Title
New Approaches to Estimating National Rates of Invasive Pneumococcal Disease

Journal
American Journal of Epidemiology, 174(2)

ISSN
0002-9262

Authors
Costa, Marcelo A
Huang, Susan S
Moore, Matthew
et al.

Publication Date
2011-07-15

DOI
10.1093/aje/kwr058

DOI: 10.1093/aje/kwr058

Practice of Epidemiology

New Approaches to Estimating National Rates of Invasive Pneumococcal Disease

Marcelo A. Costa, Susan S. Huang, Matthew Moore, Martin Kulldorff, and Jonathan A. Finkelstein*

* Correspondence to Dr. Jonathan A. Finkelstein, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Avenue, 6th Floor, Boston, MA 02215 (e-mail: jonathan_finkelstein@harvardpilgrim.org).

Initially submitted October 19, 2010; accepted for publication February 14, 2011.

National infectious disease incidence rates are often estimated by standardizing locally derived rates using national-level age and race distributions. Data on other factors potentially associated with incidence are often not available in the form of patient-level covariates. Including characteristics of patients' area of residence may improve the accuracy of national estimates. The authors used data from the Centers for Disease Control and Prevention's Active Bacterial Core Surveillance program (2004–2005), adjusted for census-based variables, to estimate the national incidence of invasive pneumococcal disease (IPD). The authors tested Poisson and negative binomial models in a cross-validation procedure to select variables best predicting the incidence of IPD in each county. Including census-level information on race and educational attainment improved the fit of both Poisson and negative binomial models beyond that achieved by adjusting for other census variables or by adjusting for an individual's race and age alone. The Poisson model with census-based predictors led to a national estimate of IPD of 16.0 cases per 100,000 persons as compared with 13.5 per 100,000 persons using an individual's age and race alone. Accuracy of, and confidence intervals for, these estimates can only be determined by obtaining data from other randomly selected US counties. However, incorporating census-derived characteristics should be considered when estimating national incidence of IPD and other diseases.

estimation; pneumococcal infections; sentinel surveillance; statistics as topic; *Streptococcus pneumoniae*

Abbreviations: ABCs, Active Bacterial Core Surveillance; CDC, Centers for Disease Control and Prevention; IPD, invasive pneumococcal disease; PCV7, heptavalent pneumococcal conjugate vaccine.

Streptococcus pneumoniae (pneumococcus) is a common cause of morbidity and mortality from local and invasive infections, in the United States and worldwide (1). Risk factors for pneumococcal disease include age, chronic lung and cardiovascular disease, smoking, and immunosuppressive disorders (2). Independently of these factors, race and other sociodemographic variables have also been recognized to be related to the risk of pneumococcal carriage (3, 4) and infection (5). For example, even after introduction of the heptavalent pneumococcal conjugate vaccine (PCV7) in 2000, differences in the incidence of invasive pneumococcal disease (IPD) among white and black Americans persisted (5), though they may now be narrowing in some locales (6). It is unclear how much of the added risk associated with race, which is seen in a wide range of health conditions (7), may be explained by variables related to socioeconomic status, environmental exposures, and access to medical care. Information on potentially important sociodemographic variables is often not collected as part of infectious disease surveillance. Therefore, surveillance systems that use age and race only to estimate national rates of disease may over- or underestimate disease burden.

The Centers for Disease Control and Prevention (CDC) monitors the incidence of IPD using Active Bacterial Core Surveillance (ABCs), an active, population- and laboratory-based surveillance system operating in 10 state health departments (8). Because ABCs sites routinely audit laboratories to ensure complete reporting, observed rates of IPD within ABCs catchment areas are believed to be quite accurate. Incidence rates of IPD have varied widely, at least in the era before the introduction of PCV7, from 9 per 100,000 persons to 19 per 100,000 persons across geographic regions

(9). National estimates for IPD have been calculated by adjusting the rates derived from ABCs surveillance areas for the age and race distributions of the United States, as reported by the US Census Bureau. However, the ABCs sites were selected through a competitive process that evaluated criteria such as the existing public health infrastructure and the availability of academic partners at each site (10). Therefore, since ABCs represents a nonrandom sample of US counties, extrapolation to the remainder of the United States may not yield optimal national estimates.

We hypothesized that incorporation of community-level variables, in a statistical model, might produce better national estimates of the incidence of IPD. Community-level data do not replace information collected on individuals. Instead, such data reflect the demographic attributes of the environment (11). For example, we have shown that independently of whether a child receives out-of-home child care, living in an area where most children attend day care increases the risk of pneumococcal carriage (12). We limited our investigation of potential predictors to those available from the US Census Bureau, since they are widely available and require no additional original data collection. Since the surveillance areas of the ABCs are counties for which we already know the number of IPD cases, we used county-level information as candidate predictors. The primary study question was whether a model for extrapolating the incidence of IPD from a nonrandom group of US counties that adjusted for census-based variables would be superior to one that adjusted for the age and race distribution of the United States alone. The result might inform calculation of other national estimates from surveillance programs based on limited and non-randomly selected geographic areas.

MATERIALS AND METHODS

Data sources

ABCs conducts active IPD surveillance in 170 geographically diverse counties in California (1 county), Colorado (5 counties), Connecticut (all 8 counties), Georgia (20 counties), Maryland (6 counties), Minnesota (all 75 counties), New Mexico (all 26 counties), New York (15 counties), Oregon (3 counties), and Tennessee (11 counties) (Figure 1). Cases are reported when pneumococcus is isolated from a normally sterile site (e.g., blood, cerebrospinal fluid) by clinical microbiology laboratories serving these counties, as described elsewhere (13). Information on an individual's age, race, ethnicity, gender, and county is collected on a standard reporting form. For this analysis, we used the 7,370 cases of IPD identified in the years 2004 and 2005 using 6 age groups: 0–<2, 2–<5, 5–<18, 18–<50, 50–<65, and ≥65 years.

The 2000 US Census contains information on the 115.9 million housing units and 281.4 million persons present in the United States in the year 2000, and its data are available in a variety of formats and media (http://www.census.gov/). We selected 24 county-level variables, a priori, as potential candidates to enter into a prediction model for the incidence of IPD. Since many measures had the potential to be correlated, we initially grouped variables into measures of population density, age distribution, race/ethnicity, education, economic factors, household crowding, and other measures (Table 1). In the 2000 US Census, the variable for race (e.g., white, black, American Indian, etc.) was separated from that for ethnicity (Hispanic vs. non-Hispanic). Therefore, while categories of race are mutually exclusive, race and ethnicity categories are not.

Household crowding was defined as the proportion of owner- and renter-occupied housing units with more than 1 occupant per room. County in-migration was defined as the proportion of the population over 5 years of age who had become new residents of the county in the previous 5 years. County-level vaccine penetration information was obtained from published CDC estimates (14) of the fraction of children aged 19–35 months who had received appropriate primary and booster vaccines (although we did not include PCV7 in the coverage fraction, to allow comparison with prior years).

Statistical methods

Model assumptions. Poisson regression was used to model the number of cases Y_{ij} in county $j = \{1, 2, \ldots, 170\}$ and age group $i = \{1, 2, \ldots, 6\}$ as

$$Y_{ij} \sim \text{Poisson}\left(\mu_{ij}\right), \tag{1}$$

where $\mu_{ij} = \text{Pop}_{ij} e^{\beta_0 + \alpha_{ij} + \sum \beta_k x_{jk}}$, Pop_{ij} is the population in age group i in county j, β_0 is the intercept, α_i is the age group effect, β_k is the census-level variable effect, and x_{jk} is the value of the kth census variable for county j.

A second model assuming a negative binomial distribution for the outcomes was also tested to account for the possibility of overdispersion in the observed data. As with the Poisson regression model, a logarithm link function was used.

$$Y_{ij} \sim \text{negative binomial}\left(\mu_{ij}\right). \tag{2}$$

US Census variable selection. In order to identify US Census variables that might improve model prediction performance, we began by fitting models for each variable within each of the 7 groups of similar socioeconomic variables and then scoring the variables on the basis of their predictive performance through a cross-validation procedure. Because many of the sociodemographic variables are highly correlated, putting more than 1 variable from each group into a model would probably not improve model fit greatly. Therefore, the single best variable from each group was tested for inclusion in the final model. In all tested models, the age group of each IPD case was always included.

Model fitting. We designed and evaluated several candidate models using predictive variables from each group of census variables, specified a priori (Table 1). Candidate models included 2 or more variables (maximum of 5) from different variable groups. The most predictive variables from each group were used in the candidate models. We

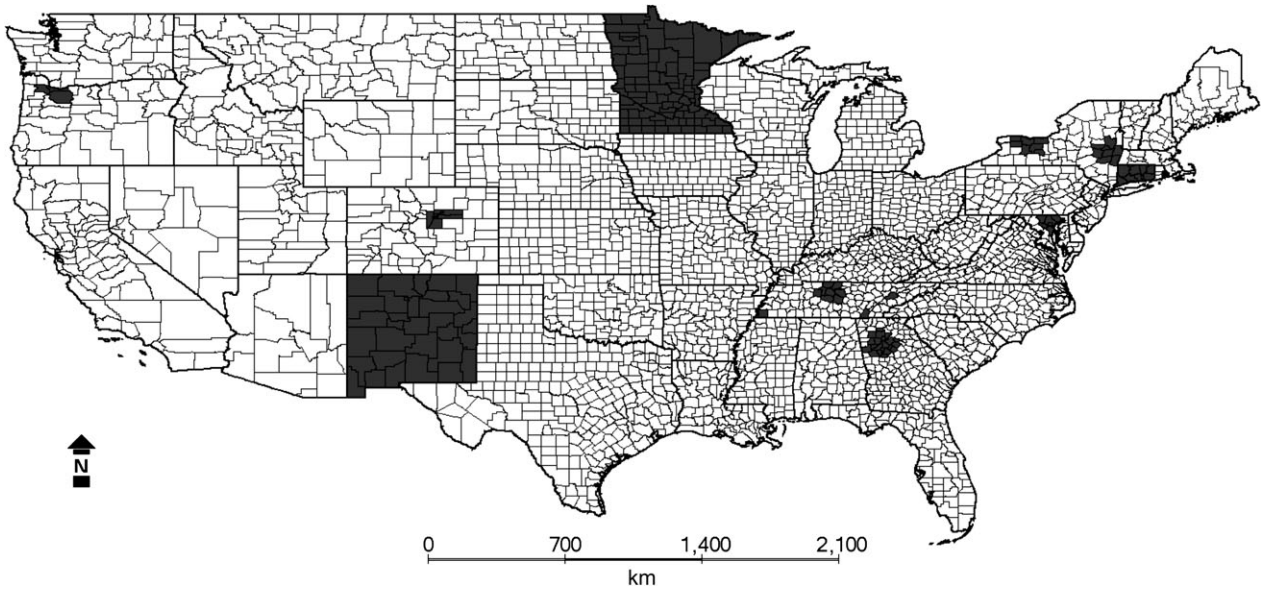

Figure 1. US counties (dark areas) included in the Centers for Disease Control and Prevention's Active Bacterial Core Surveillance (ABCs) system.

also tested some of the variables that were ranked as second-best predictors from the groups. Finally, we selected 7 final models with improved prediction as measured using cross-validation: 1) race only (proportion white); 2) race and high school education (proportion with less than a high school education); 3) race, high school education, and general poverty (proportion of persons living in poverty); 4) race, high school education, and child poverty (proportion of children under age 6 years living in poverty); 5) race, high school education, child poverty, and household income; 6) population density, race, high school education, general poverty, and vaccine penetration; and 7) population density, race, high school education, child poverty, and vaccine penetration. All candidate models were tested using both the Poisson and negative binomial distributions by cross-validation. After finding the model with the least cross-validation error, we tested whether an interaction term between age group and the census-derived variables improved the fit further.

Cross-validation procedure. Cross-validation partitions an existing sample of data (here the known cases of pneumococcal disease in the 170 ABCs counties) into subsets (15). In K-fold cross-validation, the original sample is partitioned into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K - 1$ subsamples are used to fit the model. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results can then be averaged (or otherwise combined) to produce a single error estimate. The advantage of this method over repeated subsampling is that all observations are used for both fitting the model and validation, and each observation is used for validation exactly once (16).

We used a 10-fold cross-validation in which each partition represented one of the states in the ABCs' surveillance network. A 170-fold cross-validation was also tested, where each partition represented 1 county in the sample; results were similar to those for the 10-fold state-level cross-validation, and therefore we present only the state-level cross-validation results. Models were used to provide predictions for all ABCs sites in the sample, and the predicted values were compared with the observed values using the generalized Pearson and error statistics.

Generalized Pearson and error statistics. The generalized Pearson statistic (17) and the sum of squared error (or simply "error") statistic were applied to compare the candidate models. The generalized Pearson statistic is

$$X^2 = \sum_{i,j} \frac{\left(y_{ij} - \hat{\mu}_{ij}\right)^2}{V_{ij}}, \tag{3}$$

where $\hat{\mu}_{ij}$ is the expected number of cases estimated using the 10-fold cross-validation procedure. For the Poisson model, $V_{ij} = \hat{\mu}_{ij}$, and for the negative binomial, $V_{ij} = \hat{\mu}_{ij}\left(1 + \kappa \times \hat{\mu}_{ij}\right)$, where κ is a dispersion parameter that is estimated from the data. The error statistic is

$$\text{Error} = \sqrt{\sum_{i,j} \left(y_{ij} - \hat{\mu}_{ij}\right)^2}. \tag{4}$$

The smaller the error, the better is the model's prediction. Both metrics were calculated with data stratified by age group and county. Note that either statistic can be used to

Table 1. Pearson and Error Statistics Derived From Bivariate Negative Binomial and Poisson Models Estimating the Incidence of Invasive Pneumococcal Disease in the United States, 2004–2005[a]

Variable	Negative Binomial Model		Poisson Model	
	Pearson	Error	Pearson	Error
Intercept only	*202.20*[b]	294.23	*3,041.34*	287.24
Household crowding				
Average household size (no. of persons per housing unit)	217.71	300.29	3,235.85	285.80
% living in crowded housing (>1 person per room)	*196.17*	*291.51*	*2,864.12*	*283.90*
Population density				
No. of persons per square mile[c]	203.41	494.07	3,893.15	751.64
No. of children under age 6 years per square mile	204.01	*323.23*	*3,046.24*	357.24
No. of adults aged 65 years or more per square mile	*201.97*	597.76	4,128.93	874.71
% of population in urban areas	205.04	293.28	3,135.63	*286.88*
Age distribution				
% of persons under age 6 years	213.52	298.92	3,099.25	288.41
% of persons aged 65 years or more	200.12	295.44	3,058.48	288.91
% of households with children under age 18 years	212.00	299.40	3,125.82	*286.02*
% of households with adults aged 65 years or more	*185.91*	*294.94*	*3,043.66*	289.42
Race/ethnicity, %				
White	204.48	*279.72*	*2,561.87*	*270.25*
Black	208.06	283.21	2,808.89	277.98
Asian	204.00	306.90	4,284.24	309.11
Native American	185.26	295.14	3,260.58	288.45
Other	204.19	295.49	3,131.62	289.62
Hispanic and Latino	204.01	295.48	3,132.93	289.60
% foreign-born	190.29	296.72	3,207.81	294.76
Education				
% of adults with less than a high school education	*135.62*	*264.48*	*2,141.76*	*244.59*
% of adults with less than a college education	139.26	289.56	2,727.56	279.44
Economic factors				
% of persons living in poverty (general poverty)	141.84	*259.61*	*2,201.30*	*239.69*
% of children under age 6 years living in poverty (child poverty)	156.57	266.05	2,375.83	252.91
Median household income/1,000 population	*124.51*	276.60	2,397.88	259.12
% unemployed among persons aged 16 years or more	179.76	275.74	2,584.52	265.19
Other				
Vaccine penetration[d], %	207.27	290.40	2,915.57	281.21
County in- and outmigration[e], per 1,000	*184.18*	299.33	3,100.62	299.31

[a] Results were adjusted for age group and 1 census variable.
[b] The italic entries identify the models with the best fit within each category.
[c] 1 square mile = 1.6 km^2.
[d] Estimated vaccine coverage level among children aged 19–35 months.
[e] Rate of domestic migration per 1,000 population.

compare the same probability model with different variables, but only the error statistic can be used to compare the Poisson model versus the negative binomial model with the same covariates. The Pearson statistic cannot be used in this way, since the denominator is calculated differently for the 2 types of models.

Estimates of national incidence. To estimate the national incidence, the 2 best models (those with the best cross-validation results) were used to determine the estimated number of cases for each age group in each of the 3,108 counties in the United States. The age-specific national estimate is the sum of the age-specific county-level estimates, and the overall national estimate is the sum for all age groups and counties. The age-specific national incidence rates per 100,000 persons were calculated as the national estimate of the number of cases in that age group divided by the national population in that age group, multiplied by 100,000. The overall incidence rates per 100,000 were calculated in the same way.

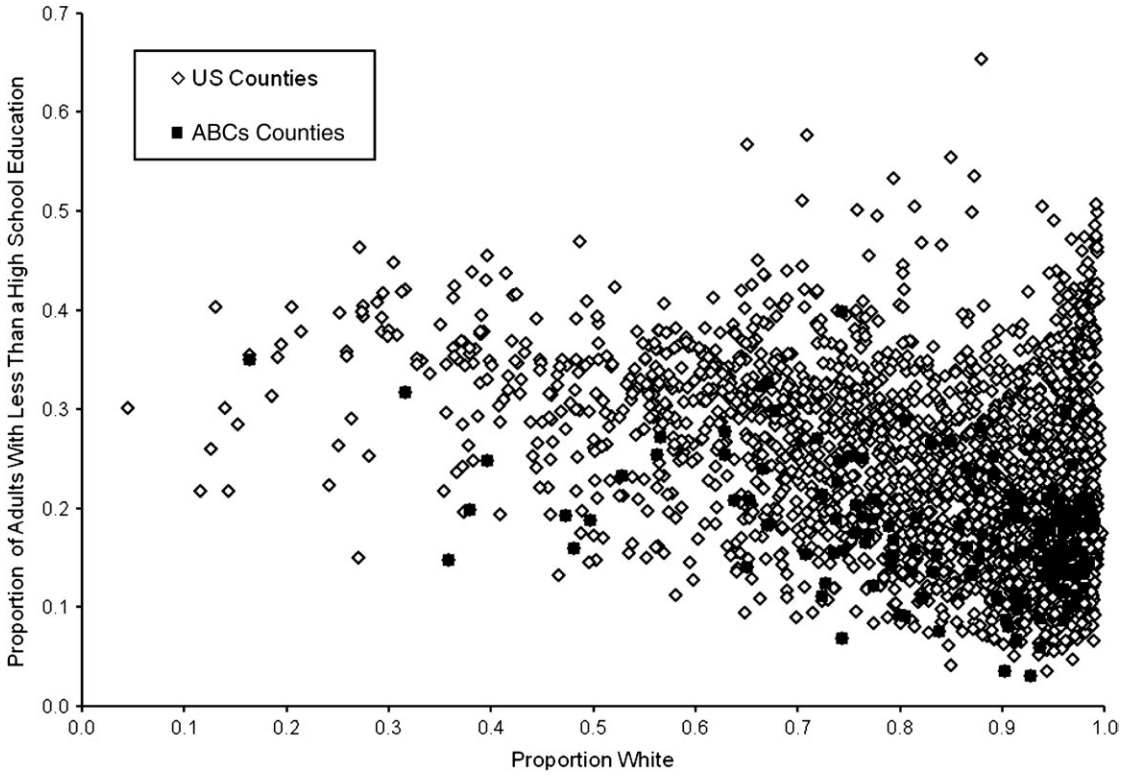

Figure 2. Proportion (%) of adults with less than a high school education according to the proportion (%) of county residents who are white, United States, 2004–2005. Black squares, counties included in the Active Bacterial Core Surveillance (ABCs) system; white diamonds, non-ABCs counties ("predicted" data).

Reference model. For comparison, we also fitted a model with only individual covariates, consisting of age group and race, without any county-level variables. Since information on race was missing for 16% of individuals, we imputed values in proportion to the nonmissing cases in the same age group, race, and county. This model is the one that is most similar to the model currently used by the CDC to estimate the number of cases of IPD in the United States, although we used a different imputation procedure for missing values.

RESULTS

We first assessed how well the catchment areas of the ABCs represented all US counties. As one diagnostic, we analyzed a scatterplot of 2 census variables shown to be highly predictive of IPD rates: the proportion of persons of white race and the proportion of persons with less than a high school education in each county (Figure 2). The ABCs counties appear as filled squares in Figure 2 and the non-ABCs counties, which need to be "predicted" in the development of national estimates, as empty diamonds. The scatterplot suggests that the ABCs sample includes counties with a similar distribution (compared with non-ABCs counties) of the proportion of residents of white race, but it includes counties with a lower proportion of people with less than a high school education than is true for US counties overall. It is this nonrandom nature of the ABCs counties that make it important to consider various county-level adjustment variables.

The results from initial age-adjusted models testing the effect of each census variable alone are shown in Table 1 for both the Poisson and negative binomial models. We have chosen to display both Pearson and error statistics to highlight the difference between model fitting (better Pearson statistics) and accurate cross-validation results (better error statistic). Results are similar regardless of whether we evaluate the variables based on the Pearson statistic or the error statistic. Note that for almost all variables, the Poisson models achieved lower (i.e., better) error statistics than the negative binomial models, and hence the Poisson models performed better.

We next evaluated 7 models that included various combinations of predictive county-level variables (Table 2). All 7 models performed considerably better than the reference model (model 0) based on the Pearson statistic, and all but 1 performed considerably better based on the error statistic. As is seen by comparing models 0 and 1, individually ascertained race can be replaced by a county-level race variable without any loss in predictive power. Across all models, the values of the error statistic are lower for the Poisson models than for the negative binomial models with the same variables. Model 3, with the proportion white, proportion of

Table 2. Candidate Models Estimating the Incidence of Invasive Pneumococcal Disease in the United States With Predictive Pearson and Error Statistics, Using Model 0 (Without County-Level Variables) as the Reference Model, 2004–2005

Model[a] and Variables	Negative Binomial Model		Poisson Model	
	Pearson	Error	Pearson	Error
Model 0	N/A[b]	N/A[b]	4,051.61	269.98
Age and individual race				
Model 1	204.48	279.72	2,561.87	270.25
% white				
Model 2	148.46	255.06	2,077.49	234.62
% white				
% of adults with less than a high school education				
Model 3	143.55	*252.81*[c]	2,078.02	*232.44*
% white				
% of adults with less than a high school education				
% of persons living in poverty				
Model 4	146.29	253.52	2,100.96	234.86
% white				
% of adults with less than a high school education				
% of children under age 6 years living in poverty				
Model 5	*125.65*	254.58	*2,045.23*	234.67
% white				
% of adults with less than a high school education				
% of children under age 6 years living in poverty				
Median household income/1,000 population				
Model 6	140.18	258.48	2,104.83	234.46
% of persons living in crowded housing				
% white				
% of adults with less than a high school education				
% of persons living in poverty				
Vaccine penetration				
Model 7	145.31	259.35	2,174.49	239.85
% of persons living in crowded housing				
% white				
% of adults with less than a high school education				
% of children under age 6 years living in poverty				
Vaccine penetration				

Abbreviation: N/A, not applicable.
[a] All models included age.
[b] The model did not fit the data with only age and race included, so it did not produce a result.
[c] The italic entries represent solutions with the least error.

adults with less than a high school education, and proportion in poverty variables, performs best on the basis of the error statistic. Model 2, with only the first 2 of these variables, performs almost as well. Model 5, with proportion white, proportion of adults with less than a high school education, proportion of persons in poverty, and median household income, has the best model fit according to the Pearson statistic. In the best predictive model (model 3), the error statistic was improved by 16% in comparison with the reference model. These findings suggest that including county-level measures of race, poverty, and educational attainment of the local population can contribute to improved prediction of national IPD counts, compared with estimates based on people's age and race alone.

Since the Poisson models performed better than the negative binomial models, the remaining results focus on the Poisson models. All of these were evaluated by adding age group $\times$ county-level variable interaction terms. Only models 2 and 3 achieved better prediction when the interaction terms were added, with both models showing improvement in prediction in every age group. Their error statistics, categorized by age group, are shown in Table 3.

Table 3. Comparison of the Error Statistics for Models Estimating the Incidence of Invasive Pneumococcal Disease in the United States, With and Without Age Group × County-Level Variable Interaction Terms Included in the Poisson Model, 2004–2005

Model and Variables	Age Group, years						
	0–<2	2–<5	5–<18	18–<50	50–<65	≥65	Total
Model 0	33.93	19.17	17.13	209.65	123.64	108.79	269.98
Age and individual race							
Models With Age Group Effect and No Age Interaction Terms							
Model 2	39.38	20.11	16.27	174.74	93.31	116.56	234.62
% white							
% of adults with less than a high school education							
Model 3	39.58	20.17	15.84	174.89	92.04	112.91	232.44
% white							
% of adults with less than a high school education							
% of persons living in poverty							
Models With Interaction Between Age Group and County-Level Variables							
Model 2	36.83	17.35	15.63	142.28	83.64	74.95	186.44
% white							
% of adults with less than a high school education							
Model 3	37.30	18.36	15.13	161.87	84.70	75.08	202.41
% white							
% of adults with less than a high school education							
% of persons living in poverty							

For example, model 2, with proportion white, proportion of adults with less than a high school education, and the age group × county-level variable interaction terms (Table 3), achieved improvement of approximately 26% in prediction when compared with the same model without the age group × county-level variable interaction terms.

The national estimates derived from the reference model and from the 2 best models are shown in Table 4. Not surprisingly, the model based on individually ascertained or imputed race provides estimates closer to current estimates for IPD, since this method is analogous to the one used currently by ABCs. For the remaining models, the estimates

Table 4. National Estimates of Expected Numbers of Cases of Invasive Pneumococcal Disease per 100,000 Persons for Models With Imputed Race, Best Census Variable Models (With Age × County Interaction), and Current Published Estimates, United States, 2004–2005

Model and Variables	Age Group, years						
	0–<2	2–<5	5–<18	18–<50	50–<65	≥65	Total
Model 0	35.6	12.1	2.5	7.4	21.9	38.8	13.5
Age and individual race							
Model 2	*36.9*[a]	*13.0*	*3.1*	*9.7*	*27.3*	*42.9*	*16.0*
% white							
% of adults with less than a high school education							
Model 3	36.9	13.2	3.1	9.7	27.4	43.0	16.0
% white							
% of adults with less than a high school education							
% of persons living in poverty							
Active Bacterial Core Surveillance system estimates[b]	35.8	12.7	2.7	7.5	19.1	39.5	13.5

[a] The italic values represent national estimates using the best model found in the cross-validation procedure.
[b] Data were provided by the Centers for Disease Control and Prevention (2004).

for age groups over 18 years were particularly high in comparison with current estimates. These were the age groups that achieved improved prediction when the model with race, education variables, and the age group × county-level variable interaction was tested (Table 3). Although results are not displayed in Table 4, models 2 and 3 (with no age interaction terms) provided similar national estimates of 16.1 cases per 100,000 persons.

DISCUSSION

We have identified a method by which to incorporate census-level variables to inform national estimates of disease incidence when surveillance is limited to a nonrandom sample of defined geographic areas in the United States. We developed and evaluated models that use publicly available US Census variables as predictors to estimate the national incidence of IPD from the sample of US counties under active surveillance through ABCs. To do this, we first tested the utility of Poisson and negative binomial models with a cross-validation procedure, using Pearson and error statistics as selection criteria. We then added US Census variables to determine the most parsimonious combination of variables that resulted in the lowest residual error. Then, we improved the 2 best models by adding interaction terms for age group × census-level variables. Overall, we found that the model that best fit the data included census variables describing county-level race and education, accounting for interaction with age group. Although overall national estimates were similar regardless of whether or not the interaction term was added to the model, models within age groups that included the interaction term provided better prediction.

As we expected, the negative binomial is better at modeling the uncertainty in the rates when there is overdispersion. However, for a predictive model, the key is to have accurate point estimates, and the Poisson model was superior in this regard. This is so because the maximum likelihood fitting alters the point estimates. As a consequence, the negative binomial fitting may not match the observed rates as well as the Poisson. In situations where accurate point estimates are the priority, cross-validation error is the appropriate statistic for model selection.

Note that the approach currently used by ABCs to develop national estimates of IPD and the proposed model from this work share race as a predictor variable. However, in the current ABCs approach, race is an attribute recorded at the level of the individual, with missing values for some persons. In the proposed model, individually ascertained race is replaced by a county-level variable from the US Census (proportion of persons of white race in each county). The observation that the county-level race distribution functioned almost as well as individual-level information on race suggests that this may be a useful option when data on race have not been collected or are missing for a substantial fraction of persons.

The approach reported here has some limitations. We used data from the 2000 US Census and made the implicit assumption that the county-level variables used in these models were relatively stable through 2004 and 2005. Estimated population denominators for 2004 and 2005 for each county were likewise obtained from the 2000 Census. As with other parameters estimated by statistical models, it would be useful to develop confidence intervals around the incidence rates for IPD that we have estimated here. In practice, that is not possible, since the estimates do not come from a randomly selected set of counties. The accuracy of the new national estimates and confidence intervals can only be evaluated by obtaining data from an independent random sample of counties. Such an external validation is beyond the scope of this study. To evaluate the predictive ability of our candidate models, we relied on a cross-validation procedure. However, our predicted rates are also subject to the same issue of potential bias because of nonrandomly selected counties. That is, if there were systematic biases in the selection of ABCs counties that are not reflected in our selected census-level variables, then the estimate of the national incidence will also be biased. A comparison of the predicted number of cases and actual counts of IPD for randomly selected counties not included in the ABCs database would allow further evaluation of the performance of this method.

We believe that the general method described here may be applied equally well to estimate national incidence for other reported diseases when only information from nonrandomly selected geographic areas is available. This is the case in many current surveillance programs (18–20). In the particular case of IPD, our analysis suggests that the use of county-level variables improves the estimation of IPD incidence. Thus, this method may provide an alternative to the labor-intensive collection of detailed individual data if surveillance systems are resource-constrained. However, in addition to the above limitations, we note that the use of this approach for other conditions is likely to reveal a different set of predictor variables that produce the smallest residual error in a cross-validation procedure. It is also possible that for other clinical conditions, individually ascertained information from cases (regarding race, for example) may be more useful than the assignment based on census-level information we employed. Exploration of the utility of these techniques in the evaluation of other conditions and surveillance programs is warranted.

ACKNOWLEDGMENTS

Author affiliations: Department of Statistics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil (Marcelo A. Costa); Division of Infectious Diseases and Health Policy Research Institute, School of Medicine, University of California, Irvine, Irvine, California (Susan S. Huang); Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, Georgia (Matthew Moore); and Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts (Martin Kulldorff, Jonathan A. Finkelstein).

This work was supported by a cooperative agreement (TS-1363) with the US Centers for Disease Control and Prevention through the Association of Prevention Teaching and Research.

Conflict of interest: none declared.

REFERENCES

1. O'Brien KL, Wolfson LJ, Watt JP, et al. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. Pneumococcal Global Burden of Disease Study Team. *Lancet*. 2009;374(9693):893–902.
2. Robinson KA, Baughman W, Rothrock G, et al. Epidemiology of invasive *Streptococcus pneumoniae* infections in the United States, 1995–1998: opportunities for prevention in the conjugate vaccine. Active Bacterial Core Surveillance (ABCs)/ Emerging Infections Program Network. *JAMA*. 2001;285(13): 1729–1735.
3. Millar EV, O'Brien KL, Zell ER, et al. Nasopharyngeal carriage of *Streptococcus pneumoniae* in Navajo and White Mountain Apache children before the introduction of pneumococcal conjugate vaccine. *Pediatr Infect Dis J*. 2009;28(8): 711–716.
4. Huang SS, Finkelstein JA, Rifas-Shiman SL, et al. Community-level predictors of pneumococcal carriage and resistance in young children. *Am J Epidemiol*. 2004;159(7):645–654.
5. Flannery B, Schrag S, Bennett NM, et al. Impact of childhood vaccination on racial disparities in invasive *Streptococcus pneumoniae* infections. Active Bacterial Core Surveillance/ Emerging Infections Program Network. *JAMA*. 2004;291(18): 2197–2203.
6. Talbot TR, Poehling KA, Hartert T, et al. Elimination of racial differences in invasive pneumococcal disease in young children after introduction of the conjugate pneumococcal vaccine. *Pediatr Infect Dis J*. 2004;23(8):726–731.
7. Smedley BD, Stith AY, Nelson AR, eds. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Washington, DC: National Academies Press; 2003.
8. Whitney CG, Farley M, Hadler J, et al. Active bacterial core surveillance of the emerging infections program network. *N Engl J Med*. 2003;348(18):1737–1746.
9. Rosen J, Thomas A, Lexau C, et al. Geographic variation in invasive pneumococcal disease [abstract]. Presented at the 48th Annual Interscience Conference on Antimicrobial Agents and Chemotherapy (ICAAC) and the Infectious Diseases Society of America (IDSA) 46th Annual Meeting, Washington, DC, October 25–28, 2008.
10. Pinner RW, Rebmann CA, Schuchat A, et al. Disease surveillance and the academic, clinical, and public health communities. *Emerg Infect Dis*. 2003;9(7):781–787.
11. Chen FM, Breiman RF, Farley M, et al. Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive *Streptococcus pneumoniae* infections. *Am J Epidemiol*. 1998;148(12):1212–1218.
12. Huang SS, Finkelstein JA, Lipsitch M. Modeling community- and individual-level effects of child-care center attendance on pneumococcal carriage. *Clin Infect Dis*. 2005;40(9): 1215–1222.
13. Schuchat A, Hilger T, Zell E, et al. Active bacterial core surveillance of the Emerging Infections Program Network. Active Bacterial Core Surveillance Team of the Emerging Infections Program Network. *Emerg Infect Dis*. 2001;7(1): 92–99.
14. National, state, and urban area vaccination coverage among children aged 19–35 months—United States, 2005. *MMWR Morb Mortal Wkly Rep*. 2006;55(36):988–993.
15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer Science + Business Media, Inc; 2001.
16. Picard RR, Cook RD. Cross-validation of regression models. *J Am Stat Assoc.* 1984;79(387):575–583.
17. McCullagh P, Nelder JA. *Generalized Linear Models* 2nd ed. London, United Kingdom: Chapman and Hall Ltd; 1989.
18. Poehling KA, Edwards KM, Weinberg GA, et al. The underrecognized burden of influenza in young children. *N Engl J Med*. 2006;355(1):31–40.
19. Musinguzi J, Kirungi W, Opio A, et al. Comparison of HIV prevalence estimates from sentinel surveillance and a national population-based survey in Uganda, 2004–2005. *J Acquir Immune Defic Syndr.* 2009;51(1):78–84.
20. Tate JE, Panozzo CA, Payne DC, et al. Decline and change in seasonality of US rotavirus activity after the introduction of rotavirus vaccine. *Pediatrics*. 2009;124(2):465–471.