ANNEX 1. Supplementary material for the study methods

1. Data

Data from the 2008, 2009, and 2010 BRFSS surveys were combined. For each of the 78 counties in Puerto Rico, sampled persons were cross-classified by age group (20–29 years of age, 30–39, ..., 70–79, 80+) and sex (male, female). This cross-classification resulted in 14 classes per county. The number of people sampled in each class that have diabetes can be determined. Specifically, let:

\[ n_{ij} = \text{the number of sampled people in county } i, \text{ class } j = 1, \ldots, 14 \]
\[ y_{ij} = \text{the number of sampled people with diagnosed diabetes in county } i, \text{ class } j \]

In some years, the \( n_{ij} \) is some counties will = 0. For these, the corresponding \( y_{ij} \) will also = 0. The United States Census Bureau publishes population estimates by demographic characteristics (unit-level auxiliary information) for all counties; the Census provides no information on diabetes status.

The 2009 Census county projections were used to obtain estimates for the number of persons in each age and sex group used to cross-classify the BRFSS data. Let, \( N_{ij} = \text{the estimated number of people in county } i, \text{ class } j = 1, \ldots, 14, \text{ in } 2009. \) Variability in Census projections was ignored.

The county-level covariates were obtained from the 2000 United States Department of Agriculture (USDA—population density) (13) and the 2005–2009 Puerto Rico Community Survey (percent of population 25 years of age and older who have completed high school; percent of population below poverty level in past 12 months) (14). The county-level covariates was centered and scaled by subtracting the overall mean from each and dividing the result by twice the standard error (SE).

2. Regression model

A Bayesian multilevel model was fit to the combined BRFSS data. This model relates observed quantities to other variables of interest. In particular:

\[ y_{ij} \sim \text{Binomial}(p_{ij} n_{ij}); i = 1, \ldots, 78 \text{ and } j = 1, \ldots, 14 \]
where \( p_{ij} \) is the prevalence of diagnosed diabetes in county \( i \), class \( j \). The regression model includes the following terms:

- (a) logit link function: \( \text{logit}(p_{ij}/(1 - p_{ij})) \)
- (b) a separate intercept for each class (age by sex group) \( \alpha_j; j = 1,\ldots,14 \)
- (c) effects of county-level predictors by sex \( \delta_{ij}; l = 1,2,3 \) and \( s = 1,2 \). Predictors include rural-urban continuum code \( x_1 \), percent of people 25 years and older who have completed high school \( x_2 \), and percent of people below poverty level in past 12 months \( x_3 \).
- (d) spatially correlated effects by county and class: \( v_{ij}; i = 1,\ldots,78 \) and \( j = 1,\ldots,14 \)
- (e) spatially unstructured effects by county and class: \( \mu_{ij}; i = 1,\ldots,78 \) and \( j = 1,\ldots,14 \)

Parameters under (b) and (c) are “fixed” effects, while (d) and (e) are random effects that borrow strength over county and class. Parameters under (d) are modeled via multivariate normal conditional autoregressive priors (of dimension 14) (37). These parameters allow spatial correlation of a county with its neighboring counties. Parameters under (e) are modeled via multivariate normal priors (of dimension 14) (38). These parameters allow for correlated effects across class without any form of spatial correlation over county. Thus, the regression model is:

\[
\text{logit}(p_{ij}) = \alpha_j + \delta_{ij1}x_{i1} + \delta_{ij2}x_{i2} + \delta_{ij3}x_{i3} + v_{ij} + \mu_{ij}
\]

Where \( [j]=1 \) if class \( j \) contains males and \( [j]=2 \) if class \( j \) contains females.

A basic model was considered as a benchmark to assess the study’s extended model. The basic model includes fixed effects for class and a spatially-unstructured random effect for county. The basic regression model is:

\[
\text{logit}(p_{ij}) = \alpha_j + \epsilon_i
\]

Where \( \epsilon_i \) are modeled via a normal prior with mean zero.

3. Estimates of diabetes prevalence

The study’s prevalence estimates of diagnosed diabetes in each county are the means of the posterior predictive distributions of the \( p_i \)’s:

\[
p_i = \frac{\sum_i p_{ij}N_i}{\sum_i N_i}
\]

This weighting by population totals is called poststratification and simultaneously corrects for nonresponse. Age-adjusted prevalence for county \( i \) is given by:

\[
\sum_k w_k \left[ \frac{\sum_{j \in \delta_k} p_{ij}N_i}{\sum_{j \in \delta_k} N_i} \right]
\]

where \( k \) indexes age group; \( w \) is a vector of standard population weights and \( \delta_k \) contains the subset of classes belonging to age group \( k \). The United States population in the year 2000 was used as the standard.

All posterior distributions were simulated in WinBUGS (17). The 2.5th and 97.5th percentiles of the posterior distributions of the \( p_i \)’s provided the 95%CI for county prevalence of diagnosed diabetes. A burn-in of 5000 was used, and a single chain for 20 000 iterations was then monitored.

4. Model comparison

Models were compared using the criterion developed by prior researchers (15). This criterion (D) is the sum of two parts, a goodness of fit measure (G), and the expected mean-square predictive error (P). Calculating D requires replicating the entire data set for each posterior draw of the model parameters. Using these replicates the posterior predictive mean and variance for each observation were computed. G is the sum over observations of the squared difference between the data and its posterior predictive mean; P is the sum over observations of the posterior predictive variances and D = G + P. Models with smaller values of D are preferred.
Objetivo. Calcular la prevalencia en el año 2009 de casos con diagnóstico de diabetes en Puerto Rico en adultos de 20 años de edad o mayores, para conocer mejor su distribución geográfica con objeto de que los responsables políticos puedan encauzar más eficientemente los programas de prevención y control.

Métodos. Se ajustó un modelo multinivel bayesiano a la combinación de datos del Sistema de Vigilancia de Factores de Riesgo del Comportamiento 2008–2010 y del Censo de los Estados Unidos del 2009 para calcular la prevalencia de la diabetes en cada uno de los 78 municipios de Puerto Rico.

Resultados. El cálculo del valor medio no ajustado para todos los municipios fue de 14,3% (intervalo por municipio de 9,9 a 18,0%). La amplitud promedio de los intervalos de confianza fue de 6,2%. Hubo poca diferencia entre los cálculos ajustados y los no ajustados.

Conclusiones. Los valores obtenidos mediante estos cálculos correspondientes a 78 municipios fueron por término medio más elevados y mostraron menor variabilidad (es decir, el intervalo era más pequeño) que los cálculos anteriormente publicados sobre la prevalencia de la diabetes en todos los municipios de los Estados Unidos en el 2008 (media, 9,9%; intervalo de 3,0 a 18,2%).

Palabras clave Diabetes mellitus; prevalencia; política social; Puerto Rico.