Appendix B.

A Proposal for a Method to Designate Communities as Underserved

Technical Report on the Derivation of Weights

This Appendix is intended to provide more technical details about the proposed methodology and how it was developed. The principal authors of this document are, alphabetically: Laurie Goldsmith, Mark Holmes, Jan Ostermann, and Tom Ricketts.

The General Approach

The overall approach for deriving an empirical, data driven system to identify underserved areas and populations is to estimate the effect of demographic factors on the population-to-practitioner ratio, using a sample of counties as proxies for a health care market. These effects are then translated to a score which is added to an adjusted ratio for a total "need" measure. Thus, the implementation is similar to the current IPCS or MUA method in that it creates a "score" or "index" of underservice, however, the proposed system's score is based on an adjusted ratio that is meant to represent an "effective" or "apparent" population and its primary health care needs.

There are eight steps to the project, which we divide for expository purposes into two distinct "Tasks". Please note that the specific steps described earlier in the preamble to this rule may not match up to the steps described below (for example, "step 4" in the preamble matches up with "steps 4-5" and "step 7" in this appendix).

Task One: Calculate The Weights that will be used to Adjust Ratios ("Analysis")

This is the analytical portion of the project in which we explore the degree to which observable demographic characteristics tend to be associated with population to provider ratios. The specific steps in this task include:

1. *Create an age-sex adjusted population.*
2. *Calculate the base population-provider ratio for regression to determine weights for need variables.*
3. *Select study sample primary care service area proxies.*
4. *Create factor scores to control for interactions of variables.*

5. *Run regression models to create weights for community variables.*

Task Two: Calculate The Scores Based On These Factors ("Computation")

This is the portion of the process in which scores are assigned to geographic areas based on the weights calculated in Task One.

6. *Calculate the base population-practitioner ratio for designation determination*
7. *Calculate the scores for each area based on the values for each variables for each area and add to the ratio.*
8. *Step 8: Compare the ratio to a designation threshold ratio.*

We describe each of these steps in detail in the following sections.

Task 1: Analysis Steps

Step 1: Create an age-sex adjusted population

Using estimated visit rates from individual-level surveys, we weight the population to create a "base population." In this manner, populations can be compared across areas. The use of these data for this adjustment are discussed in detail in reports and background papers for the proposal including the report that estimates the national impact of the NPRM-2 proposal, "National Impact Analysis of a Proposed Method to Designate Communities as Underserved" dated September 7, 2001; the background paper, "Designating Underserved Populations. A Proposal For An Integrated System Of Identifying Communities With Multiple Access Challenges," which is in draft form; and the "Executive Summary" of the "Designating …" paper which has been circulated in draft form to the Bureau of Primary Health Care.

The weights are summarized in Table 1.

Table 1: Visit weights for age-sex adjustment

	0-4	5-17	18-44	45-64	65-74	75 and ove
Female	4.046	2.256	5.007	5.480	6.710	8.160
Male	5.164	2.499	2.867	4.410	6.052	8.056

these are the original weights using 1996 data)

The weighted sum of these populations is calculated as 4.046 * (# Females 0-4) + 2.256 * (# Females 5-17) +…+ 8.056 *(# Males 75 and over) and equals an age-sex adjusted number of visits for a particular population. Dividing this number of visits by the mean visit rate (3.741) creates a "base population". Areas with equal base populations (and equal demographics) have an equal need for primary care visits per year. This adjustment allows us to compare, say, the population-based visit differentials between an area with a high concentration of elderly (with a higher need for visits) and an area with a high population of middle aged individuals (with a lower need for visits). The visit rates were obtained from the Medical Expenditure Panel Survey (1996) and were calculated for non-poor, white, non-Hispanic individuals. Employment status, which was included in the MEPS survey and was a significant correlate of use of service, was also intercorrelated with the other variables and was not included in the final visit calculation.

Step 2: Calculate the base population-provider ratio for regression to determine weights for need variables

With the base population in hand, we calculate the population-provider ratio to use in the regression to determine factor weights. When applying the formula for the initial estimation of weights, the number of practitioners is calculated as:

$$\begin{aligned} Providers = {} & physicians - (J1\_physicians + NHSC\_physicians + SLRP\_physicians) + \\ & .5 * \left[midlevels - (NHSC\_midlevels + SLRP\_midlevels) \right] + \\ & .1 * \left[residents - (NHSC\_residents + SLRP\_residents) \right] \end{aligned}$$

where all practitioners are measured in FTE units and the practitioner total includes NPs, PAs and CNMs weighted according to agency guidelines. The number of practitioners used in the regression to determine weights for the need variables represents only those practitioners that are considered to be the "private" supply. That is, the practitioners who would choose to practice in the community without federal support or incentives to practice in state- or federally-operated facilities. As such, government practitioners (whether federal or state) are not counted here. Community Health Center practitioners

who are not federal employees, however, are counted since many of these are not "placed" into communities but are practitioners already located in the area that are "reclassified" as CHC practitioners for later subtraction from the practitioner supply at a later step. For the estimation of the formula, an area with no practitioners is dropped from use in the regression analysis to determine weights for the need variables as a ratio is undefined (not calculable).

Step 3: Select study sample

A sample of counties and county equivalents that serve as proxies for a health care market are then selected for analysis to derive formula weights. This step was done to identify places which functioned as primary care service areas and which reported stable, reliable, usable data. According to 2000 Census data, the median county land area is 616 square miles, corresponding to an approximate radius of 14 miles. The tenth and ninetieth percentiles are 288 and 1847 square miles, corresponding to approximate radii of 10 and 24 miles respectively. The approximate radius of a county that is between the tenth and ninetieth percentiles in land area reflects a consensus of the extent of distances traveled for primary care services. The report describing PCSAs developed by Dartmouth and VCU did not identify a median or mean size rather they indicated that "A land area of 1,256 square miles or a radius of 20 miles (assuming a circular shape) was used as a crude indicator of geographically large PCSAs." (Good,man, et al., 2003 p. 297). The population threshold we proposed of 125,000 was chosen based on a perception that cities and counties with populations greater than this level were likely to have many more specialists and tertiary care services structure that would substitute for primary care alone, thus skewing the relationship between primary care practitioners and population. No specific studies were done to further support this assumption. The PCSA project reported a median population of 17,276 with multiple PCSAs exceeding that threshold. Many U.S. counties meet these general qualifications and the process selected a range of counties that met three criteria, including:

i. populations below 125,000 (410 eliminated*)
ii. area below 900 square miles (856 eliminated)

iii. base population to provider ratio below 4250 (336 eliminated)

*some counties had combinations of both values
(From file: /`area level tables – US2.xls`)

The third criterion effectively eliminated very small counties and counties with unusual distributions of health practitioners. The goal was to determine the relationship of area characteristics to practitioner supply under "normal" conditions in order to create stable estimates of those relationships in order to apply them to all appropriate populations and areas.

These sample selection criteria were varied; we tested over 2000 combinations in the estimation process described in the next step to test for robustness and sensitivity. The variations included testing within the following ranges: population 80,000-150,000; area 700-1200 sq. miles; ratio 3000-4250. Overall, the estimations derived from the models were not substantially different among the different samples The study sample contained 1643 counties. Counties were chosen because they are well-defined and are not endogenous to the current system.

Using currently designated areas would lead to biased conclusions due to the fact the subcounty areas are carefully and deliberately constructed for purposes of designation. Furthermore, dividing a county into a subcounty-designated and subcounty-undesignated would generate an extremely large number of possible observations in the analysis since the county could be divided in many different ways and into many subsets of county parts. Finally, since some data are calculated and available primarily on a county level, measurement error is minimized by using counties. Using other units of analysis requires interpolating values for subcounty and multicounty areas based on the constituent geographic units.

Step 4: Create factors

The proposed designation process, in keeping with the original MUA/MUP and HPSA approaches, identified commonly available statistics that correlated with a small number of primary care practitioners-to-population ratio. The selection of the measures was based on reviews of the scientific literature on access to care and preliminary work on the development of an alternative measures of underservice conducted by Donald H. Taylor, Jr. (Taylor & Ricketts, 1994). Candidate statistics were also suggested by a working group of State Primary Care Associations (PCAs) and Primary Care Offices (PCOs) convened by the Division of

Shortage Designation (DSD) to gather state-level input into the process of revising the method. The staff and leadership of the DSD also provided extensive input into the design. More than 20 specific variables were suggested during this process. Some candidate variables could not be used, despite being highly correlated with low access and poor health outcomes, due to lack of availability of data for small areas (e.g. lack of health insurance). Ultimately, the high intercorrelations among candidate variables restricted the calculation to 7-9 individual indicators (the actual number to be tested depended upon the specific combination of variables). The final choice of variables and the priority for inclusion in the analysis was based on the degree to which the variables best reflected underlying components of access as qualitatively assessed by the UNC-CH team, the PCA/PCO group, and staff of Bureau of Primary Health Care (BPHC). The final measures consist of demographic, economic and health status indicators (presented in Table 2).

Demographic: Population characteristics, especially racial and ethnic characteristics, have been consistently shown to affect access to primary care (Berk, Bernstein, & Taylor, 1983; Berk, Schur, & Cantor, 1995; Schur & Franco, 1999). Measures of the percent of population that is non-White and percent of population that is Hispanic were used to further adjust the ratio. The inclusion of the percentage of population older than 65 years was also included because communities with higher percentages of elderly have different community characteristics not captured in the initial population adjustment. This is likely due to the relative lack of younger people to provide supportive care and the fact that communities with declining economies, especially rural communities, have older age profiles that combine with other factors to create overall lower access.

Economic: Income and employment are very strong indicators of ability to access primary health care and to afford health insurance (Mansfield, Wilson, Kobrinski, & Mitchell, 1999; Prevention, 2000; Robert, 1999). The unemployment rate and the percent of population below 200 percent of the poverty level were used to further adjust the ratio.

Health Status: Certain populations and communities have higher than average need for health care services based primarily on their health status independent of other factors. Therefore, health status measures used to adjust the ratio include the standardized mortality ratio (General Accounting Office, 1996) and either the infant

mortality rate or the low birthweight rate (Matteson, Burr, & Marshall, 1998; O'Campo, Xue, Wang, & Caughy, 1997). These special epidemiological conditions that increase need are not fully represented in the age-gender adjustment.

Table 2. Variables Used in Creating Proposed Method

Demographic	Economic	Health Status
Percent Non-white "NONWHITE" Percent Hispanic "HISPANIC" Percent population >65 years "ELDERLY"	Percent population <200% FPL "POVERTY" Unemployment rate "UNEMPLOYMENT"	Actual/expected death rate (adj) "SMR" Low birth weight rate "LBW" Infant mortality rate "IMR"
Population density "DENSITY"		

These measures are highly intercorrelated. Table 3, below shows the Pearson-product moment correlations. The first column shows that poverty and unemployment are positively correlated (+0.64), meaning, in counties with high proportions of the population living in poverty there is usually a higher unemployment rate. Poverty and density are negatively correlated (–0.55), meaning that where there is higher density there are lower percentages of the population living in poverty. The correlation matrix is population-weighted.

Table 3: Percentile Correlation Matrix

	Poverty	Unemp	Density	Elderly	Hispanic	NonWhite	SMR	IMR	LBW
Poverty	1.00								
Unemp	0.64	1.00							
Density	-0.55	-0.21	1.00						
Elderly	0.36	0.28	-0.47	1.00					
Hispanic	-0.32	-0.23	0.22	-0.25	1.00				
NonWhite	0.10	0.12	0.22	-0.29	0.25	1.00			
SMR	0.57	0.55	-0.04	0.04	-0.26	0.42	1.00		
IMR	0.33	0.25	-0.10	0.08	-0.08	0.41	0.43	1.00	
LBW	0.40	0.37	0.05	-0.05	-0.14	0.63	0.69	0.54	1.00

Variable definitions

Variables were assigned a percentile based on the distribution of values of all US counties to all U.S. counties. This allows for continuity in the use of the proposed scores

if variables are defined differently in the future (e.g. the poverty measure is changed to 100 percent below poverty instead of 200 percent). It also allows policymakers a choice of how often (or whether) to update the percentile values without having to change the weights. If poverty conditions improve markedly across the nation, scores will tend to fall unless the percentile tables are updated. For all variables except DENSITY the theoretically worst value corresponded to the 99th percentile. At first glance, it might appear that places with very low population density would be worse off with regard to primary care access and health service needs. Places with extremely high density may also have problems caused by overcrowding and the population density may reflect problems that are commonly encountered in inner-cities. For this variable there is no apparent "right" direction for the weights. We arbitrarily specified the functional form such that lower population density corresponds to a worse off (higher percentile score) community. Accounting for the negative effects of very high density is described below.

We combined low birth weight and infant mortality into one measure (called *HEALTH)*, defined as the maximum percentile of low birth weight and the infant mortality rate for a given area. This is due to a medium level of correlation between the two and the fact that not all areas report both measures. Finally, the use of the infant mortality rate in measures of underservice is required by existing law and there is precedent for using these measures as rough substitutes. The original Index of Primary Care Shortage described in NPRM-1 of September 1, 1998 used them interchangeably.

We defined nonwhite as the maximum of zero or the percentile minus 40, so that only the top (most nonwhite) 60 percent of areas get "points" for the nonwhite variable. In other words, all areas less than the 40th percentile are treated equally. There were two main reasons for this. The first is that many of the areas have low nonwhite percentages (the 40th percentile is about 2.6 percent nonwhite). By not making this adjustment, we are differentiating areas that have little difference in the underlying measure. The second reason is that without this adjustment, the scores were not stable; small differences in the definition of this variable resulted in wide swings in the magnitude of the nonwhite variable when testing multiple randomly chosen samples. We experimented with a multitude of cutoff points (0-50 in 10 unit increments). In the final specification, small changes in the definition of NONWHITE had little substantive effect.

With the corresponding percentiles in hand, the associated scores were transformed to a logarithmic scale so that the highest derivative corresponded to the theoretically worst end of the scale. For example, the independent variable corresponding to poverty (*lnpcpov)* was defined as $lnpcpov = \ln(100 - pcpov)$ so that the fastest acceleration in the poverty score occurs at high levels of poverty rather than at low levels. In other words, we specified the model to allow a greater score to accrue to areas "moving" from the 95th percentile to the 96th percentile than to areas "moving" from the 5th percentile to the 6th percentile. All variables were assumed to have this shape (so that the theoretically worst values have the largest derivative). A more detailed description of the regression approach is included at the end of this appendix (Notes to Appendix B).

Basing the Scores on the Population-Practitioner Ratio

Although this approach specifies the *shape* of the function as logarithmic and this constrains the rate of change in the scoring as variables differ from one percentile to another, it does not constrain the *sign* nor the absolute *magnitude of the parameters that create the weights.* That is, the regression models are indifferent to whether a parameter comes out positive or negative or how large or small it is when the statistical model is run to create the weights. The magnitude is the most important parameter of the three and will be used for estimating the scores but the potential effects of the size and sign of the weights must fit into our logic of additivity of factors. The magnitude of the weights are expressed as a synthetic unit which cannot be compared to any other unit—the weight for UNEMPLOYMENT, for example, when transformed to the log-normal form and constrained to a positive value in the course of the estimation, is not a "percent of workforce not working but seeking work" but an abstract number that describes the relative contribution of that factor to a total access score at that percentile of unemployment given all the value of all the other variables and the population structure. The final model creates an estimate for the weight for each set of variables using this abstract number but that number has to be brought back into a logical relationship with the key unit of access we are using—the population portion of a practitioner-to-population ratio. The final combined sum of these abstract values has to be adjusted back to an interpretable relationship with the practitioner-population ratio. This requires that

some form of restraint on the parameter (weight) values be imposed or the solution set may produce a "best result" that causes one or two variables to dominate the weighting and others to vary from positive indicators of barriers to access to negative in various combinations.

In the application of the process this means that the parameter is used along with the intercept of the regression models to generate the specific weight for each variable. This was done to normalize the scores so that the minimum score was zero. This is done by adding a fixed number to the log result.

In an unconstrained solution of the regression models this is, indeed, the case. There are possible solution sets that include mixes of positive and negative values; in statistical parlance the functions are "two-sided." The logic of the scoring system anticipated this when we stipulated that factors which restrain use of services by creating barriers to access, also create subsequent higher levels of need likely to be met by higher levels of use, use of services that was preventable but now necessary. In the real community, both things are happening, an access program is promoting appropriate utilization by overcoming access barriers and all practitioners are involved in caring for people who are using the system because emergent conditions were not treated appropriately. The amount of the increase in use brought about by delayed care must be added into the reduction in use to produce a sum of the access "problem" in a community. To account for the "mirror" effects of these variables, the final value, the sum of the weights are doubled, to produce a population estimate that is scaled to represent the overall effect on the population need.

Factor analysis

Because many of these measures are highly correlated, we perform factor analysis in order to compute factors for the independent variables defined above. Essentially, factor analysis provides a method to translate highly correlated variables into orthogonal measures to obtain more precise estimates and minimize the impact of multicollinearity in the variables of interest. Often used as an end product statistical tool, we use it here to improve the precision of the estimates.[1]

[1] Greene (2003) (Greene W. *Econometric Analysis*, 5th Ed. Prentice Hall, New Jersey) acknowledges that the use of principal components regression is sometimes used in the presence of multicollinearity. One of

Our procedure here was to decompose the independent variables into factors and then create scores based on these factors. The factor scores follow in Table 3. The bold elements are the largest weight in the row, or on which factor the variable weighs most heavily (except for SMR, which has two maximum weights of almost equal magnitude). Four factors might be interpreted as structuring the data:

I. High health risk, nonwhite

II. Geo-demographics

III. Economic conditions

IV. Hispanic

Table 2: Factor Scores

	Factor			
Variable	1	2	3	4
Poverty	-0.005	0.208	**-0.423**	0.044
Unemp	-0.044	-0.074	**-0.338**	0.009
Elderly	-0.039	**0.355**	0.021	-0.226
Density	0.042	**0.440**	0.051	0.189
Hispanic	0.018	-0.002	0.046	**0.291**
NonWhite	**0.408**	-0.012	0.136	0.099
SMR	**0.206**	-0.107	**-0.226**	-0.124
Health	**0.353**	0.066	0.100	-0.046

Step 5: Run Regressions

We regress the base population-to-private supply practitioner ratio on the scores obtained from the factor analysis (Ratio = Factor I + Factor II … + error). By combining the scores from the factor analysis with the estimated coefficients from the regression, we obtain the effect of our underlying variables on the ratio.

As an example, the factor analysis might yield a result such as:

Variable	**factor1**	**factor2**
Poverty	.2	.4
Unemployment	.3	-.1

his criticisms is the inability to interpret the underlying regression parameters (p. 59), although this criticism is not very applicable here (the underlying parameters are never considered by the applicants.) More importantly, Greene lays out the tradeoffs: "If the data suggest that a variable is unimportant in the model, the, the theory notwithstanding, the researcher ultimately has to decide how strong the commitment is to that theory." One of the guiding principles was face validity, which essentially says conventionally accepted wisdom on important determinants of access should suggest included variables.

Which we could translate into a matrix

$$\begin{bmatrix} .2 & .4 \\ .3 & -.1 \end{bmatrix}$$

Suppose regressing the ratio onto these two scores yields estimates of

Variable	beta
Factor1	1
Factor2	-.4

which would translate to a vector

$$\begin{bmatrix} 1 \\ -.4 \end{bmatrix}$$

By multiplying these two matrices, we can obtain the total effect of one variable on the ratio:

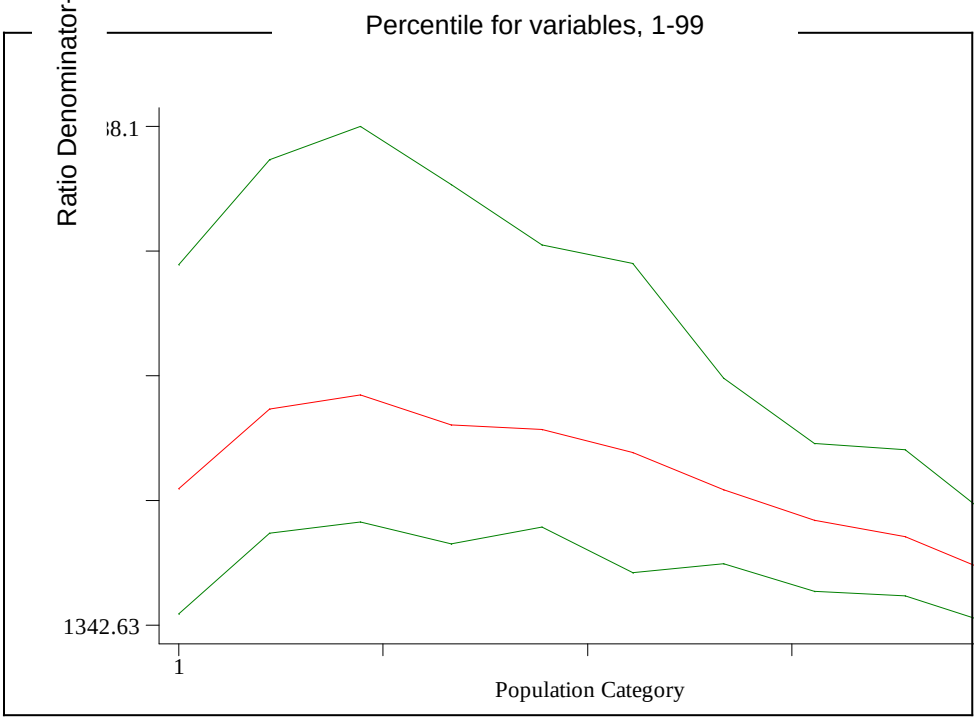
$$(1) \qquad \begin{bmatrix} .2 & .4 \\ .3 & -.1 \end{bmatrix} \times \begin{bmatrix} 1 \\ -.4 \end{bmatrix} = \begin{bmatrix} .04 \\ .34 \end{bmatrix}$$

Thus, (in this simple example) the overall effect of Poverty on the ratio is calculated as .04 and the overall effect of Unemployment is .34. We use the rightmost matrix for computing the scores (see the next section) except for one correction (see below).

Weights/Heteroskedasticity

Because the dependent variable is a ratio with population in the denominator, we are concerned about possible heteroskedasticity in the dependent variable. This is the property that the sampling variability in the dependent variable is not constant across the sample. Specifically, we expect the ratio to be estimated more precisely as the population grows. See Figure 1 below for support of this hypothesis—the ratio tends to become less variable as the population increases (population category 1 is the lowest population category and population category 10 is the highest population category). (The upper and lower bands are the values for the 25th and 75th percentiles). The consequence of this violation is that the standard errors from the regression are biased and a more efficient estimator may exist. As such, we weight the regressions by the total population of the county.

Figure **leteroskedasticity in Ratio**



There is a question of whether we are even dealing with a "sample" in the conventional statistical sense. If our analysis is composed of the population of interest, then classical statistical inference is a bit artificial; there is no uncertainty if we have data on all the units of interest. We argue that this is a sample in the conventional sense, for reasons including but not limited to the following:

a. Measurement error occurs more often then we expect. County population values are estimated in 1997 and the accuracy of provider supply is not 100 percent. As the nation observed in the presidential vote count in Florida, even simple computations are not immune from error. Thus, because the data used here are affected by measurement error we have a sample drawn from the possible data for the population of counties.

b. The units used here are a sample of a much bigger population of interest. Not only are we interested in counties other than those included in the analysis due to s ˈ criteria, ultimately we are using counties as approximations for "health c arkets" or rational primary care service areas, whether they follow the b aries of a county or not. These methods are designed to be applied to data

Population-"Private" Practitioner Ratio

for future years and the construction of the areas may vary from one based on geography to ZIP code boundaries.

Other considerations, such as errors in model specification or the discrete "lumpiness" associated with using a dependent variable like this one provide support for the use of factor scores.

Sampling error in the regression

We wish to reduce the error in predicting the designation of communities. As such, we seek to incorporate the precision with which the regression parameters are estimated into the scoring procedure. As an example, it is entirely possible, given two factors, to have one coefficient be estimated as 100 with a standard error of 1 and the other coefficient to be estimated as 400 with a standard error of 1000. If asked which factor is more important, most people would probably admit that although the 400 is a larger point estimate, the 100 is probably more important given its statistical significance. As such, the regression estimates are adjusted for the statistical significance by the algorithm defined below.[2]

1. Obtain the variance-covariance matrix V of the parameter estimates from the regression.
2. Compute the weighting matrix W defined as the inverse of the Cholesky transformation of a zero matrix except for the diagonal, which consists of the diagonal of V. (This is identical to a zero matrix with diagonal elements equal to the reciprocal of the standard errors of the parameter estimates).
3. Transform the vector of parameter estimates (omitting the constant) b by $b^* = b * W *$ number of factors/trace(W). The trace() portion of the expression ensures the weights sum to the number of factors.
4. Compute $F = S\,b^*$ as above.

As an example, return to the hypothetical results for poverty and unemployment above. Suppose the (estimated) variance-covariance matrix from the regression was

[2] An alternative treatment would be to discard any statistically insignificant estimates. We have strong conceptual biases against employing such stepwise procedures.

$$V = \begin{bmatrix} .04 & .01 \\ .01 & .49 \end{bmatrix}$$

$$\text{then } W = \begin{bmatrix} 1/\sqrt{.04} & 0 \\ 0 & 1/\sqrt{.49} \end{bmatrix} = \begin{bmatrix} 1/.2 & 0 \\ 0 & 1/.7 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 1.42857 \end{bmatrix}$$

so

$$\begin{aligned} F &= SWb * 2/trace(W) \\ &= \begin{bmatrix} .2 & .4 \\ .3 & -.1 \end{bmatrix} \times \begin{bmatrix} 5 & 0 \\ 0 & 1.42857 \end{bmatrix} \times \begin{bmatrix} 1 \\ -.4 \end{bmatrix} * \frac{2}{5 + 1.42857} \\ &= \left(\frac{2}{6.42857} \right) \begin{bmatrix} .7714 \\ 1.557 \end{bmatrix} \\ (1) \qquad &= \begin{bmatrix} .24 \\ .4844 \end{bmatrix} \end{aligned}$$

The estimated scores in equation (2) differ from those obtained in equation (1) (page 17) due to the weight. Because the regression estimate for the first factor is estimated with roughly three times the precision as the estimate for the second factor ($5/1.42 \approx 3$), the estimate for the first factor (1) is weighted more heavily than the estimate for the second factor (-.4). In this case, this has the end result of increasing the scores from .04 to .24 for poverty and .34 to .4844 for unemployment. Vector *F* is the scoring vector used in the next step.

Although the process for obtaining matrix *F* is complex and multi-stage the process was completed for all possible values of the variables. Having done this, data describing a service area can be translated readily into percentile scores using a look-up table, a simple spreadsheet, or a web-based application. This parallels the existing MUA scoring process. Applicants do not need to perform Cholesky transformations or any other mathematical calculations. Fundamentally, the "weighting" step rescales the regression parameters, placing more weight on more precisely estimated parameters. We are not aware of other published research performing this reweighting, but there are at least two reasons this approach has intuitive appeal. The reweighted models performed better empirically in the sense of minimizing disruption to current designation status. We considered dropping statistically insignificant principal components from the regression and not weighting. Although this would be a more traditional use of principal components regression (with both its advantages and disadvantages), in addition to subpar performance, the omission of insignificant components drops factors that theory suggests should contribute to access barriers. At its core, this unconventional approach

represented the best tradeoff we could devise between health care access theory, statistical theory, and empirical performance.

Task 2: Computation

Step 6: Calculate the base population-provider ratio for designation determination

Using the same age-sex adjusted population from Step 1, we calculate the population-practitioner ratio. All primary care practitioner FTEs in the area are counted to initially determine designation, this is termed the "Tier 1 designation ratio" and follows the FTE allocation of

*Providers = active non-federal, primary care physicians + 0.5 * primary care NPs, PAs, and CNMs + 0.1 * medical residents in training*

For applicants not meeting the threshold criterion, the FTEs for practitioners who are supported by safety net programs (e.g., NHSC providers, J-1 visa practitioners, CHC providers) are subtracted from the supply total and the applicant ratio is compared to the threshold. That step is termed " Tier 2 designations." The formula for that calculation follows the same logic as in Step 2, above:

$$\begin{aligned} Providers = {} & physicians - (J1\_physicians + NHSC\_physicians + SLRP\_physicians) + \\ & .5 * \left[midlevels - (NHSC\_midlevels + SLRP\_midlevels) \right] + \\ & .1 * \left[residents - (NHSC\_residents + SLRP\_residents) \right] \end{aligned}$$

Step 7: Calculate Scores

With row vector *F* in hand, we then turn to computing scores for geographical units. We compute the ratio of population to providers using the algorithm outlined above. We use the percentile scores as computed above for the counties. See the document "Completing the NPRM2 Application" for these percentiles.

We then calculate the score for the communities and add this score, upweighted by 2 to account for the 2-sided properties of the regression estimates so the total score for the community equals

*ADJUSTED RATIO (or "INDEX") = RATIO + 2 * SCORE*

This is the total score for the community and determines its designation status. The applicants never see the regression multiplier; it is embedded in the tables.

Because the use of the multiplier for the score is applied at this stage of the process, it may be seen as an ad-hoc adjustment. The statistical logic for this has been described above, the policy logic for applying this adjustment is supported by these points:

1. The multiplier is used to account for the fact that the existing measures and processes including: the HPSA formula, the IPCS/MUA formulae, and the practical application of the CHC/RHC clinic placement process—all recognize the importance of the basic population-to-practitioner ratio in determining need. Indeed, some simple models run on the study sample provide evidence that the multiplier should be closer to 10 rather than 2 if the goal were to include every area containing a CHC under the proposed designation process (this assumes that the presence of a CHC is an indicator of need in and of itself as opposed to the result of the calculation of pre-existing unmet need). The IPCS mechanism provided for a maximum score *from the population-practitioner ratio* of 35 points. The maximum score available from other factors (poverty 35 points, IMR/LBW 5 points, minority 5 points, Hispanic 5 points, LI 5 points, density 10 points = 65 points) are, collectively, almost twice that in terms of potential contribution. Thus, the weighted contribution of the factors besides the ratio is roughly twice that of the ratio itself. Multiplying the ratio denominator by two intensifies the relative effect of the underlying, basic population to practitioner ratio in the designation process providing continuity with prior policy.
2. The multiplier functions as a scale /weighting factor. The score has a much smaller variance than the ratio. This is not just an annoyance—it is used to generate a prediction, and thus will have smaller variance than the dependent variable. The dependent variable and the score used here have some sort of meaning, a person per provider, although the various adjustments make this unit of measurement not as meaningful as we might think. One alternative we considered is rescaling the ratio and the score into *z*-scores and using these standardized measures rather than the unscaled measures. This rescaling would involve multiplying the score by a larger factor than the ratio.
3. The multiplier helps control for the (observed) low ratios in, (eg, metro) areas with high scores. The following example illustrates this:

Table 3: Example score and ratios

County of HPSA	State	Ratio	Score	IPCS	IMR	LBW	Poverty
Bronx	NY	1357.2	1043.5	54	10.1	10.1	77.8
Coconino	AZ	1266.8	1005.6	56	8.1	7.2	65.1
Kings	NY	1634.7	897.8	52	10.3	9.2	59.2
East Baton Rouge	LA	1660.5	874	46	11.3	10.2	69
St. Lucie	FL	1138.5	873	44	10	7.3	67
Philadelphia	PA	1055.9	861.2	47	13.3	11.4	61.1
Mahoning	OH	1505.3	839.3	44	10.7	8.9	67.5

The (unmultiplied) maximum score is about 1300. The areas listed above are all in the worst 10 percent of scores. Note that these areas would not qualify without the "score x 2" multiplier rule (see below). Perhaps the ratio is a misleading measure in some circumstances.

4. The multiplier fills a statistical role. The score is (likely) more stable across years; e.g., if one physician moves out of a rural area, the ratio varies dramatically. The score is not going to change drastically across years. Thus, it should be given more weight.
5. The multiplier creates a standard which designates roughly the same number of people as the IPCS and the current HPSA designations.
6. It performs better than without the doubling. Although this particular argument has little theoretical basis, it is still compelling

Why is a portion of the density score function negative?

The astute reader will note that the constant from the regression was dropped and never used. The reason for this is that the constant has no clear meaning in this context. We decided to norm the scores so that the minimum score—that is, the best area in the country—was zero. Thus, although *in theory* an area could receive a negative score if it had very favorable demographics and had a high population density, *in practice* no area had a negative score (by definition).

Step 8: Compare to Threshold

Areas are designated if and only if the "adjusted ratio" (or ratio+score) is greater than 3000. This threshold was adopted for its reflection of the clear need for a single full-time equivalent primary care physicians, its consistency with prior threshold values, and its familiarity to stakeholders.

Areas with No Practitioners

The problem of how to treat areas with zero providers emerged early in the process of ranking areas as medically underserved. There is an informative treatment of the phenomenon in Black and Chui (1981).* For areas with zero providers, we have not made any firm recommendations and have treated them in one of three ways for various parts of the analysis

(a) Every area with zero providers automatically gets an adjusted ratio of 3000 (which guarantees them designation), to which a score for community need indicators are added. This results in all areas having a NPRM2 score, including areas with zero providers. This method was used in early tabulations and compilations.

(b) Automatically designate areas with zero providers without assigning an adjusted ratio or a score for community need indicators. Therefore, areas with zero providers will not have a NPRM2 total score. This has occurred when calculations and tabulations of the database using the NPRM2 scoring system was applied. The places with no score were dropped. This method was used in the final impact analysis.

(c) Assigning an arbitrarily small FTE to the area, such as 0.1 to create a score that is primarily dependent upon the denominator population. This was used only in selected tests of the scoring system as an alternative.

* Black, R. A., and Chui, K.-F. (1981). Comparing schemes to rank areas according to degree of health manpower shortage. Inquiry, 18(3), 274-280.

Notes to Appendix B: Regression approach for assignment of weights to correlates of "shortage"

The basic method for assigning weights to individual variables involved the estimation of a county-level linear regression model with the adjusted population-to-physician ratio as the left-hand side variable, and the variables described in step 4 as right-hand side variables. Coefficients on the right-hand side variables can be interpreted directly as average differences in the population-to-physician ratio for counties with specified characteristics relative to counties without those characteristics.

To reduce the effects of extreme outliers (e.g., population density in New York City, or per capita income in Silicon Valley), all variables were converted into percentages. To allow for non-linear relationships between each variable and the ratio, the variables were further converted from a linear variable, ranging from 1 to 100, into twenty five-percentile categorical variables, i.e., one each for 1-5th percentile, 6-10th percentile, … 96th-100th percentile. When all but one of these variables are entered on the right-hand side of a regression with the population-to-physician ratio as the dependent variable, the coefficients on each variable represent the average difference in the adjusted population-to-physician ratio relative to the omitted reference category. In most cases, the omitted reference category is the 1-5th percentile, i.e., the five percent of counties with the lowest values for a particular variable.

Entering highly collinear variables, such as income and poverty, into a single regression model usually results in one coefficient being positive, and the other being negative. In order to develop a "user-friendly" scoring system in which all weights are positive, variables were added sequentially to the regression model, with the effects of previously entered variables constrained to their estimated effects. As a result, coefficients on all variable other than the first represent the "marginal differences" in the ratio, after controlling for all previously included variables.

A decision was made to use a population-to-physician ratio of 3000:1 as a cutoff criterion for designation. The following analysis was restricted to counties with adjusted population-to-physician ratios between 500:1 and 3000:1, for which the dependent variables was not missing (N=2,493).

Income was the single most important correlate of the ratio. It was entered first, and estimates were obtained for each of 19 categories; counties in the 95-100th percentile were the excluded category. Each of the estimated coefficients represents the average difference in the ratio for counties in the respective percentile range relative to the omitted group of counties with the highest income. Coefficients were graphed and examined visually, and differences between the coefficients for "neighboring" categories were evaluated for statistical significance. Categories with no statistically significant differences were combined into single variables. As a result of this process, three categories (plus reference category) remained, one each for the 1-75th, 76-85th, and 86-95th percentiles. The regression was run again, suggesting that counties in these categories had higher ratios by 628, 344, and 216 "units", respectively. (These units are the average differences in the population-to-physician ratio).

Constraining the coefficients on these variables to these values, 19 percentile ranges for the next-highest correlate of the ratio, population density, were added to the analysis. Visual inspection pointed to clear non-linearities in the relationship. There appeared to be a statistically significant difference between counties in the 95-100th percentile relative to all other counties. Furthermore, the effect was increasing up to the 35th percentile of counties, and then decreased between the 36th and 95th percentiles. Note that these relationships describe the relationship between population density and the population-to-physician ratio *after controlling for the effects of income*. Consistent with the observed relationship, three variables were defined, a categorical variable for the 1-95th percentile range, and two splines for the 1-35th and 36-95th percentiles, respectively.

These three variables describing population density were entered into the model together with the income variables, and the estimated coefficients were used to analyze the marginal effect of unemployment according to the same method. Relative to the omitted reference group of counties in the 1-5th percentile, counties in the 6-20th and 21-100th percentile ranges had significantly higher population-to-physician ratios, after controlling for income and population density. Consequently, two dummy variables for counties in these categories were entered into the model. The process was repeated for percent of the population under 200% FPL, which suggested that — after controlling for income, population density, and unemployment — the ratio was lowest for counties with

a percentage of the population below 200% poverty around the 20th percentile of all counties. Below this threshold, the average ratio was higher by about 110 "units", above that, the ratio gradually increased by about 2.5 "units" per percentile increment.

Table 2 shows the results of the final regression model containing the four variables described above. After controlling for these variables, none of the remaining variables was significantly associated with shortage. This finding is consistent with other studies of the effects of community characteristics on access to health care, in that the economic/barrier variables have been shown to have much greater impact than other characteristics. However, legislation requires the use of selected morbidity and mortality measures such as infant mortality and, even if marginal in their net effect, these measures are tied closely to the logic of need for primary care and access to primary care.

To comply with this requirement, the analysis was repeated for actual/expected deaths, the maximum of low birth weight / infant mortality rate, and the percentage of the population over the age of 65. Table 3 shows the results of the final regression model and the specification of each variable. The coefficient estimates in Tables 2 and 3 were used to create a single table containing the weights associated with each variable, for each percentile increment, usually rounding to the nearest increment of 5.

Table 2. Coefficient estimates for economic / barrier correlates of shortage

Correlate of Shortage	Cutoffs (percentiles)	Specification	Coefficient	SE	t
Income	0 - 74	Dummy Variable	355.9	59.3	5.997
	75 - 84	Dummy Variable	186.0	59.6	3.121
	85 - 84	Dummy Variable	69.7	53.6	1.301
Population Density	0 - 95	Dummy Variable	318.6	51.4	6.197
	0 - 35	Spline	4.23	0.95	4.432
	35 - 95	Spline	-3.73	0.84	-4.467
Unemployment	5 - 19	Dummy Variable	167.8	52.0	3.228
	20 - 99	Dummy Variable	245.4	48.0	5.110
Below 200% FPL	0 - 14	Dummy Variable	109.0	38.8	2.807
	15 - 99	Spline	2.36	0.54	4.406
Constant			732.0	78.7	9.297

Table 3. Coefficient estimates for health / demographic correlates of shortage

Correlate of Shortage	Cutoffs (percentiles)	Specification			
Actual/Expected Deaths	6 - 15	Dummy Variable	66.4	64.0	1.038
	16 - 55	Dummy Variable	121.6	57.2	2.124

	56 - 75	Dummy Variable	211.2	59.4	3.554
	76 - 100	Dummy Variable	278.5	60.2	4.625
Infant Morality	81 - 100	Dummy Variable	65.73	27.41	2.398
Percent 65+	1 - 100	Continuous	1.93	0.37	5.161
Constant			1364.4	57.2	23.872