. Alternative models, such as conditional autoregressive models, would account for such spatial effects and might offer a competing model. We did not include county level covariates beyond the state in which the county was located. Several county level variables are available through the U.S. Census and could potentially improve model fit. However, including county level covariates in estimates would limit the ability to consider these variables as explanatory variables in ecological analyses.

The model-based approach provides estimates that are highly likely to be useful to national, state, and local public health agencies. Only through such combined efforts can we hope to achieve goals of *Healthy People 2010* and "ensure that good health, as well as long life, are enjoyed by all" (U.S. Department of Health and Human Services, 2000).

Acknowledgments

The authors would like to acknowledge Roberta Hilsdon, Perot Systems for her assistance with data preparation and data management and Randall Young, Agency for Toxic Substance and Disease Registry, for producing Figure 1.

References

American Diabetes Association. (2008). Economic costs of diabetes in the U.S. in 2007. *Diabetes Care* **31**, 1-20.

Boyle, J. P., Honeycutt, A. A., Narayan, K. M. V., Hoerger, T. J., Geiss, L. S., Chen, H. and Thompson, T. J. (2001). Projection of diabetes burden through 2050. *Diabetes Care* **24**, 1936-1940.

Centers for Disease Control and Prevention. (2007). Surveillance for certain health behaviors among states and selected local areas-United States, 2005. *MMWR.* **56**, 1-160.

Centers for Disease Control and Prevention. (2008a). Behavioral Risk Factor Surveillance System survey data. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Retrieved September 22, 2008, from Web site: http://www.cdc.gov/BRFSS/.

Centers for Disease Control and Prevention. (2008b). National Diabetes Surveillance System. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Retrieved June 30, 2008, from Web site: http://www.cdc.gov/diabetes/statistics/index.htm.

Centers for Disease Control and Prevention. (2008c). Stroke Fact Sheet. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Falorsi, P. D., Falorsi, S. and Russo, A. (1994). Empirical comparison of small area estimation methods for the Italian Labour Force Survey. *Survey Methodology* **20**, 171-176.

Fay, R. E. and Herriott, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269-277.

Gelfand, A. E., Hills, S .E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972-985.

Gelman, A., Carlin, B. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis.* 2nd ed. Chapman and Hall.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**, 55-76.

Lawson, A. B., Biggeri, A. B., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P. and Divino F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine* **19**, 2217-2241.

Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* **10**, 325-337.

MacNab, Y. (2003). Hierarchical Bayesian spatial modeling of small-area rates of nonrare disease. *Statistics in Medicine* **22**, 1761-1773.

Malec, D., Sedransk, J., Moriarity, C. L. and LeClere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association* **92**, 815-826.

Malec, D., William, W. D. and Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine* **18**, 3189-3200.

Narayan, K. M. V., Boyle, J. P., Thompson, T. J., Sorensen, S. W. and Williamson, D. F. (2003). Lifetime risk for diabetes mellitus in the United States. *JAMA* **290**, 1884-1890.

National NAGPRA. Indian reservations in the continental United States. Washington, D.C.: National Park Service. (n.d.). Retrieved September 11, 2008, from Web site: http://www.nps.gov/history/nagpra/documents/resmap.htm.

Pickle, L. W. and Su, Y. (2002). Within-state geographic patterns of health insurance coverage and health risk factors in the United States. *American Journal of Preventative Medicine* **22**, 75-83.

Purcell, N. J. and Kish, L. (1979). Estimates for small domain. *Biometrics* **35**, 365-384.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., and Feuer, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association* **102**, 474-486.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley.

Sarndal, C. E. (1984). Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association* **79**, 624-631.

Spiegelhalter, D. J., Abrams, K. R. and Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Wiley.

U.S. Census Bureau. The American Indian population: 2000. Washington, D.C.: U.S. Census Bureau. (2001). Retrieved September 11, 2008, from Web site: http://www.census.gov/prod/2001pubs/mso01aian.pdf.

U.S. Department of Health and Human Services. (2000). *Healthy People 2010: Understanding and Improving Health.* 2nd ed. U.S. Government Printing Office.

University of Michigan. Social Science Data Analysis Network. Ann Arbor, MI: University of Michigan. (n.d.) Retrieved September 11, 2008, from Web site: http://www.censusscope.org/us/map_poverty.html.

Xie, D., Raghunathan, T. E. and Lepkowski, J. M. (2007). Estimation of the proportion of overweight individuals in small areas-a robust extension of the Fay-Herriot model. *Statistics in Medicine* **26**, 2699-2715.

Received April 30, 2009; accepted August 26, 2009.

Betsy L. Cadwell
Division of Diabetes Translation
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
1600 Clifton Road
Mailstop E92, Atlanta, GA 30333, USA
BCadwell@cdc.gov

Theodore J. Thompson
Division of Diabetes Translation
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
4770 Buford Highway, NE
Mailstop K10, Atlanta, GA 30341, USA
TThompson@cdc.gov

James P. Boyle
Division of Diabetes Translation
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
4770 Buford Highway, NE
Mailstop K10, Atlanta, GA 30341, USA
JBoyle@cdc.gov

Lawrence E. Barker
Division of Diabetes Translation
National Center for Chronic Disease Prevention and Health Promotion
Centers for Disease Control and Prevention
4770 Buford Highway, NE
Mailstop K10, Atlanta, GA 30341, USA
LBarker1@cdc.gov