Bayesian Small Area Estimates of Diabetes Incidence by United States County, 2009

Lawrence E. Barker*, Theodore J. Thompson, Karen A. Kirtland,
James P. Boyle, Linda S. Geiss, Mary M. McCauley and Ann L. Albright
Centers for Disease Control and Prevention

Abstract: In the United States, diabetes is common and costly. Programs to prevent new cases of diabetes are often carried out at the level of the county, a unit of local government. Thus, efficient targeting of such programs requires county-level estimates of diabetes incidence−the fraction of the non-diabetic population who received their diagnosis of diabetes during the past 12 months. Previously, only estimates of prevalence−the overall fraction of population who have the disease−have been available at the county level. Counties with high prevalence might or might not be the same as counties with high incidence, due to spatial variation in mortality and relocation of persons with incident diabetes to another county. Existing methods cannot be used to estimate county-level diabetes incidence, because the fraction of the population who receive a diabetes diagnosis in any year is too small. Here, we extend previously developed methods of Bayesian small-area estimation of prevalence, using diffuse priors, to estimate diabetes incidence for all U.S. counties based on data from a survey designed to yield state-level estimates. We found high incidence in the southeastern United States, the Appalachian region, and in scattered counties throughout the western U.S. Our methods might be applicable in other circumstances in which all cases of a rare condition also must be cases of a more common condition (in this analysis, "newly diagnosed cases of diabetes" and "cases of diabetes"). If appropriate data are available, our methods can be used to estimate proportion of the population with the rare condition at greater geographic specificity than the data source was designed to provide.

Key words: Bayesian estimates, diabetes, small area estimates.

1. Introduction

In the United States, diabetes has an enormous social cost (American Diabetes Association, 2008). Persons with diabetes are at increased risk of devastating

---

*Corresponding author.

complications that include blindness (Kertes and Johnson, 2007), hearing loss (Kakarlapudi *et al.*, 2003), and kidney failure (Centers for Disease Control, 2011), and a person with diabetes is at about twice the risk of death as a person of similar age without diabetes (Centers for Disease Control, 2011). Having diabetes approximately doubles one's medical costs (Centers for Disease Control, 2011).

Diabetes is common, and becoming more so. In 2010, the national incidence (fraction of the population diagnosed in the preceding twelve months) was 8.6 per 1000 person years and the national prevalence (fraction of population who have the condition) of diagnosed diabetes was 6.0% (Centers for Disease Control and Prevention, 2012a). By 2050, the national diabetes incidence is projected to be 15 per 1000 and the national prevalence is projected to be at least 14.8% (Boyle *et al.*, 2010). A child born in 2000 in the United States has a more than 1 in 3 risk of developing diabetes sometime in his or her life (Narayan *et al.*, 2003).

Programs to prevent new cases of diabetes and to prevent complications and diabetes mortality are generally conducted locally, often at the level of the county (a unit of local government in the United States) or county equivalent (hereafter "county"). State-level estimates of the prevalence of diagnosed diabetes have long existed, based on data from the Behavioral Risk Factor Surveillance System (BRFSS), an ongoing, state-based, random-digit-dialed telephone survey of non- institutionalized adults aged > 18 years in all 50 states, the District of Columbia, and some territories (Centers for Disease Control and Prevention, 2012b). County-level prevalence estimates, pinpointing the counties with the greatest need for programs to prevent complications and death among persons with diabetes, have recently been developed, using model-based small area estimation techniques (Cadwell *et al.*, 2010, Congdon and Lloyd, 2010, Srebotnjak, Mokdad and Murray, 2010).

Diabetes incidence and prevalence are obviously related; a person who develops diabetes, with rare exceptions, has the condition for life. However, incidence might or might not be high in areas of high prevalence, with differences stemming from spatial variability in mortality among persons with diabetes (a county with high incidence could have low prevalence, if mortality among persons with diabetes was large) and in mobility; for example, if incident cases in county A relocate to county B, they influence the incidence rate in A but not B, and the prevalence rate in B but not A.

To effectively target programs to prevent diabetes, we must identify areas of high incidence. However, historically, only state-level estimates of diabetes incidence have been available, and only for states that opted to conduct more in-depth interviews of persons with diabetes to determine when they were diagnosed.

Here, expanding on the work of Cadwell *et al.* (2010), in which BRFSS data were used to estimate county-level prevalence, we provide the first estimates of

diabetes incidence for all 3,143 U.S. counties. Our underlying method is based on noticing that all incident cases of diabetes must also be prevalent cases. We first estimate county-level prevalence. We then estimate the county-level proportion of prevalent cases that are incident cases. Multiplying these together yields an estimate of incidence.

2. Methods

2.1 Data Sources

We use BRFSS data to model diabetes incidence. The BRFSS is a collection of cross-sectional random-digit-dialing telephone surveys, conducted in all U.S. states and the District of Columbia. Every year, states conduct monthly telephone surveillance using a standardized questionnaire to determine the distribution of risk behaviors and health practices among noninstitutionalized adults. We use U.S. Census data to post-stratify the modeled estimates to be representative of U.S. counties in 2009.

County of residence for BRFSS respondents, while not included in the all versions of the data, was obtained for this study. For respondents with missing county of residence, we used the county most likely associated with the respondent's telephone number. To make inferences from the sample to the population, post-stratification weights were attached to estimates. This use of post-stratification weights guarantees that our estimates are consistent with U.S. Census estimates of county population size.

For respondents indicating that they had diagnosed diabetes, an optional diabetes module was administered in some states in some years. On average, 41 states administered the diabetes module during 2008 through 2010, the years of data we used (combining three years of data follows the precedent set by Cadwell *et al.*, and this approach was chosen because using fewer years resulted in some unstable estimates, due to smaller sample size). This results in some states having one year of data, some two years, and some three. One state had no diabetes module data.

The U.S. Census Bureau publishes population estimates by demographic characteristics (unit-level auxiliary information) for all counties (U.S. Census, 2012); the Census provides no information on diabetes status. We used the 2009 U.S. Census county projections to obtain estimates for the number of persons in each of the classes into which we group BRFSS data.

2.2 Diabetes Information Gathered

All BRFSS respondents were asked, "Have you ever been told by a doctor

that you have diabetes?" Those who responded "no" were considered to not have diagnosed diabetes. Males who responded "yes" were considered to have diabetes. Females who responded "yes" were asked if they were told only during pregnancy. Those who indicated this had happened were considered have had gestational diabetes, which resolves at delivery, and were not considered to have diabetes. Females who had been told at any other time were considered to have diabetes. Respondents who did not know or refused to answer the question were considered to have missing diabetes status.

Those who indicated that they had diagnosed diabetes and resided in states that administered the optional diabetes module were asked, "How old were you when you were told you have diabetes?" Age at the time of the survey and age at diagnosis, both self-reported, are measured in years and used to estimate the number of incident cases (2012a). If the difference between a respondent's age and age at diagnosis is two or more years, the person is not considered an incident case. If the difference is zero years, the person is considered an incident case. If the difference is one year, the person is weighted as half an incident case; this method has been used previously to estimate diabetes incidence (Centers for Disease Control and Prevention, 2012a).

2.3 Incidence Modeling with Missing Data

The general framework for our model-based approach includes multilevel modeling followed by post-stratification (Gelman and Hill 2007, Chapter 14). Because of missing data, we use an approach analogous to "factoring the likelihood" (Little and Rubin 2002, Chapter 7). We specify a full Bayesian probability model that allows the posterior distribution to be factored into two parts that are estimated separately. Unlike the perhaps more familiar complete case analysis, this methods uses all the information on diabetes in the responses for which diabetes incidence status is missing.

Let $Z_1 = 1$ if an individual was diagnosed with diabetes within the past year; $Z_1 = 0$ otherwise. Let $Z_2 = 1$ if an individual was diagnosed with diabetes; $Z_2 = 0$ otherwise. Since the distribution of prevalence given incidence, $p(Z_2|Z_1)$, equals one, the distribution of incidence, $p(Z_1)$, can be written as $p(Z_1, Z_2) = p(Z_1|Z_2)p(Z_2)$. We develop separate models for incidence conditioned on the person being a prevalent case, $p(Z_1 |Z_2)$, and for prevalence, $p(Z_2)$. The estimates from the two models are combined to obtain an estimate of incidence, $p(Z_1)$. Assuming the parameters in the conditional incidence model and the prevalence model are distinct allows the distribution of incidence to be split into these two parts. See Appendix 1 for details.

For each of the 3,143 U.S. counties, sampled persons were cross-classified by age group (20-44, 45-64, 65+ years), sex (male, female) and race/ethnicity

(non-Hispanic white, other); sample sizes did not support a finer division of race/ethnicity. This results in 12 classes per county. The number of people sampled in each class who have diabetes can be determined. Specifically, let n_{ij} = the number of sampled people in county i, class $j = 1, \cdots, 12$, and y_{ij}= the number of sampled people with diagnosed diabetes in county i, class j. In some years, in some counties, $n_{ij} = 0$. For each of these, the corresponding $y_{ij} = 0$. Let $y_{ij}*$ = the number of sampled people with diagnosed diabetes in county i, class $j = 1, \cdots, 12$, whose incident status is known and r_{ij} = the number of sampled people with incident diabetes in county i, class j.

2.4 Prevalence Model

We fit a Bayesian multilevel model to the three years of BRFSS data, using methods similar to Cadwell *et al.*, 2010, but differing in several ways (e.g., we used a Binomial distribution with a logit link instead of the Poisson distribution and logarithmic link that Cadwell *et al.* used). Our model and Cadwell *et al.*'s model yield similar, but not identical, estimates. Our changes reduce the number of approximations (e.g., Cadwell *et al.* used the Poisson distribution to approximate the Binomial, while we used the Binomial).

Our model relates observed quantities to the twelve classes. In particular,

$$y_{ij} \sim \text{Binomial}(p_{ij}, n_{ij}); \quad i = 1, \cdots, 3143, \quad j = 1, \cdots, 12, \qquad (2.1)$$

where p_{ij} = the prevalence of diagnosed diabetes in county i, class j. Let $s(i)$ denote the state s that contains county i. The regression model includes:

(a) logit link function $\log(p_{ij}/(1 - p_{ij}))$.

(b) fixed effect intercept for each class (age by sex by race/ethnicity group) α_{p_j}; $j = 1, \cdots, 12$.

(c) random effects by county and class $\mu_{p_{ij}}$; $i = 1, \cdots, 3143, j = 1, \cdots, 12$.

(d) random effects by state and class $\nu_{p_{s(i)j}}$; $s(i) = 1, \cdots, 51, j = 1, \cdots, 12$.

(e) spatial effects by state and class $\omega_{p_{s(i)j}}$; $s(i) = 1, \cdots, 51, j = 1, \cdots, 12$.

Parameters under (c) and (d) are modeled via 12-dimensional multivariate normal priors (Gelfand, Hills, Racine-Poon and Smith, 1990). Parameters under (e) are modeled via 12-dimensional multivariate normal conditional autoregressive priors (Besag, York and Mollie, 1991). The subscript "p" is used to indicate prevalence. Thus the regression model is

$$\text{logit}(p_{ij}) = \alpha_{p_j} + \mu_{p_{ij}} + \nu_{p_{s(i)j}} + \omega_{p_{s(i)j}}. \qquad (2.2)$$

To assess this extended model, we consider a basic model (Rao, 2003) as a benchmark. The basic model includes fixed effects for class and a spatially unstructured random effect for county. The basic regression model is

$$\text{logit}(p_{ij}) = \alpha_{p_j} + \varepsilon_i. \tag{2.3}$$

All prior distributions appear in Appendix 2.

2.5 Conditional Incidence Model

The conditional incidence model is similar to the prevalence model:

$$r_{ij} \sim \text{Binomial}(q_{ij}, y_{ij}^*); \; i = 1, \cdots, 3143, \quad j = 1, \cdots, 12, \tag{2.4}$$

where $q_{ij} =$ the probability of incident diabetes among people with diabetes in county i, class j. Let $s(i)$ denote the state s that contains county i. The regression model includes:

(a) logit link function $\log(q_{ij}/(1 - q_{ij}))$.

(b) fixed effect intercept for each class (age by sex by race/ethnicity group) α_{c_j}; $j = 1, \cdots, 12$.

(c) random effects by county and class $\mu_{c_{ij}}$; $i = 1, \cdots, 3143$, $j = 1, \cdots, 12$.

(d) random effects by state and class $\nu_{c_{s(i)j}}$; $s(i) = 1, \cdots, 51$, $j = 1, \cdots, 12$.

(e) spatial effects by state and class $\omega_{c_{s(i)j}}$; $s(i) = 1, \cdots, 51$, $j = 1, \cdots, 12$.

The subscript "c" is used to indicate "conditional incidence", and to clearly distinguish these parameters from the ones in the prevalence model. As before, parameters under (c) and (d) are modeled via 12-dimensional multivariate normal priors. Parameters under (e) are modeled via 12-dimensional multivariate normal conditional autoregressive priors. Thus the regression model is:

$$\text{logit}(q_{ij}) = \alpha_{c_j} + \mu_{c_{ij}} + \nu_{c_{s(i)j}} + \omega_{c_{s(i)j}}. \tag{2.5}$$

We again consider a basic model as a benchmark. The basic model includes fixed effects for class and a random effect for county. The basic regression model is

$$\text{logit}(q_{ij}) = \alpha_{c_j} + \varepsilon_i. \tag{2.6}$$

All prior distributions appear in Appendix 2.

2.6 Estimates of Diabetes Incidence

Let $N_{ij} =$ the estimated number of people in county i, class $j = 1, \cdots, 12$,

in 2009 from U.S. Census Bureau. Let $\tilde{p}_{ij}$ and $\tilde{q}_{ij}$ be draws from the posterior distributions of p_{ij} and q_{ij}, respectively. Our estimate of the number of incident diabetes cases in county i is the mean of the posterior predictive distribution of:

$$\sum_j \tilde{p}_{ij}\tilde{q}_{ij}N_{ij}. \tag{2.7}$$

The estimate of the annual incidence rate in county i is the mean of the posterior predictive distribution of:

$$\frac{\sum_j \tilde{p}_{ij}\tilde{q}_{ij}N_{ij}}{\left[\sum_j N_{ij} - \sum_j \tilde{p}_{ij}(1 - \tilde{q}_{ij})N_{ij}\right]}. \tag{2.8}$$

Note that the denominator in the above expression is an estimate of the number of nondiabetic persons in county i at the start of 2009. The estimate of the age specific number of incident diabetes cases in county i, age group k is the mean of the posterior predictive distribution of:

$$\sum_{j \in A_k} \tilde{p}_{ij}\tilde{q}_{ij}N_{ij}, \tag{2.9}$$

where A_k is the set of classes for age group k. The estimate of the age specific annual incidence rate in county i, age group k is the mean of the posterior predictive distribution of:

$$\hat{p}_{ik} = \frac{\sum_{j \in A_k} \tilde{p}_{ij}\tilde{q}_{ij}N_{ij}}{\left[\sum_{j \in A_k} N_{ij} - \sum_{j \in A_k} \tilde{p}_{ij}(1 - \tilde{q}_{ij})N_{ij}\right]}. \tag{2.10}$$

Then the age-adjusted annual incidence rate for county i is the mean of the posterior predictive distribution of:

$$\hat{p}_i = \sum_k w_k \hat{p}_{ik}, \tag{2.11}$$

where w_k is the proportion of the U.S. population in age group k as measured by the 2000 U.S. Census.

All posterior distributions were simulated in WinBUGS (Lunn *et al.*, 2000). The 2.5th and 97.5th percentiles of the posterior distributions provide the 95% posterior intervals for county incidence of diagnosed diabetes. We treat the posterior intervals as confidence intervals. We run two chains, each with a burn-in of 1000 iterations followed by an additional 12,000 iterations. We check convergence by visual inspection of history and autocorrelation plots and using the Gelman-Rubin statistic, as modified by Brooks and Gelman (1998).

2.7 Model Comparisons

We evaluate the frequentist properties of the Bayesian estimates that we calculate using the model selection criterion developed in Gelfand and Ghosh (1998). This criterion depends on a statistic D, which combines goodness-of-fit and variability. Calculating D requires replicating the entire data set for each posterior draw of the model parameters. Using these replicates, we compute the posterior predictive mean and variance for each observation. We calculate G, a goodness of fit measure, as the sum over all observations of the squared difference between the data and its posterior predictive mean. We calculate P, a measure of expected mean square measure, as the sum over all observations of the posterior predictive variances. The statistic D is defined as G+P. Models with smaller D values, meaning that modeled values are more similar to observations, are preferred.

2.8 Model Checking

We implemented posterior predictive checks to examine the consistency of the model with the data (Gelman *et al.*, 2004, Chapter 6.3). In posterior predictive checking, the entire data set is replicated for each posterior draw of the parameters. A discrepancy or test measure that reflects relevant aspects of the model is calculated for each replicate. We used the deviance (McCullagh and Nelder, 1989, Chapter 2.3) as the discrepancy measure. A Bayesian *p*-value associated with the test measure is calculated; Bayesian *p*-values differ from the more familiar frequentist *p*-values in that a value between 0.1 and 0.9 indicates reasonable model fit.

We check our model by comparing it to direct estimates. The BRFSS provides direct, design-based, complete-case estimates of diabetes incidence for 48 states and DC for 2009 (Centers for Disease Control and Prevention, 2012a). These estimates are based on the combined 2008, 2009, 2010 BRFSS data. We aggregate our modeled county estimates to the state level and plot them against the direct estimates, allowing a comparison of modeled and direct estimates.

In an additional model check, we set up a simulation study to evaluate the two-part, prevalence and conditional incidence, modeling approach. Fifteen years of BRFSS data (1996-2010) were combined to create the test population. This test population includes 3,183,766 people, 285,851 with diabetes and 29,048 with incident diabetes. This population is too small to evaluate our model at the county level so we use the fifty states and DC as our "small areas". The evaluation proceeds as:

1. Sample 20,000 individuals from our population. This approximates the average number of observation per county in three years of BRFSS times 51.

2. Fit the prevalence and conditional incidence models to obtain incidence estimates by the 51 small areas. These models exclude the μ_{ij} terms.

3. Calculate direct design-based incidence estimates for the 51 small areas.

4. Repeat steps 1, 2 and 3 a total of fifty times.

5. Compare the modeled estimates to the true values from the test population.

6. Compare the direct design-based estimates to the true values from the test population.

Comparisons statistics in step 5 include the difference between the modeled and true values and the root mean squared error divided by the true incidence. Values of the root mean squared error divided by true value less than 0.25 are considered evidence of useful estimates. We generate box plots to visually present the dispersion in the difference between the modeled and true values. In step 6, we calculate the root mean squared error of the design-based estimates divided by the true incidence.

3. Results

3.1 Description of Data

Table 1 provides details (BRFSS data unweighted in this table, since we describe the sample and, at this stage, do not yet make inference to the population) on the number counties and the number of adults 20 years of age or older enumerated in the 2009 Census and surveyed in the 2008-2010 BRFSS. The number of respondents who self-reported diabetes and whose incident status is known also appears. Twelve percent of the BRFSS sample participants reported diagnosed diabetes. Of those, 73% have known incidence status.

Table 1: Descriptive statistics of the Census and BRFSS data

		BRFSS 2008-2010	
	Census 2009	Diabetes Prevalence	Conditional Incidence
Counties	3143	3140	2981
Responses	-	152,391	9676
median (min, max) per county	-	18 (0, 13000)	1 (0, 88)
Observations	223,585,859	1,255,029	110,765
median (min, max) per county	18885 (36, 7057285)	142 (2, 11967)*	14 (1, 1127)*

*Among counties with greater than zero observations.

3.2 Estimates

There was relatively little variation in county-level conditional incidence (proportion of prevalent cases which are also incident cases); estimated county-level conditional incidence ranged from 0.07 to 0.12, with a median value of 0.10. We looked for spatial patterns in conditional incidence, such as geographic clustering or a tendency to correlate with some demographic, but found no evidence of such patterns (data not shown).

Estimates for incidence for all counties appear in Figure 1. Incidence estimates of diabetes were 0.38%-2.1% (mean 1.1%). The average coefficient of variation was 18% (range, 8%-26%). We also calculate age-adjusted estimates (standardized to the 2000 U.S. population). These differed little from non-age-adjusted ones, and are not reported.
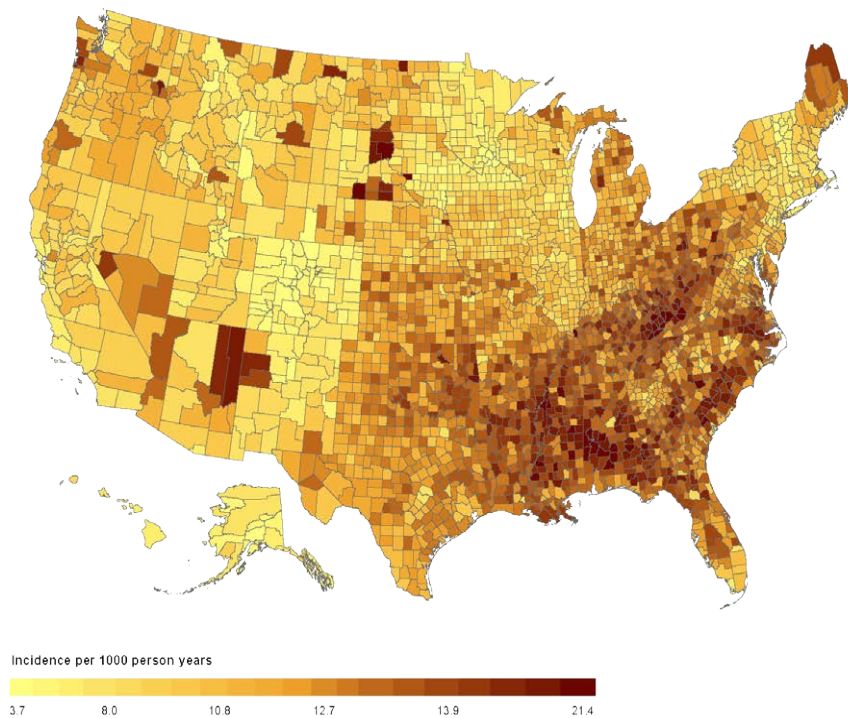

Figure 1: Map of 2009 diabetes incidence for 3143 U.S. counties

3.3 Model Comparisons

Table 2 displays G, P, and D, the model selection criteria. For both prevalence and conditional incidence, the basic model provides a much poorer fit than its

competitors. For both prevalence and conditional incidence, the extended model without the spatial effects provides a slightly better fit than the model including the spatial effects. Spatial effects require substantially longer simulation runs for little or no gain. Therefore, we did not include spatial effects in our final estimates.

Table 2: Model selection criterion

Outcome	Model	Goodness of Fit G	Predictive Error P	D*
Prevalence	Basic	153,335	143,990	297,225
	Extended	65,368	175,465	240,833
	Extended, without spatial	66,310	174,444	240,755
Conditional Incidence	Basic	7084	8765	15,849
	Extended	5594	9783	15,377
	Extended, without spatial	5635	9727	15,362

*Smaller is better.

The lack of spatial effects in the final model could be viewed as counter-intuitive, since the prevalence of diabetes in the United States (Barker *et al.*, 2011) exhibits strong spatial correlation. However, age, race/ethnicity, and state of residence, factors explicitly considered in our model, are also spatially correlated. These factors also serve as surrogates for factors that directly contribute to the incidence of type 2 diabetes (90-95% of diabetes in the United States [Centers for Disease Control, 2011]), such as obesity, physical inactivity, and lower income/ educational achievement. Thus, it is likely that the spatial autocorrelation that remained, after accounting for age, race/ethnicity, and state of residence, was small.

3.4 Model Checking

All models considered had Bayesian p-values between 0.1 and 0.9 (minimum 0.36, maximum 0.81), indicating reasonably good model fit. Figure 2 is a scatterplot of incidence rate per 1000 person-years for direct, design-based estimates (available for 48 states and DC) versus modeled (aggregated up to the state level) estimates. There is good agreement between the estimates with the modeled estimates being slightly higher, on average, than the direct estimates.
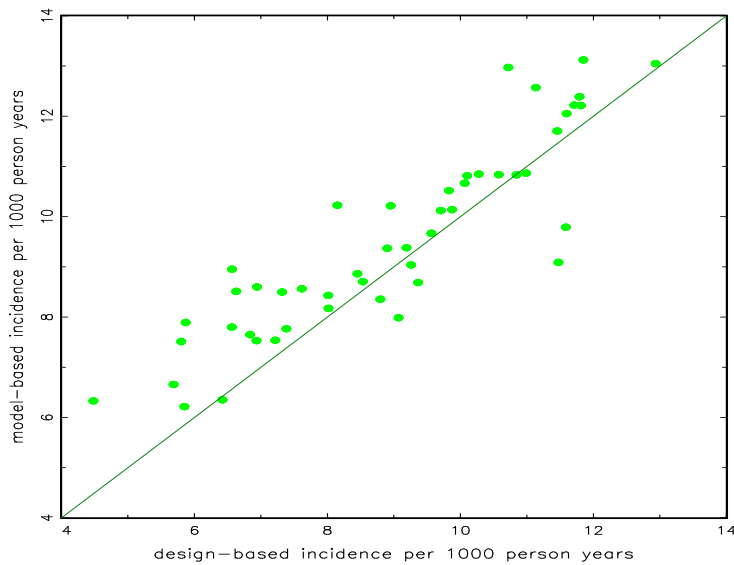
Figure 2: Scatterplot of design-based state incidence versus model-based county incidence aggregated to the state level for 48 states and the District of Columbia

The simulation study showed the modeled estimates to be quite good. Figure 3 displays the root mean squared error over the true value. For all but six of the estimates, the value is less than 0.25. All but four are less than 0.30, and, for most, the ratio is close to 0.15.
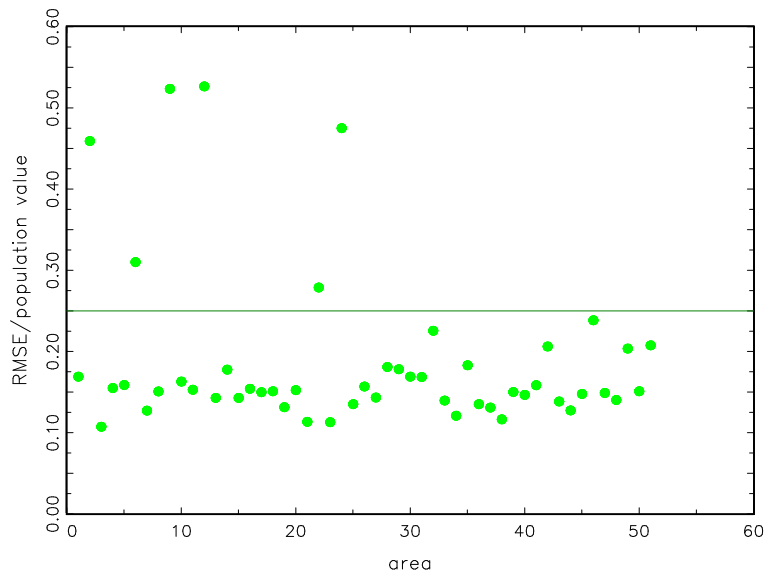

Figure 3: Scatterplot of root mean squared error of model-based incidence estimates over population value versus area. Results are averaged over 50 random samples from the test population

Figure 4 displays boxplots of differences (population minus modeled estimate). Only four states, out of 51 (counting the District of Colombia), have differences that are statistically significantly different from zero at the 0.05 level. The direct, design-based estimates from the simulation study are not usable because of large relative error. For the design-based estimates, the root mean squared errors divided by true values were over 0.30 for all 51 areas, and the median value was 0.54.
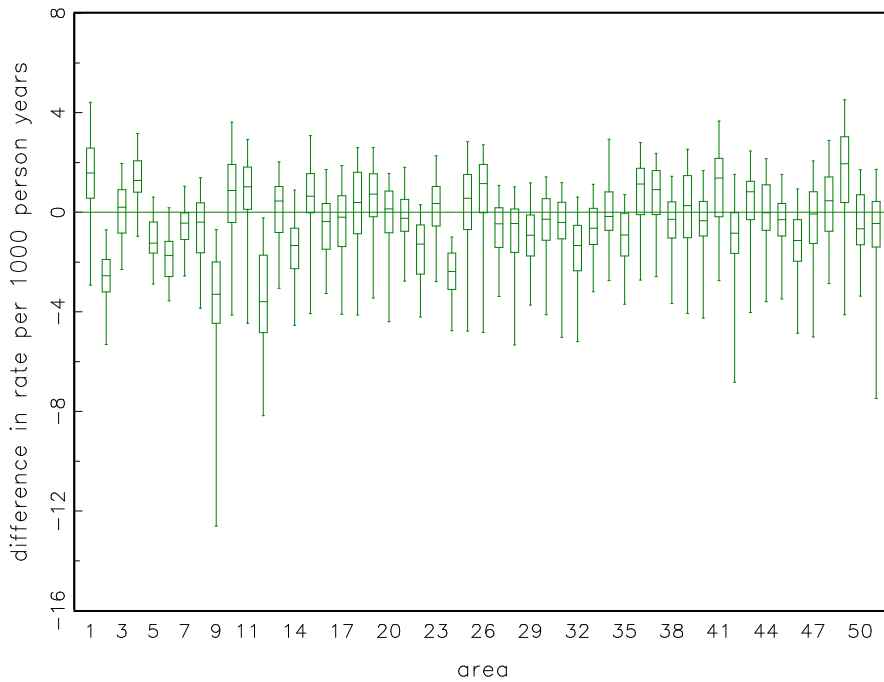

Figure 4: Box plots of difference in incidence, population minus model-based, by area. Each plot box represents 50 random samples from the test population

4. Discussion

We found a concentration of high-incidence counties in the U.S. southeast and in Appalachia, with a scattering of high-incidence counties in the U.S. west, primarily in counties with large American Indian populations. The map, Figure 1, also shows a concentration of low-incidence counties in the state of Colorado. The pattern of county-level diabetes incidence in the U.S. is roughly similar to the pattern of county-level prevalence, in the sense that high-prevalence counties tend to be high-incidence counties as well. Both high- incidence and prevalence, and low- incidence and prevalence counties correlate well with the county-level prevalence of obesity (CDC, 2012), a strong risk factor for type 2 diabetes. The prevalence of obesity is high in the southeast and in Appalachia, and low in

Colorado (CDC, 2012). The scattered counties in the west with high diabetes incidence have high prevalence of poverty, another risk factor for type 2 diabetes, and have relatively large populations of American Indians, who are at greater risk of type 2 diabetes than non-Hispanic whites, the majority of the U.S. population.

The strong association of prevalence and incidence estimates suggests that mortality among those with diagnosed diabetes and the rate of relocation to another county after developing diabetes, both of which could cause incidence and prevalence to look very different, are likely be relatively constant across the U.S. This was not known prior to this analysis.

Modeled estimates tended to be slightly larger than direct estimates, as depicted in Figure 2. This is probably due to direct estimates not accounting for those whose incidence status is unknown. Thus, the two estimates have somewhat different denominators, and therefore might be expected to differ.

Among the values of root mean squared error over the true value, Figure 3, there were four possible outliers: Alaska, District of Colombia, Hawaii, and Minnesota. Alaska, District of Colombia, and Hawaii are geographic/demographic outliers in the U.S. We have no explanation for Minnesota's being on this list.

Our study is subject to several limitations. First, instead of "incidence of diabetes", we are actually estimating "incidence of diagnosed diabetes". An estimated 27% of all U.S. diabetes cases are undiagnosed (Centers for Disease Control, 2011). People can have type 2 diabetes, which comprises 90-95% all diabetes (Centers for Disease Control, 2011), for years and be unaware of their condition. No national data set allows us to distinguish "incidence of diabetes" from "incidence of diagnosed diabetes". Second, our study is subject to the limitations of the BRFSS, such as not adequately representing households that have no land-line telephone (Centers for Disease Control, 2012b). Another limitation of BRFSS is recall bias: people might have simply misremembered the date at which their diabetes was diagnosed (although, for this study, respondents need only categorize diagnosis as: within the current calendar year; in the preceding calendar year; or before the preceding calendar year). In any survey, social desirability bias can exist, although self-report of diagnosed diabetes is usually correct (Okura *et al.*). Some of the participants in the study whose county of residence was imputed from telephone number might have been incorrectly categorized. This would tend to bias county estimates toward the overall mean. Also, BRFSS does not allow us to distinguish type 2 diabetes from the much less common type 1 diabetes. In our analysis, uncertainty in U. S. Census projections was ignored. This uncertainty is probably small, but not zero; however, the U.S. Census provides what is, by far, the best available estimates of N_{ij}. Finally, diabetes incidence rates are likely to vary over time. Our methods require at least three years of BRFSS data to produce stable estimates. Thus, we are estimating

a running average over time of incidence rates instead of the incident rate for a single year.

Incidence estimates stratified by demographics, such as age or race/ethnicity would have been desirable, since they would have provided insights into the causes of variability among counties in diabetes incidence. However, due to sample size limitations, such estimates have coefficients of variation that we consider unacceptably large. Therefore, we do not report such estimates.

While identifying counties of high and low diabetes incidence is of utility in directing public health interventions in the U.S., our methods have a broader applicability. Neither the methods developed in Cadwell *et al.* (2010) nor the modified version introduced here can be used to estimate rates of sufficiently rare events, such as diabetes incidence in a single state. However, if the outcome of interest is rare, but all cases of that outcome must also be cases of a more common outcome, our methods might be useful. The fraction of the population with the more common outcome can be estimated. Then the proportion of persons with the rare outcome who also have the more common outcome can be estimated. These estimates can be combined to estimate the faction of the population with the rare outcome. This opens up the possibility of data from BRFSS, or similar surveys in other countries, being used to provide small-area estimates of events that are too rare to be estimated in a one-stage model.

Appendix 1. Factoring the Full Probability Model

Let $Z_1 = 1$ if an individual was diagnosed with diabetes within the past year; $Z_1 = 0$ otherwise. Let $Z_2 = 1$ if an individual was diagnosed with diabetes; $Z_2 = 0$ otherwise. Let $\theta = (\theta_1, \theta_2)$ be the parameters associated with conditional incidence and prevalence, respectively. Assuming θ_1 is distinct from θ_2 and their prior distributions are independent then

$$
\begin{aligned}
(\theta|Z_1, Z_2) &= \frac{p(Z_1, Z_2|\theta)p(\theta)}{\int p(Z_1, Z_2|\theta)p(\theta)d\theta} \\
&= \frac{p(Z_1|Z_2, \theta)p(Z_2|\theta)p(\theta)}{\int p(Z_1|Z_2, \theta)p(Z_2|\theta)p(\theta)d\theta} \\
&= \frac{p(Z_1|Z_2, \theta_1)p(\theta_1)p(Z_2|\theta_2)p(\theta_2)}{\int\int p(Z_1|Z_2, \theta_1)p(\theta_1)p(Z_2|\theta_2)p(\theta_2)d\theta_1 d\theta_2} \\
&= \frac{p(Z_1|Z_2, \theta_1)p(\theta_1)}{\int p(Z_1|Z_2, \theta_1)p(\theta_1)d\theta_1} \times \frac{p(Z_2|\theta_2)p(\theta_2)}{\int p(Z_2|\theta_2)p(\theta_2)d\theta_2}.
\end{aligned}
\tag{A.1}
$$

Therefore the posterior distribution factors into two independent posterior distributions. Each part of the posterior can then be evaluated separately.

Appendix 2. Prior Assumptions

The intercepts by class are assigned improper flat priors, $\alpha \prec 1$.

The spatially correlated effects by county and class, ω, are assigned a multivariate normal (MVN) conditional autoregressive prior (Besag, York and Mollie, 1991). Let $\omega_i = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{i12})'$. Then

$$\omega_i | \omega_{(-i)} \sim \mathcal{MVN}(\bar{\omega}_i, \Sigma_\omega), \qquad\qquad (A.2)$$

where $\omega_{(-i)}$ equals the matrix $\boldsymbol{\omega}'$ with the ith column removed, $\bar{\omega}_i = \sum_{j \in \delta_i} \omega_j / n_i$, and δ_i and n_i denote the set of labels of the neighbors of county i and the number of neighbors, respectively, where counties are consider to be "neighbors" if they have a common border. The inverse of Σ_ω is assigned a Wishart prior with scale matrix S_ω and 12 degrees of freedom. The matrix S_ω has ones along the main diagonal and 0.001 for all other elements (Rao, 2003).

The spatially unstructured effects by county and class, μ, are assigned a multivariate (of dimension 14) normal prior with mean zero and variance matrix Σ_μ. The inverse of Σ_μ is assigned a Wishart prior with scale matrix S_μ and 14 degrees of freedom. The matrix S_μ has ones along the main diagonal and 0.001 for all other elements (Rao, 2003). The spatially unstructured effects by state and class, ν, are assigned the same types of priors as μ.

The error terms, ε, in the basic models are assigned a proper half-Cauchy (Gelman and Hill, 2007, Chapter 19) prior distribution with median equal to one. This is a diffuse prior. For this model, its use greatly speeds convergence.

References

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

Barker L. E., Kirtland K. A., Gregg E. W., Geiss L. S. and Thompson T. J. (2011). Geographic distribution of diagnosed diabetes in the United States: a diabetes belt. *American Journal of Preventive Medicine* **40**, 434-439.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1-59.

Brooks, S. P. and Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* **7**, 434-455.

Boyle, J. P., Thompson, T. J., Gregg, E. W., Barker, L. E. and Williamson, D. F. (2010). Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence. *Population Health Metrics* **8**, 29.

Cadwell, B. L., Thompson, T. J., Boyle, J. P. and Barker, L. E. (2010). Bayesian small area estimates of diabetes prevalence by U.S. county, 2005. *Journal of Data Science* **8**, 173-188.

Centers for Disease Control and Prevention. (2011). National diabetes fact sheet, 2011. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Atlanta, Georgia. Retrieved February 7, 2012, Web site: http://www.cdc.gov/diabetes/pubs/pdf/ndfs_2011.pdf

Centers for Disease Control and Prevention. (2012a). National diabetes surveillance system. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Atlanta, Georgia. Retrieved February 6, 2012, Web site: http://www.cdc.gov/diabetes/statistics/

Centers for Disease Control and Prevention. (2012b). Behavioral risk factor surveillance system survey data. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. Atlanta, Georgia. Retrieved February 6, 2012, Web site: http://www.cdc.gov/brfss/

Centers for Disease Control and Prevention. (2012c). Diabetes data and trends. U.S. Department of Health and Human Services, Centers for Disease Control and prevention. Atlanta, Georgia. Retrieved February 8, 2012, Web site: http://apps.nccd.cdc.gov/DDTSTRS/default.aspx

Congdon, P. and Lloyd, P. (2010). Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science* **8**, 235-252.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1-11.

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, J. B., Stern, H. S. and Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd edition. Chapman & Hall, New York.

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

Kakarlapudi, V, Sawyer, R and Staecker, H. (2003). The effect of diabetes on sensorineural hearing loss. *Otology & Neurotology* **24**, 382-386.

Kertes, P. J. and Johnson, T. M. (2007). *Evidence Based Eye Care.* Lippincott Williams & Wilkins, Philadelphia, Pennsylvania.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Wiley, New York.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edition. Chapman & Hall, New York.

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. *Journal of the American Medical Association* **290**, 1884-1890.

Okura Y, Urban, L. H., Mahoney, D. W., Jaconbsen, R. J. and Rodeheffer, R. J. (2004). Agreement between self-report and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not heart failure. *Journal of Clinical Epidemiology* **57**, 1096-1103.

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

Srebotnjak, T., Mokdad, A. H. and Murray, C. J. L. (2010). A novel framework for validating and applying standardized small area measurement strategies. *Population Health Metrics* **8**, 26.

U.S. Census Bureau. (2012). Population estimates, county characteristics: vintage 2009. U.S. Census Bureau. Washington, District of Columbia. Retrieved February 6, 2012, Web site: http://www.census.gov/popest/data/counties/asrh/2009/CC-EST2009-alldata.html

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
lsb8@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
tat5@cdc.gov

Karen A. Kirtland
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
gon6@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
hzb0@cdc.gov

Linda S. Geiss
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
lsg2@cdc.gov

Mary M. McCauley
Office of the Director
National Center for Immunization and Respiratory Diseases
Centers for Disease Control and Prevention
1600 Clifton Avenue, Mailstop A27, Atlanta, GA 30329, USA
zmf5@cdc.gov

Ann L. Albright
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
aga62@cdc.gov