A 'Virtual Population' Approach To Small Area Estimation

Michael P. Battaglia[1], Martin R. Frankel[2], Machell Town[3] and Lina S. Balluz[3]
[1] Abt Associates Inc., Cambridge MA 02138
[2] Baruch College, CUNY, New York City, NY 10010
[3] Centers for Disease Control and Prevention, Atlanta GA 30329

Abstract

A small area estimation system is developed for the Behavioral Risk Factor Surveillance System. The BRFSS is a state-based health survey but there is a need for county estimates. The 2005-2009 American Community Survey PUMS is used to create a "population" of adults for each county. Iterative probability adjustment is used to ensure that the county estimates agree with the direct tabulated "official" estimates from the survey for key state and sub-state domains.

Key Words: Iterative Probability Adjustment, American Community Survey, Behavioral Risk Factor Surveillance System

1. Introduction and Background

One of the most difficult problems associated with the development of small area estimates is related to the "inconsistency" of the various small area estimates when compared with those produced by the primary survey from which they are derived. For example, consider a state that is subdivided into 60 counties. Suppose further that the sample design calls a total sample size of 5000 interviews, with stratification consisting of 7 regions, with a minimum sample size of 500 within each region. The post-stratification/weighting process explicitly recognizes these 7 regions, so it is possible to produce region level estimates which are weighted for region specific population characteristics of Age, Gender, and Race. By making use of explicit region level weighting, users are assured that when separate survey estimates are produced for these regions they will be viewed as "representative" with respect to the characteristics used in the weighting process. Furthermore, when users combine a subset of these regions, direct survey tabulations will be consistent with those obtained by "adding up" separate region estimates. This will be true in total and when the tabulation or addition is carried out on the basis of the categories used in the weighting process. For example, estimates for females on a region by region basis will be consistent with estimates produced by tabulation from various aggregations of regions up to and including the full state.

Suppose now, that estimates are desired for geographic sub-areas (e.g., counties) within each of the regions used in stratification, sample control and weighting. There are several approaches that may be applied when considering the development of estimates within geographic sub-region. They are as follows:

A. Direct tabulation (using the existing weights) from the interviews within each of the geographic sub-regions.
B. Re-weighting the interviews within the sub-regions, using some of the region level population estimates as controls.

C. Development of model based estimates within the geographic sub-regions using method that do not involve direct tabulation of survey observations.
D. Development of estimates based on a weighted combination of survey data (either A or B) and model based estimates.
E. Suppression of codes required for geographic sub-region tabulation. While this does not produce estimates, it in fact prevents these estimates from being produced.

It should be noted that each of the above approaches has obvious and subtle drawbacks. Some of these drawbacks are statistical while others impact the "perceived" validity and credibility of the survey.

In addition to small sub-region sample size problems, approach A, (involving direct tabulation from the weighted data) produces within sub-region estimates that will "add-up" to the total region but may have obvious credibility related flaws (e.g. the weighted survey data for a certain sub-region may have a 80:20 female to male ratio, when it is known that the true gender ratio within the sub-region is close to 50:50. Further, the projected population size within any sub-region may be quite different from the actual population size.

Approach B, which involves within region sub-region re-weighting will produce estimates that may be credible, at the sub-region level but when these re-weighted estimates from the sub-regions are added together to the region level they will differ from the region estimates using the original weights. Thus there will different estimates for the total region. While this situation may be statistically acceptable, it will often produce serious questions about overall survey credibility on the part of final data users and analysts. Approach B is also limited by small sub-region sample size problems.

Approach C, which involves the development of model based estimates will generally result in sub-region estimates that have certain desirable statistical properties, but will often fail in the add-up or consistency test. For example, if we produce age by gender sub-region estimates they will not add up to (i.e. be consistent with) the model based sub-region estimates. Furthermore, when these sub-region estimates are added to form a region level estimate, they will generally not agree with the published (i.e., "official") estimates produced by direct tabulation. Again, this may be "acceptable" to technically trained survey statisticians, but will generally produce strong questions about validity on the part of policy makers and subject matter analysts.

Approach D suffers from the same drawbacks as approaches A – C. Approach E avoids these issues, but limits the "usefulness" and suitability of survey.

We have developed a method for producing estimates on a sub-region (Small Area level that addresses some of the drawbacks and limitations outlined above. It allows for the production of estimates for sub-regions (Small Areas) that are both internally consistent and consistent with survey based estimates at the region level. The methodology for producing these estimates involves the development of a "virtual population" based on an externally available large scale survey and/or census and the use of model based probability estimates that are iteratively adjusted (Iterative Probability Adjustment) to conform with pre-specified control totals obtained from direct survey tabulation. As a result if users of these small area estimates add them up to areas where direct survey estimates are produced, the two estimates will be in agreement (i.e. they will be consistent).

While the methods used in this system are not new to survey statistics, we believe that the particular combination of these methods in order to produce consistent small area estimates has not been previously published.

In this paper we will first provide a general description of the steps used in developing the estimates. This will be followed by a more detailed explanation of the specific steps followed to produce various county estimates. It is assumed that the basic sample design, sample weighted process and final estimation process have been specified.

Step 1. Determine the geographic level for which small area estimates are desired and specify the structure of these small areas within areas for which direct estimates are deemed appropriate. Determine if modifications to the overall survey weighting process should be undertaken in order to establish appropriate control constraints for aggregations of the small area estimates. For example, within a state it might be reasonable to obtain estimates from direct tabulation for all Metropolitan Areas (and within Metropolitan Area sub-components) where the sample size exceeds 500. If this is the case, then the overall weighting process should be modified to assure that appropriate post-stratification controls are applied at this level. The reason for this step is to make sure that in the process of aggregation of the small area estimates to areas where direct tabulations are possible are not inconsistent because of lack of controls at the level for which direct estimates are to be produced. For example, if a particular estimate is highly correlated with gender, then if small area estimates are to be consistent with tabulated estimates, post-stratification by gender is desirable at the level for which direct estimates are to be produced.

Step 2. Identify a "larger" data set to be used as the "virtual population." While it is possible to develop a large virtual population by simulation, we have found that it is more desirable to make use of a larger data set which for which basic demographic and other variables are based on actual observations. In developing small area estimates of health conditions, we have found that either the decennial census or American Community Survey (ACS) public use micro data (PUMS) are quite suitable to form the basis of a "virtual population", which we define as a data set with reasonably large numbers of observations with the small areas for which estimates are to be produced.

Step 3. If appropriate apply additional weights to the elements of the virtual population in order to assure consistency with the overall and sub area weighting controls used for the survey. Create "simulated" sub-areas if necessary.

Step 4. Develop basic prediction models for the desired estimates based on prediction variables that are available in the survey and the "virtual population". For example these might be demographic variables collected on a respondent level or may be variables available from outside sources. For example, demographic variables might be age, gender, education, marital status, while external variables might be "number of doctors per person" within the county. We have restricted our estimate to binary (0,1) variables and have made use of logistic regression for the prediction process.

Step 5. Apply the model to the elements of the virtual population in order to obtain a predicted probability for each element. Aggregate these probabilities to the control levels for which direct estimates are produced and apply iterative probability adjustments to these probabilities. In general this adjustment process will involve the total sample and various demographic subgroups which estimates are to be available. For example, we

have produced small area estimates of binary health conditions by gender, age group, education and marital status.

Step 6. Using the adjusted probabilities, produce the small area estimates.

County estimates are developed for behavioral risk factors, health conditions, and access to care measures using the 2009 Behavioral Risk Factor Surveillance System. The BRFSS is the largest health survey in the U.S. The BRFSS is conducted annually in each of the 50 states and the District of Columbia by the Centers for Disease Control and Prevention. This state-based survey is conducted by telephone with a sample of adults (age 18+) using random-digit-dialing. The BRFSS questionnaire consist of a core module that collects basic risk factor and health condition data such as general health, health care coverage, smoking, alcohol use, asthma and BMI, as well as socio-demographic characteristics such as age gender race/ethnicity and education. The core section is followed by one or more topic-specific modules.

2. Creating an Unified Weight for the 2009 BRFSS

Purpose: Provide direct sample health risk factor prevalence estimates for all BRFSS domains and subclasses using one set of weights.

The weighting methodology for the BRFSS landline telephone state samples involves the calculation of a design weight for each completed adult interview. The design weight incorporates the selection probability of the telephone number, the number of voice-use landline telephone numbers in the household, and the number of adults in the household. The design weights are then post-stratified to control totals obtained from Claritas Inc. and from the latest 3-year American Community Survey PUMS. The control totals for each state include age by gender, race/ethnicity, marital status, education, landline telephone service interruption, age by race/ethnicity, and gender by race/ethnicity. For those states that are divided into regional strata additional control totals are employed for region, region by age, region by gender, and region by race/ethnicity (Battaglia et al. 2008). Category collapsing rules are used to avoid small sample size categories. Raking ratio estimation is used to calculate the final interview weights for each state. The raking algorithm incorporates a weight trimming procedure to prevent extreme high or low weights (Izrael et al. 2009).

The *SMART BRFSS* identifies counties and metropolitan statistical areas that meet minimum annual sample size criteria. Direct sample estimates are provided for these geographic areas. For our purposes we defined SMART geographic areas as counties or MSAs with a minimum of 500 adult interviews which is very similar to the definition used by the BRFSS. For an MSA that cuts across state boundaries, the part inside one of the five states needed to contain at least 500 interviews

The adult interviews in each SMART geographic area are weighted to Claritas Inc. control totals for age, gender and in some cases race/ethnicity using cell poststratification. A separate set of SMART weights are used to provide risk factor and health condition prevalence estimates for each SMART geographic area. Because the weighting methodology for the state BRFSS is different from the weighting methodology for the *SMART BRFSS*, the prevalence estimates will differ. For example, if a region contained four counties that were all classified as SMART counties, then the regional prevalence estimates from aggregating the four counties would not agree with the region prevalence estimates from the BRFSS state sample weighting.

Before developing the small area estimates for each county in a state we first developed a unified BRFSS weighting methodology that extended the raking methodology described above by adding additional margins for: 1) SMART county (with a residual category for the balance of the state), 2) SMART county by age, 3) SMART county by gender, 4) SMART county by race/ethnicity, 5) SMART MSA (with a residual category for the balance of the state), 6) SMART MSA by age, 7) SMART MSA by gender, and 8) SMART MSA by race/ethnicity. Category collapsing rules are used to avoid small sample size categories. We also used weight trimming during the raking avoid extremely large or small weights. The weight trimming during the raking (Izrael et al. 2009) involves placing global and individual constraints on the weights:

Global low weight cap value = mean raking input weight times 0.091
Global high weight cap value = mean raking input weight times 11.0
Individual low weight cap value = respondent's weight times 0.2
Individual high weight cap value = respondent's weight times 5.0

We calculated the unified weight for the adult interviews in each of the five states.

3. Creating a "Virtual Population" of Adults for Each County

Purpose: Create large sample size county data sets of adults for use in small area estimation and ensure that the county and state distributions for key socio-demographic variables align with population control totals.

The 5-year 2005-2009 ACS PUMS provides an extremely large sample of adults living in 9,689,251 households in the 50 states and the District of Columbia. For the five states the sample sizes are:

State	Sample Size of Adults
California	1,266,424
Connecticut	129,485
Minnesota	193,799
North Carolina	339,707
Texas	821,980

We created a "virtual population" of adults for each county in the five states. The external ACS PUMS does not contain county identifiers. The PUMAs included in the ACS PUMS were therefore mapped to counties using information available from the Missouri State Data Center website (http://mcdc2.missouri.edu/websas/geocorr2k.html). They provide population estimates for PUMA/county intersections. Using their information we created a mapping of PUMAs to counties that consisted of four categories:

One-to-one mapping between the PUMA and the county,
Two or more PUMAs cover the county,
One PUMA covers two or more counties, and
All remaining relationships between PUMAs and the county.

An example of the fourth category is a county covered by three PUMAs. Two of the PUMAs only cover the county but the third PUMA also covers an adjacent county. For

each of the three intersections we have the estimated population from the Missouri State Data Center website.

Using Connecticut as an example we show the size of the "virtual population" of each county and the mapping category number:

County FIPS Code	Sample Size of Adults	Mapping Category Number
09001	32,085	2
09003	32,733	2
09005	8,771	1
09007	6,059	1
09009	29,533	2
09011	10,477	2
09013	5,212	1
09015	4,615	1

4. Reweighting the ACS PUMS

For each county in the U.S. the Census Bureau Population Estimates Program provides annual population estimates by age, gender and race/ethnicity. However, as discussed above the BRFSS has used Claritas Inc. as the source of its age, gender and race/ethnicity control totals for the state and SMART geography poststratification. These control totals are also available at the county level.

Using the ACS person weight as the raking input weight we carried out a separate raking for each county in a state to bring the weighted distribution of the virtual population of adults into close agreement with the Claritas Inc. control totals for age by gender and for race/ethnicity with collapsing of small race/ethnicity categories in a county. For mapping categories 1 to 3 we used the ACS person weight as the raking input weight. For mapping category 4 we first adjusted the ACS person weights using the estimated population of the PUMS/county intersection as a proportional weighing factor, where the proportions sum to one.

For a given county the ACS person weights can exhibit considerable variability due to the ACS sub-county geographic sample allocation procedures. To avoid ending up with adjusted ACS person weights that exhibited considerable variability, we developed a weight trimming approach that involved conducting a minimum of two rakings in each county using the following rules for trimming high weight values:

1. Raking # 0: No weight trimming. Convergence criterion = maximum percentage point difference of 0.1.

Calculate Upper half way point 1 = (Maximum weight / Mean weight) / 2.

2. Raking # 1: Raking with Global High Weight Cap Value.

Global high weight cap value = Upper half way point 1

If the raking converges for the county go to step 3.

If the raking does not converge within 25 iterations go to step 4.

3. Calculate Upper half way point -1 = Upper half way point 1 / 2.

Raking # -1: Raking with Global High Weight Cap Value.

Global high weight cap value = Upper half way point -1

If the raking converges go to step 5.

If the raking does not converge within 25 iterations stop and save the raking weight from raking # 1.

4. Calculate Upper half way point 2 = Upper half way point 1 + (Upper half way point 1 / 2).

Raking # 2: Raking with Global High Weight Cap Value.

Global high weight cap value = Upper half way point 2

If the raking does not converge within 25 iterations go to step 6.

If the raking converges stop and save the raking weight from raking # 2.

5. Calculate Upper half way point -2 = Upper half way point -1 / 2.

Raking # -2: Raking with Global High Weight Cap Value.

Global high weight cap value = Upper half way point -2

If the raking converges go to step 7

If the raking does not converge within 25 iterations stop and save the raking weight from raking # -1.

6. Calculate Upper half way point 3 = Upper half way point 2 + (Upper half way point 2 / 2).

Raking # 3: Raking with Global High Weight Cap Value.

Global high weight cap value = Upper half way point 3

If the raking does not converge within 25 iterations stop and flag this county as a non-convergence county.

If the raking converges stop and save the raking weight from raking #3.

7. Calculate Upper half way point -3 = Upper half way point -2 / 2.

Raking # -3: Raking with Global High Weight Cap Value.

Global high weight cap value = Upper half way point -3

If the raking converges stop here and save the raking weight from raking # -3.

If the raking does not converge within 25 iterations stop and save the raking weight from raking #-2.

After completing all of the county raking in a state a state level raking was conducted to obtain a final adjusted ACS person weight for each adult in each county. This ensures that the BRFSS and the ACS PUMS have the same weighted distributions on key socio-demographic variables (Battaglia et al. 2009). Using the adjusted county ACS raked weight input weight, we raked to the BRFSS control totals for:

Age by gender
Race/ethnicity
Education
Marital status
Gender by race/ethnicity
Age by race/ethnicity
Region

To maintain the county level controls we also included raking margins for:

County by age by gender
County by race/ethnicity

No weight trimming was employed for the state level raking, because the weighted ACS margins were typically very close to marginal the control totals.

5. Eleven Risk Factor and Health Condition Dependent Variables

Purpose: Develop state level logistic regression models predicting health risk factor dependent variables.

The BRFSS questionnaire includes several key health risk factors, obtains information on access to health care, and also obtains self-reports on specific health conditions. Eleven key variables were selected for the development of county prevalence estimates:

Current smoking

Current asthma

Binge drinking

Obese

Fair/Poor health

Diabetes

No physical activity

No health care coverage

No medical home

Delayed medical care due to cost reasons

No checkup in past 12 months

We created eleven dichotomous dependent variables. The BRFSS has a very low level of missing data on most of the health-related questions in the survey. In order to have the same sample count in state for all eleven dependent variables we used a single imputation hot deck procedure with imputation cells formed on the basis of age group, gender and race/ethnicity.

Logistic regression predictor variables fell into two categories. The first category included individual level socio-demographic predictors that are available in the BRFSS and in the ACS PUMS. The second category consisted of county level variables obtained from the Area Resource File and from County Business Patterns. The predictor variables are shown in the table below.

Individual Level Predictors:	
Gender	1 Male
	2 Female
Number of adult males in household	0 Adult Men
	1 Adult Men
	2+ Adult Men
Number of adult females in the household	0 Adult Women
	1 Adult Women
	2+ Adult Women
Marital status	1 Married
	2 Divorced/Separated
	3 Widowed
	4 Never married
Number of children in the household	1: 0 Children
	2: 1 Child
	3: 2 Children
	4: 3+ Children
Education	1 Less than High School
	2 High School graduate
	3 Some College
	4 College graduate
Race/ethnicity	White, nonHispanic
	Black, nonHispanic
	Hispanic
	Asian, American Indian, Other, nonHispanic
Region (example of state with 5 regions)	Region 1
	Region 2
	Region 3
	Region 4
	Region 5

Age Group	1: 18-24
	2: 25-29
	3: 30-34
	4: 35-39
	5: 40-44
	6: 45-49
	7: 50-54
	8: 55-59
	9: 60-64
	10: 65-69
	11: 70-74
	12: 75-79
	13: 80-99
County level predictors:	
Low Education Typology (25 percent or more of residents 25 through 64 years old had neither a high school diploma nor GED)	0 (county does not meet typology) 1 (county meets typology)
Low Employment Typology (Less than 65 percent of residents 21 through 64 years old were employed)	0 (county does not meet typology) 1 (county meets typology)
Housing Stress Typology (30 percent or more of households had one or more of these housing conditions: lacked complete plumbing, lacked complete kitchen, paid 30 percent or more of income for owner costs or rent, or had more than 1 person per room)	0 (county does not meet typology) 1 (county meets typology)
Population Loss Typology (Number of residents declined both between the 1980 and 1990 censuses and between the 1990 and 2000 censuses)	0 (county does not meet typology) 1 (county meets typology)
County Births	Total births per 100,000 population
County Black Population	Percent of population African-American
County Deaths	Total deaths per 100,000 population
County Dentists	Count of total private practice, non-Federal dentists per 100,000 population
County Emergency Room Visits	Count of emergency room visits at short-term general hospitals per 100,000 population
County Hispanic Population	Percent of population Hispanic/Latino
County Hospital Admissions	Count of admissions at short-term general hospitals per 100,000 population
County Hospital Beds	Count of hospital beds at short-term general hospitals per 100,000 population

County Hospitals	Count of short-term general hospitals per 100,000 population
County MDs	Count of general practice MDs (General Practice and Family Medicine, patient care, office based, non-Federal) per 100,000 population
County Poverty	Percent of persons in poverty
County Medical Specialists	Count of medical specialist MDs (Allergy and Immunology, Cardiovascular Disease, Dermatology, Gastroenterology, Internal Medicine, Pediatrics, Pediatric Cardiology and Pulmonary Disease, patient care, office based, non-Federal) per 100,000 population
County Liquor Stores	Beer, wine & liquor stores per 100,000 population
County Fitness Establishments	Fitness & recreation sports centers per 100,000 population
County Fast Food Establishments	Limited-service eating places per 100,000 population

The BRFSS state sample sizes are large enough to allow for the fitting of state specific logistic regression models. For each state main effect weighted logistic regression models were fit for each dependent variable. For the 55 models we found that the individual level predictors were much more likely to be statistically significant than the county level predictors, reflecting the primary importance of individual characteristics in determining risk factors and health conditions.

6. Iterative Probability Adjustment

Purpose: Adjust health risk factor predicted probabilities of adults in each county in a state so that the small area estimates are in agreement with the direct sample estimates for all key domains in that state.

For a given state we can use the coefficients from a logistic regression model to assign predicted probabilities, $prob_i$, for that dependent variable to each adult in the "virtual population" of each county in the state. Because each adult in the "virtual population" also has an adjusted ACS person weight, $WTACS_i$, we can estimate the state prevalence of that risk factor:

$$\hat{P}_{ACS} = \frac{\sum_i WTACS_i \, prob_i}{\sum_i WTACS_i} = \frac{\hat{Y}_{ACS}}{X}$$

as well as the county prevalence for county h:

$$\hat{P}_{ACSh} = \frac{\sum_i WTACS_{hi}\, prob_{hi}}{\sum_i WTACS_{hi}} = \frac{\hat{Y}_{ACSh}}{X_h}$$

Estimates for key domains (e.g., gender) within a county or at the state or region level can also be calculated. One problem with using the predicted probabilities from the logistic regression models is that the aggregation of the small area estimates will typically not agree with the direct sample estimates. For example, if the aggregate the prevalence estimates for all the counties in the state it is very unlikely that the state prevalence will agree with the direct sample estimate from the BRFSS. Also, if we aggregate the estimates for males within each county, the state level prevalence estimate for males will likely differ from the direct sample estimate for males from the BRFSS.

For a single domain such as gender one can calibrate the predicted probabilities to ensure that the prevalence estimates for males and females are in agreement with the direct sample BRFSS estimates. The BRFSS however has several key domains:

SMART counties
SMART MSAs
Regions
Marital status
Education
Race/ethnicity
Gender
Age group

We developed an iterative probability adjustment (IPA) algorithm to calculate adjusted predicted probabilities for the adults in each state for a given risk factor variable. The approach is iterative in that it terminates after 10 iterations or when the maximum difference between an adjusted prevalence estimate and the direct BRFSS prevalence estimate for the domain is 0.1 percentage points or less.

IPA consists of two main steps. First we ratio adjust the predicted probabilities of all adults in a state so that $\hat{Y}_{ACS}$ equals $\hat{Y}_{BRFSS}$, and therefore $\hat{P}_{ACS} = \hat{P}_{BRFSS}$. If any adults have a predicted probability that is ratio adjusted to a value greater than one, we truncate that adjusted predicted probability to one, and proportionately allocate the reduction in $\hat{Y}_{ACS}$ to the remaining adults in the state. The adjusted predicted probability of adult i in county h is $pred_{hi0}$.

Second, an iteration is defined as the sequential adjustment of the predicted probabilities in the order of 8 domain variables listed above. Starting with the adjusted predicted probabilities from step 1, we ratio adjust the predicted probabilities for the adults in each

SMART county domain d (plus a residual category for the balance of the state) so that the ACS prevalence estimate is inexact agreement with the BRFSS direct sample prevalence estimates:

$$pred_{dhi1} = pred_{dhi0}\left(\frac{\hat{P}_{BRFSSd}}{\left(\frac{\sum_d \sum_i WTASC_{dhi}\, pred_{dhi0}}{\sum_d \sum_i WTASC_{dhi}}\right)}\right)$$

We then use these adjusted predicted probabilities as the input probabilities into the ratio adjustment step for the SMART MSAs in the state (plus a residual category for the balance of the state). At this point the ACS prevalence estimates for the SMART MSAs and the balance of the state are in exact agreement with the BRFSS direct sample prevalence estimates. We continue the first iteration by moving on to the region variable and work our way to the end of the first iteration by adjusting the predicted probabilities within each of the age groups. For each domain variable if any adults have a predicted probability that is ratio adjusted to a value greater than one, we truncate that adjusted predicted probability to one, and proportionately allocate the reduction in $\hat{Y}_{ACSd}$ to the remaining adults in domain d. We continue this process with iteration two by returning to the SMART county domain variable. The IPA continues until we complete 10 iterations or the maximum difference for any domain is less than 0.1 percentage points.

7. County Prevalence Estimates

The IPA yields an ACS PUMS data set with 11 predicted probability variables for each adult in a county. The ACS PUMS also contains the final adjusted ACS person weight. The county prevalence estimates are calculated as weighted proportions for all adults. Our approach also allows for the calculation of county prevalence estimates for key domains such as age group, gender, race/ethnicity, education and marital status.

Using Texas as an example we first aggregate the county estimates for three of the 11 risk factor variables to each of the eight key domains listed above. We then compare the aggregated estimates with the BRFSS direct sample estimates. As shown in Table 1 all of the differences are zero or very small, indicating that the small area estimates "add" to the published estimates.

Table 1: Difference Between Aggregated County Estimates and Direct Sample Estimates for Key State and Sub-State Domains

	Current Smoking Prevalence			Current Asthma Prevalence		
Domain	County Aggregation	BRFSS Estimate	Difference	County Aggregation	BRFSS Estimate	Difference
Age:						
18-24	20.7574	20.7574	-0.00000	9.7883	9.7883	-0.00000
25-34	22.1642	22.1642	0.00000	5.4996	5.4996	-0.00000

35-44	24.2395	24.2395	0.00000	5.7456	5.7456	0.00000
45-54	21.6495	21.6495	0.00000	6.9607	6.9607	-0.00000
55-64	18.5290	18.5290	-0.00000	8.3375	8.3375	0.00000
65+	9.6464	9.6464	0.00000	7.6868	7.6868	0.00000
Gender:						
Male	23.9505	23.9505	-0.00001	5.2661	5.2660	0.00007
Female	16.1329	16.1329	0.00001	8.9270	8.9271	-0.00007
Race/Ethnicity:						
White nonHispanic	22.2224	22.2228	-0.00041	8.3429	8.3434	-0.00043
Black nonHispanic	18.6209	18.6213	-0.00035	7.7140	7.7142	-0.00021
Hispanic	17.5094	17.5098	-0.00034	4.9450	4.9450	-0.00002
Asian, American Indian, Other nonHispanic	17.0017	17.0020	-0.00033	7.8241	7.8242	-0.00012
Education:						
Less than HS	23.4926	23.4893	0.00332	6.8501	6.8506	-0.00050
High School graduate	26.4108	26.4072	0.00363	7.5722	7.5721	0.00009
Some College	19.0975	19.0949	0.00264	6.8715	6.8712	0.00027
College graduate	10.4191	10.4176	0.00146	7.1478	7.1478	-0.00003
Marital Status:						
Married	15.2333	15.3211	-0.08784	5.1489	6.8506	-0.00050
Divorced/Separated	32.1284	32.3137	-0.18536	10.9451	7.5721	0.00009
Widowed	14.4539	14.5372	-0.08330	9.7238	6.8712	0.00027
Never married	25.2497	25.3956	-0.14586	8.8001	7.1478	-0.00003
Region:						
1	19.4770	19.4773	-0.00024	6.5387	6.5388	-0.00005
2	16.3861	16.3864	-0.00027	7.5073	7.5074	-0.00010
3	17.8142	17.8144	-0.00017	6.9251	6.9251	-0.00002
4	21.2723	21.2724	-0.00011	6.0186	6.0185	0.00016
5	15.4479	15.4479	-0.00001	9.9822	9.9828	-0.00056
6	17.1135	17.1135	-0.00002	6.6098	6.6101	-0.00037
7	13.6567	13.6567	-0.00000	8.4959	8.4960	-0.00006
8	16.2499	16.2499	0.00001	6.3020	6.3027	-0.00068
9	16.3166	16.3166	-0.00003	2.5042	2.5045	-0.00026
10	26.6284	26.6308	-0.00238	13.5830	13.5827	0.00031
11	18.8144	18.8144	-0.00003	13.0704	13.0702	0.00020
12	16.7505	16.7507	-0.00013	5.6429	5.6424	0.00044
13	24.0131	24.0132	-0.00008	6.6902	6.6900	0.00024
14	22.6273	22.6274	-0.00006	8.1496	8.1491	0.00048
SMART MSA:						
Austin, TX MSA	16.3294	16.3293	0.00011	7.0376	7.0380	-0.00037
Dallas, TX MD	17.8142	17.8144	-0.00017	6.9251	6.9251	-0.00002
El Paso, TX MSA	16.2499	16.2499	0.00001	6.3020	6.3027	-0.00068
Fort Worth, TX MD	21.2723	21.2724	-0.00011	6.0186	6.0185	0.00016
Houston, TX MSA	19.7179	19.7405	-0.02259	6.5128	6.5014	0.01138
Lubbock, TX MSA	23.1242	23.0476	0.07659	14.2194	13.9646	0.25485

McAllen, TX MSA	16.3166	16.3166	-0.00003	2.5042	2.5045	-0.00026
San Antonio, TX MSA	15.2570	15.1641	0.09283	9.7508	9.9103	-0.15953
Rest of State	24.0046	24.0167	-0.01205	7.5562	7.5290	0.02723
SMART County:						
Bexar County TX	15.4479	15.4479	-0.00001	9.9822	9.9828	-0.00056
El Paso County TX	16.2499	16.2499	0.00001	6.3020	6.3027	-0.00068
Fort Bend County TX	16.3861	16.3864	-0.00027	7.5073	7.5074	-0.00010
Harris County TX	19.4770	19.4773	-0.00024	6.5387	6.5388	-0.00005
Hidalgo County TX	16.3166	16.3166	-0.00003	2.5042	2.5045	-0.00026
Lubbock County TX	22.6873	22.7011	-0.01384	14.4114	14.1331	0.27830
Travis County TX	17.2574	17.2928	-0.03539	7.6827	8.0817	-0.39905
Williamson County TX	13.6567	13.6567	-0.00000	8.4959	8.4960	-0.00006
Rest of State	21.3675	21.3655	0.00204	6.9912	6.9686	0.02259

8. Conclusions

Small area estimation techniques often produce estimates that do no "add up" to the "official" direct sample estimates. This may raise validity concerns among the "consumers" of the county estimates. To address this issue we have integrated two techniques – 1) creating a large sample "virtual population" using the 2005-2009 ACS PUMS for each county and assigned predicted probabilities to the adults in that sample, and 2) using Iterative Probability Adjustment (IPA) to constrain the predicted probabilities so that the county estimates add to the domain estimates from the state-based sample survey (i.e., the BRFSS). Although these techniques have been used individually, we are not aware of the combined use of these techniques to develop county estimates. Because we are assigning adjusted predicted probabilities to the individuals in the in the "virtual population" in each county, our approach also yields county estimates for key domains such as age group, gender, race/ethnicity, education and marital status.

References

Battaglia, M.P., M.R. Frankel, and M.W. Link. 2008**.** Improving Standard Poststratification Techniques For Random-Digit-Dialing Telephone Surveys**.** *Survey Research Methods*. Vol 2, No 1.

Battaglia, M.P., D. Izrael, D.C. Hoaglin, and M.R. Frankel. 2009. Practical Considerations in Raking Survey Data. *Survey Practice*, June 2009.

Izrael, D., M.P. Battaglia, and M.R. Frankel. 2009. Extreme Survey Weight Adjustment as a Component of Sample Balancing (a.k.a. Raking), *2009 SAS Global Forum*, http://www.abtassociates.com/Page.cfm?PageID=40858&FamilyID=8600.