Small Area Estimation for the Virginia 2017 BRFSS

Submitted to:

Virginia Department of Health
Office of Family Health Services
109 Governor Street
Richmond, Virginia 23219
Attn: Freda Williams, MPH

Submitted by:

Abt Associates Inc.
6130 Executive Boulevard
Rockville, Maryland 20852

January 30, 2019

VDH VIRGINIA DEPARTMENT OF HEALTH

Contents

1. Background

The Virginia Behavioral Risk Factor Surveillance System Survey (BRFSS) is designed to provide stable and accurate estimates on key health outcomes for the household population of adults 18 years of age and older who reside in the Commonwealth of Virginia. The survey is not designed, however, to provide estimates by county. As such, the survey sample sizes are not sufficiently large to provide stable direct estimates of important health outcomes by county (or, in the case of Virginia, by independent city).

Small area estimation (SAE) may be able to provide more reliable estimates for counties than can be obtained from direct survey estimates alone. SAE methods combine direct estimates of outcomes from the survey with estimates modeled using auxiliary or outside data. This auxiliary data can come either from a larger survey such as the U.S Census Bureau's American Community Survey (ACS) or from administrative data such as vital statistics records. Through statistical models of the mathematical relationships between the outcome of interest and area-level characteristics, we can develop estimates for small areas that "borrow strength" from data about other, similar areas. From these models we can get synthetic estimates that represent the expected value of the outcome by county with similar (modeled) characteristics.

There are several important caveats to keep in mind in presenting and using SAEs:

- Whereas direct survey estimates contain sampling error and other sources of survey error, SAEs additionally contain error from the auxiliary data sources as well as the model error. The quality of SAEs thus depends not only on the quality of the underlying survey data but also on the quality of the auxiliary data and the model used to create the estimates. As such, just as survey estimates are presented with information about the survey methodology and measures of uncertainty such as standard errors or confidence intervals, SAEs should be presented with additional information about the auxiliary data and modeling procedures along with information about the quality of the estimates.

- Related to the above, SAEs include modeled or synthetic estimates. As such, an area's SAE may differ from the true value of the outcome when there are key explanatory variables that impact the outcome that are not included in the model. As a hypothetical example, two areas with similar demographic characteristics and administrative data (and thus similar synthetic estimates) may truly differ on fruit and vegetable consumption due to the presence of a farmer's market in one county but not in the other. Users of SAEs should consider the data included in the model along with local conditions or characteristics not included in the model that may impact the accuracy of a given SAE.

The goal of the SAE for the 2017 Virginia BRFSS was to provide estimates of 36 key health outcome measures for the Commonwealth of Virginia's 133 counties and independent cities.

Exhibit 1: County/City Summary Table

County/City	Unweighted N
Accomack County	90
Albemarle County	121
Alleghany County	34
Amelia County	40
Amherst County	37
Appomattox County	24
Arlington County	189
Augusta County	126
Bath County	10
Bedford County	136
Bland County	15
Botetourt County	34
Brunswick County	21
Buchanan County	35
Buckingham County	39
Campbell County	89
Caroline County	17
Carroll County	37
Charles City County	6
Charlotte County	39
Chesterfield County	408
Clarke County	19
Craig County	6
Culpeper County	56
Cumberland County	29

County/City	Unweighted N
Dickenson County	29
Dinwiddie County	44
Essex County	32
Fairfax County	1,034
Fauquier County	59
Floyd County	42
Fluvanna County	37
Franklin County	48
Frederick County	73
Giles County	39
Gloucester County	85
Goochland County	26
Grayson County	30
Greene County	20
Greensville County	13
Halifax County	62
Hanover County	124
Henrico County	298
Henry County	58
Highland County	7
Isle of Wight County	36
James City County	101
King and Queen County	13
King George County	22
King William County	17
Lancaster County	34

County/City	Unweighted N
Lee County	25
Loudoun County	282
Louisa County	41
Lunenburg County	28
Madison County	20
Mathews County	24
Mecklenburg County	82
Middlesex County	26
Montgomery County	113
Nelson County	22
New Kent County	23
Northampton County	40
Northumberland Count	35
Nottoway County	33
Orange County	39
Page County	25
Patrick County	23
Pittsylvania County	143
Powhatan County	28
Prince Edward County	48
Prince George County	73
Prince William County	373
Pulaski County	34
Rappahannock County	16
Richmond County	51
Roanoke County	146

County/City	Unweighted N
Rockbridge County	41
Rockingham County	130
Russell County	40
Scott County	42
Shenandoah County	51
Smyth County	42
Southampton County	20
Spotsylvania County	106
Stafford County	133
Surry County	14
Sussex County	40
Tazewell County	55
Warren County	43
Washington County	88
Westmoreland County	42
Wise County	64
Wythe County	36
York County	70
Alexandria city	132
Bristol city	5
Buena Vista city	9
Charlottesville city	39
Chesapeake city	264
Colonial Heights city	34
Covington city	3
Danville city	32

County/City	Unweighted N
Emporia city	2
Fairfax city	82
Falls Church city	26
Franklin city	24
Fredericksburg city	30
Galax city	4
Hampton city	190
Harrisonburg city	14
Hopewell city	33
Lexington city	2
Lynchburg city	91
Manassas city	22
Manassas Park city	6
Martinsville city	9
Newport News city	141
Norfolk city	319
Norton city	7
Petersburg city	65
Poquoson city	6
Portsmouth city	139
Radford city	14
Richmond city	224
Roanoke city	110
Salem city	19
Staunton city	9
Suffolk city	95

County/City	Unweighted N
Virginia Beach city	499
Waynesboro city	11
Williamsburg city	14
Winchester city	12
Total	9,627

Please note that the unweighted (raw) count of completed interviews by county represent the total of the 9,627 2017 Virginia BRFSS cases. They do not represent the number of unweighted cases used in the small area estimation (SAE) for each of the thirty-six health indicator analyses. Both unweighted and effective sample sizes are indicator specific – because of the item missing data for all outcomes (due to either nonresponse when the respondent provided a "Don't know" answer or refused to answer, or due to the case coming from a different state, and not filling in the VA-specific modules) and because of the age or sex or other subsetting for some indicators (e.g. shingles vaccination age 50+). Therefore effective sample sizes by county are not presented in this report, because data files of different sizes (number of cases) were used through the SAE process, utilizing all available data.

2. Small Area Estimation Procedures

Small area estimation (SAE) is an area of active growth and research in survey statistics that emerged since the 1990s. This followed greater demands for detailed estimates from data users accompanied with greater availability of computing power. SAE addresses the problem of obtaining reasonable estimates for domains where small sample sizes do not allow direct estimation using survey data only (including domains with zero sample observations), e.g., at the levels of a county or a metro area in national and state samples. An encompassing reference on SAE is Rao (2003)[1].

The modern approach to SAE involves the use of statistical models to predict the outcome of interest, such as current smoking. Direct estimates from the survey data are combined with synthetic estimators from statistical models to create a composite SAE.

Each of the thirty-six health indicator output files contain results by county. These output include the one-sided lower and upper bounds at the 95% confidence interval and the point (best) estimate as well as bell curve charts with the two-tailed 90% confidence interval represented by gray markers.

[1] Rao, J. N. (2003). *Small Area Estimation.* Hoboken, NJ: Wiley.

3. CDC BRFSS SAE Methodology

3.1 BRFSS SAE Estimation

The Centers for Disease Control and Prevention (CDC) has developed a SAE system for a portion of the BRFSS data referred to as SMART (Selected Metropolitan/Micropolitan Area Risk Trends) for counties that have sample sizes of at least 500. Pierannunzi et. al. 2016[2] outlines the CDC method to model health outcomes at the county level using this procedure (referred to as SMART-SAE), which is described below. *Our comments regarding the implementation of these procedures and potential improvements, are provided as sub-items and are italicized.*

1. BRFSS data are re-raked to state-level demographic variables (age by gender, race/ethnicity, education, marital status, (housing) tenure, gender by race-ethnicity, age by race-ethnicity, region by age, gender, and race-ethnicity), as well as county-level targets (county by sex, age, and race) using Nielsen Claritas data. Weights are then rescaled to the nominal sample size for the county.

    a. *The Nielsen Claritas data set was apparently chosen for historic reasons, as it was used before by the CDC for similar purposes. The American Community Survey (ACS) based margins appear to be more reliable.*

    b. *There is evidence (Pfeffermann et. al. 1998) that scaling by the effective sample size, rather than by the nominal sample size, works better in reducing small sample biases of variance parameters in mixed models.*

2. BRFSS outcomes are imputed using a hot-deck procedure.

    a. *The hot-deck procedure is somewhat restrictive in that it has a certain low-dimensional structure in mind. Namely, that the missingness is conditionally independent of the outcomes within the imputation cells. We believe that the regression model for the outcome that will be proposed at a later stage does a similar or a better job incorporating the missing data in the outcomes.*

    b. *In addition, imputation introduces a source of variation in the data that needs to be accounted for in the standard errors of the SAEs down the line. To incorporate the imputation variation correctly, multiple imputation procedures and Rubin (1978) rules should be used. It is unclear whether that was done in SMART-SAE by the CDC.*

    c. *Given this, for our work with the 2014 Virginia BRFSS SAE, we chose to forego the imputation step, as the demographic predictors were used in the model for the outcome.*

[2] Pierannunzi C, Xu F, Wallace RC, Garvin W, Greenlund KJ, Bartoli W, et al. A Methodological Approach to Small Area Estimation for the Behavioral Risk Factor Surveillance System. *Prevention of Chronic Diseases* 2016;13:150480. DOI: http://dx.doi.org/10.5888/pcd13.150480.

3. ACS public use microdata series (PUMS) data are obtained.

    a. *Pierannunzi et. al. (2016) describes this step as creating a single-year data set from multiple years of ACS data. We believe it is better to use the data set created by the Census Bureau specifically for the purposes of providing sufficient sample sizes at low levels of geography, namely the 5-year data set. For 2014 Virginia BRFSS SAE we used the ACS 5-year data set for 2010–2014.*

4. ACS data are approximately subset to county levels. The finest level of geography provided in ACS PUMS data is that of the public use microdata area (PUMA), a contiguous geographic area with a total population of about 100,000. Large counties can be split into several PUMAs, and smaller counties are aggregated into a single PUMA, so there is no 1:1 relation. Population fraction weights (% of the PUMA population found in a county) are used to distribute the total population of a PUMA to its component counties if needed. SMART-SAE used the data from Missouri Census Data Center (http://mcdc.missouri.edu/websas/geocorr14.html) to obtain these fractions.

    a. *Alternatively, accurate county variables are available in the protected ACS data available to researchers through research data centers.*

5. ACS data are re-raked to Nielsen Claritas data.

    a. *The utility of this step is unclear to us. ACS is a better quality data set than any commercial data set. The only reasonable justification is to align the totals to the same ones used in BRFSS raking. Since we do not have the Nielsen Claritas data set we cannot perform the SMART-SAE raking steps 1 and 5. However, we can rely on using the low level ACS data (as detailed in 3.a. above).*

6. ACS and BRFSS data are stacked together to prepare for modeling.

7. A logistic random effects model with county as a random effect, and race, gender, and age as the main effects, is fit to the data.

    a. *The model is known as the unit-level model (Rao and Molina 2015, Sec. 4.2), where modeling happens at the level of each individual. The use of the model for SAE assumes that the values of the explanatory variables are known for all units in the population – (see 5a in Section 3.2 Challenges and Solutions below). No data source exists for the U.S. population that can act as a complete population register with the required race, gender, and age information.*

    b. *An alternative is a model in which the response variable is the direct estimate for an area, and all explanatory variables are at the area level. This approach is known as an area-level model in SAE literature (Rao and Molina 2015, Sec. 4.3). Area-level models do not require the knowledge of the values of explanatory variables for all units in the population, and thus are easier to use in the U.S. This model is what we used for the 2014 Virginia BRFSS SAE*

    c. *The SMART-SAE model used by the CDC can arguably be improved by adding county-level contextual variables.*

8. Predicted probabilities are obtained for the age-gender-race cells within a county.

   a. *This was not spelled out as a separate step by Pierannunzi et. al. 2016, arguably because its implementation is through the same SAS PROC GLIMMIX procedure call utilized in step 7 listed above. However, there are a variety of ways to create predictions. In Stata, which we use for modeling, the model fitting steps and generating prediction steps are separated in syntax.*

   b. *Pierannunzi et. al. (2016) use the language of "best linear unbiased prediction". However, these do not exist for the nonlinear models like logistic regression used here. The appropriate concept for the binary data are empirical best predictions (Jiang and Lahiri 20013).*

   c. *Aggregation of predictions to the county level is not described in the Pierannunzi et. al. paper in sufficient detail. Several implementations are possible.*

      i. *The predicted probabilities (incorporating the random effect of the county) could be obtained for the units in the ACS PUMS using the regression coefficients obtained on BRFSS data. These predicted probabilities can then be added up with their appropriate ACS weights (split between counties as needed, as explained above in item 4c) to form county-level SAE.*

      ii. *The predicted probabilities for all age-gender-race groups within a county can be obtained (incorporating the random effect of the county), and then these predicted probabilities can be added up using the estimated proportions of these population groups using aggregate ACS data such as the FactFinder tables. This approach may produce more precise estimates, as only a fraction of ACS data is released as the public use data, while the FactFinder calculations are based on the complete ACS data set. However, computation of the standard errors for these predictions through aggregation of cell-level standard errors is difficult.*

   d. *In a strict sense of SAE methodology, SMART-SAE appears to be synthetic estimates, in that they only use the estimated model. Better estimates can be obtained by combining the synthetic estimates with the direct survey estimates into composite estimates. The latter can increase the effective sample size used in estimation, and better protect against possible model violations, such as important covariates not included in the model.*

Previous work conducted by the CDC[4] compared a number of possible approaches to SAE, including:

---

[3] Jiang, J., and Lahiri, P. (2001) Empirical Best Prediction for Small Area Inference with Binary Data. *Annals of the Institute of Statistical Mathematics*, **53**, 217–243.

[4] Gotway-Crawford, C., D. Ford and C. Pierannunzi (2014) "Comparison of Small Area Estimation Methods for use by the BRFSS". Presentation at the annual AAPOR conference.

1. Unweighted logistic random effects model;
2. Weighted logistic random effects model;
3. Multi-level model;
4. Aggregation over time (7-year window);
5. Empirical best linear unbiased prediction, weighted and benchmarked;
6. Constrained and benchmarked model (essentially, a highly detailed raking procedure); and
7. Bayesian SAE.

The weighted logistic regression performed best, and was chosen as the backbone of the procedure outlined above.

3.2 Challenges and Solutions

From the total survey error perspective, aside from measurement errors, the errors in representation and modeling in the proposed SAE estimation plan are detailed below.

Challenges:

1. Coverage error for the non-telephone population. Extrapolation to the non-telephone population is implicitly performed by using the ACS areal frame data. However, the limited use of the person-level demographic variables effectively assumes that phone coverage is sufficiently well explained by these variables only.
2. BRFSS nonresponse error (partially compensated by weighting).
3. For Steps 1 and 5 of the CDC's SMART-SAE procedure as detailed in *Section 3.1*, the county-level population estimates used as targets for raking are likely to have nonzero levels of uncertainty. The resulting BRFSS estimates thus partially inherit sampling error from the ACS which can be nontrivial for small counties where ACS sample sizes may be in the low hundreds. Sampling errors in targets can be quantified for the ACS data; however, properly accounting for them downstream becomes a complicated exercise (Dever and Valliant 2010[5]).
4. Overall, the utility of step 5 of the CDC's SMART-SAE procedure as detailed in *Section 3.1* is unclear, as the direct estimates for which the re-calibrated weights may arguably provide improved inference with lesser biases are never obtained.

Solution:

[5] Dever, J. A. and R. Valliant (2010). A comparison of variance estimators for poststratification to estimated control totals. Survey Methodology 36 (1), 45-56.

5. If county-level variables are to be used in modeling the outcomes, a very broad range of possibilities opens up for the data sources. We identified the following data sets that can be used as sources for the county-level explanatory variables:

    a. **ACS Tables** (the possibilities are really endless, as FactFinder can produce several thousand tables at the county level)**.** It can be reasonably expected that some of the demographic variables may have predictive power. Also, socioeconomic status variables (poverty rates and the use of government assistance programs) and limited health-related variables (health insurance status) may turn out to be helpful.

    b. **Census Planning Database.** While this database mostly reuses the ACS data, it also contains additional variables such as the Census mail return rates. It is somewhat unclear whether these variables can be useful in modeling health outcomes, but they may serve as proxies for the willingness of the population to cooperate with the government (which in turn may be associated with the health outcomes in programs that have explicit health policies, such as vaccinations or smoking).

    c. **National Center for Education Statistics Data** for school districts that, in case of Virginia, align with counties. Variables such as percent of students eligible for free or reduced price lunch are proxies for the county-level socio-economic status, while teacher-to-student ratio, availability of the local government funding.

    d. **U.S. Environmental Protection Agency Environmental Quality Index (EQI).** This includes the measured concentrations of certain known pollutants, aggregated to the air quality, water quality and land quality indices, sociodemographic domain and built environment domain quality indices, and the overall environmental equality index.

    e. **County-level health estimates from the 2000 SAE project of the National Cancer Institute** (http://sae.cancer.gov; Raghunathan et. al. 2007). Four estimates are available: current smoking, ever smoked, Pap smear, and mammography.

    f. **County Health Rankings,** a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute (http://www.countyhealthrankings.org/app/virginia/2018/overview).

    Additionally, the Virginia Department of Health provided the following data sets which were used in the modeling:

    a. Age-Adjusted Malignant Cancer Incidence Rates and Counts for Selected Cancer Sites by sex, VA 2013.

    b. Mortality rates by source and age, VA 2010–2014.

    c. Hospitalization rates, by source, VA 2005–2013.

    Finally, we used the SAE estimates from the 2014 round.

Challenges:

6. Given the multitude of the possible area-level variables available, the issue of selecting the best ones will arise. Unfortunately, the current methodological literature on SAE is scarce on the topic. Pfeffermann (2013[6]) proposes to use conditional Akaike information criterion to find the best fitting model given the eventual focus on prediction for the existing set of areas (rather than a generalization to a fictional universe of all possible areas implied by the Gaussian distribution of the random effects). Application of this proposal would require fitting a broad number of random effects models, which does not appear to be computationally feasible (especially given the interest of the current project in 36 health indicators, each of which will likely require its own model).

    a. A faster search can be conducted with more aggregated data that would not require mixed modeling.

        i. Fay-Herriot (1978[7]) area-level regression model is formulated, where the dependent variable is given by the direct estimates of the outcome at the area level (transformed to stabilize the error variance), and explanatory variables are observed at the area level, as well.

        ii. The space of possible regressors is narrowed down using an elimination procedure, in which a candidate variable is used as a regressor along with a fixed set of demographic variables (e.g., age, gender and race, as in the CDC procedure), and is omitted from the further consideration if it failed to achieve significance at a conservative level of 0.20.

        iii. An exhaustive search in the space of the possible models of a given complexity (e.g., the base demographics + 3 additional regressors) is conducted, and the best model selected using the traditional AIC. This space would usually include 1,000 to 4,000 regression models to run, and the time scale for model search is single digit hours per outcome.

    b. To account for the potential model selection error, a limited number of the best fitting models can be retained for each outcome. Variation between the estimates based on the different models may serve as a measure of the error due to the model selection step.

7. The sampling error in the regression coefficients (step 4 in the SMART-SAE procedure) needs to be accounted for.

8. The sampling error in the ACS aggregation (step 5 in the SMART-SAE procedure) needs to be accounted for. We utilized a faster Taylor series linearization using the strata and cluster variables provided with the IPUMS data.

---

[6] Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* 28 (1), 40-68.

[7] Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74 (366), 269-277.

9. Self-contained analytical expressions for the Mean Squared Error (MSE) of the SAEs (Rao and Molina 2015) account for the sampling and the model fit error when a fixed model is fit to the (area-level or unit-level) outcomes. In other words, these expressions would be appropriate if a fixed model was fit to the BRFSS data, and predictions were obtained directly from that model. As far as we can see, additional steps will likely be required:

    a. Extrapolation to the ACS data (step 7 of the SMART-SAE procedure);

    b. Model selection (issue #6 in the current list).

4. Virginia BRFSS Small Area Estimation

4.1 Health Outcomes of Interest

The Virginia Department of Health identified the 36 outcomes of interest, listed here:

Exhibit 2: Indicator Summary Table

Indicator #	Indicator Name
1	Binge Drinking- 2014, 2017
2	Colorectal Cancer Screening- 2014, 2017
3	Overweight or Obese – 2014, 2017
4	No Physical Activity in the Past Month- 2014, 2017
5	Poor Mental Health- 2014, 2017
6	Regular Healthcare Provider- 2014, 2017
7	Diabetes- 2014, 2017
8	Hypertension- 2017
9	Arthritis- 2014, 2017
10	Current Smoker- 2014, 2017
11	Ever Had Asthma- 2014, 2017
12	Dental Visit in the Past Year- 2014, 2017
13	Poor Health: Physical or Mental Health – 2014, 2017
14	Could not afford prescribed medicines
15	Depressive Disorder- 2014, 2017
16	Ever Had Stroke- 2014, 2017
17	Ever Had Heart Attack- 2014, 2017
18	High Cholesterol- 2017
19	Diabetes Test Past Year- 2014, 2017
20	Current E-Cigarette Use- 2017
21	Ever Had COPD- 2014, 2017
22	Has dental insurance
23	Adverse Childhood experiences, percentage who reported three or more ACE experiences - 2017
24	Heart disease, 2014, 2017
25	Pre-diabetes, 2014, 2017
26	No Health coverage, 2014, 2017
27	No doctor due to cost, 2014, 2017
28	Any tobacco use – 2014, 2017
29	Routine Checkup- 2017
30	TDAP 2014, 2017
31	Shingles Vaccine- 2014, 2017
32	HIV Test Ever- 2014, 2017

Indicator #	Indicator Name
33	Fruit and Vegetable Consumption – FV 5+ times a day, 2017
34	Pneumonia Shot Ever- 2014, 2017
35	Tetanus vaccination in the past 10 years
36	Flu vaccination past year

*During the data analysis and SAE preparation, two of the original health indicators (17 'Tested for Diabetes in the Past 3 Years' and 32 'Driven when you've had too much to drink (During the past 30 days)') were removed from the list and replaced with indicators 37 and 38.

4.2 Data

The data set used for the SAE was the 2017 Virginia BRFSS, of which there were 9,627 cases. In the demographic section of the survey respondents are asked what county they live in. For the 2017 Virginia BRFSS, 8,699 (90.3% unweighted) respondents provided a response of a county within Virginia (`ctycode2` variable); for other cases, an open text entry was encoded by CDC (`cpcounty` variable). According to the BRFSS SMART procedure, a respondent's missing county was imputed by CDC according to the following process:

- For landline numbers, the frame county (based on the most prevalent county in a given 100-block) was used.
- For cell phone numbers, the county was coded based on:
  - **a.** An open-ended self-reported location;
  - **b.** Self-reported ZIP code; or
  - **c.** For the records lacking the above, the largest county population by age and race/ethnicity was used.

The accuracy of the latter imputation step is likely to be low, and it may bias estimates for the counties affected (mostly, the counties into which the imputation was made). Three cases had any and all of the geographic information missing, and did not contribute to direct estinates (but were used in model fitting).

4.3 Virginia BRFSS SAE Procedures

Retaining the main steps and methods of the CDC SMART-SAE procedure outlined above in *Section 3. CDC BRFSS SAE Methodology*, we used the following SAE procedures for the 2017 Virginia BRFSS.

1. Produced direct estimates of the health outcomes at the county level.
2. Performed selection of a concise predictive Fay-Herriot type area-level model.
3. Fitted a weighted logistic regression model to the BRFSS data with outcomes from the survey data, age, race, and gender from the survey data, and the county-level predictors selected in step 2. (The procedure was streamlined due to 0/1 nature of the recoded outcomes.)

4. A mixed logistic model was used that utilized both area level and unit level predictors. (For most outcomes, the variances of random effects were estimated to be zero, and the model reduced to the generalized linear model (GLM) weighted logistic regression model with these predictors.)

5. Obtained the (unit-level) predicted values from the final model for the matched ACS PUMS data.

6. Obtained the (area-level) point estimates using the ACS PUMS weights.

7. Obtained the (area-level) variance component due to the regression model parameters being estimated using the delta method and the variance-covariance matrix of regression parameter estimates.

8. Obtained the (area-level) variance component due to the posterior empirical Bayes distribution of area effects.

9. Obtained the (area-level) ACS sampling standard errors using the Successive Difference Replicates (SDR) variance estimation procedure recommended for the ACS.

10. Assuming that the model selection error is independent of both the ACS sampling error and the SAE MSE, calculate the between-model error as a variance within a small subset of the best-fitting candidate models from step 2.

11. Combine the sources of error assuming independence, and use the square root of thus estimated variance to report as the standard error of the SAEs obtained in step 6.

This procedure was repeated for all 36 indicators of interest for all 133 counties and independent cities.

5. Reporting Small Area Estimation Results

The small area estimation output for each of the 36 outcomes is provided in four subsections: 1) definition of the variable and coding; 2) small area regression models; 3) estimates and confidence intervals; and 4) graphical representation of the results. Each of these is detailed below.

The project consists of 32 human-created Stata do-files, 36 automatically created do-files for model search, 36 automatically created do-files for mixed model estimation, 205 intermediate data files and about 970 graphic files.

5.1 Definition of the Variable and Coding

The first subsection of the results gives the definition of the variable and how it is coded into a binary 0/1 variable. Note that some outcomes are negatively worded. For example, the poor health days variable has a value of 1 for those who report nonzero number of days when poor health prevented the respondent from daily activities. Another example is for the health insurance coverage variable, which has a value of 1 for those who do not have insurance.

5.2 Small Area Regression Models

The second subsection provides the selected SAE models that resulted from the model search. All models include demographic covariates following Pierannunzi et al (2016). For each outcome, three models are reported.

1. The first model is a weighted mixed logistic regression model with demographic variables only, and area effects included as random effects (i.e., no area level covariates are being used). This is the model that would have resulted from the SMART-SAE approach of Pierannunzi et al (2016).
2. The second model is a logistic regression model without area effects.
3. The third model is the mixed model with both demographic unit-level variables and aggregated area-level covariates in the best fitting Fay-Herriot model. This third model "Mixed with area covariates" is the final model used for the SAE.

In all but five of the final models with both unit-level demographic variables and area-level aggregated variables, the area-level random effect variance was estimated to be zero. The results are summarized in Exhibit 1. Rao and Molina (2015, Sec. 9.3) discuss this as an undesirable artefact of mixed models, as in this case, compositing of the model-based and direct-estimates is limited, and the estimates for the areas with larger effective sample sizes do not benefit from greater precision of direct estimates. With zero area variances, model-based estimates should be interpreted as interpolation/extrapolation of the area-level effects using the selected covariates. If such model approximation is subject to specification error, interpolation and especially extrapolation errors may result. The issue is partially controlled for by incorporating the variance component over the range of several best performing models. The number of such models is reported under the model table, and ranges from 1 to 15 (the number was capped at 15 to reduce the computational time). To the extent that a given area may be outlying on its area-level covariates, and thus at risk of extrapolation biases, it will also be likely to have a greater variability of estimates obtained from the

competing models. As the latter component is incorporated into the SAE standard errors, we believe that risks of extrapolation errors are mitigated. Outcomes that had non-trivial variances of random effects were:

19. Diabetes test past year

20. Current e-cigarette use

26. Health insurance coverage

27. Doctor inaccessible due to cost

32. Ever tested for HIV

Note that in the 2014 VA BRFSS SAE project, all models had zero final random effect variances.

Estimated Intraclass Correlation Coefficient (ICC) for counties varies from 0 to 8.3% (insurance coverage) in the models with demographic variables only. Here, ICC is estimated as

$$\text{ICC} = \frac{\sigma_u^2}{\sigma_u^2 + \pi^2/3}$$

where $\pi^2/3$ is the variance of logistic distribution. In other words, the reported estimated standard deviation of the area variance should be compared to the standard deviation of the logistic distribution, $\pi/\sqrt{3} = 1.814$, to gauge the relative importance of the area effects, on top of the demographic effects, unaccounted for in that model. Note that some important person-level covariates that have not been used either in Pierannunzi et al (2016) or here were socio-economic status (which can be proxied by income and education) or interactions of demographic variables (e.g., age by gender).

Exhibit 3: Estimated Standard Deviations of County-Level Random Effects.

	Demographic v variables only			**Demographic + area variables**		
Outcome	**st. dev. of area effects**	**(std. error.)**	**ICC**	**st. dev. of area effects**	**(std. error.)**	**ICC**
outc1_binge_drink	0.2985	(0.0784)	0.0264	0.0000	(0.0000)	0.0000
outc2_colon_test	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc3_oweight_obese	0.1384	(0.0383)	0.0058	0.0000	(0.0000)	0.0000
outc4_phys_activity	0.2929	(0.0672)	0.0254	0.0000	(0.0000)	0.0000
outc5_ng_ment_hlth	0.1987	(0.0658)	0.0119	0.0000	(0.0000)	0.0000
outc6_doctor_person	0.1632	(0.1201)	0.0080	0.0000	(0.0000)	0.0000

	Demographic v variables only			Demographic + area variables		
Outcome	**st. dev. of area effects**	**(std. error.)**	**ICC**	**st. dev. of area effects**	**(std. error.)**	**ICC**
outc7_diabetes_diag	0.3103	(0.0673)	0.0284	0.0000	(0.0000)	0.0000
outc8_highbp	0.2100	(0.0388)	0.0132	0.0000	(0.0000)	0.0000
outc9_arthritis	0.2400	(0.0612)	0.0172	0.0000	(0.0000)	0.0000
outc10_curr_smoker	0.4963	(0.0850)	0.0697	0.0000	(0.0000)	0.0000
outc11_asthma	0.2048	(0.0599)	0.0126	0.0000	(0.0000)	0.0000
outc12_dentist	0.4776	(0.0570)	0.0648	0.0000	(0.0000)	0.0000
outc13_poorhlthdays	0.0987	(0.0586)	0.0029	0.0000	(0.0000)	0.0000
outc14_medscost	0.3630	(0.0829)	0.0385	0.0000	(0.0000)	0.0000
outc15_depression	0.2558	(0.0871)	0.0195	0.0000	(0.0000)	0.0000
outc16_stroke	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc17_heart_attack	0.1924	(0.1111)	0.0111	0.0000	(0.0000)	0.0000
outc18_highchol	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc19_diabetes_test	0.1786	(0.0695)	0.0096	**0.0433**	**(0.2023)**	**0.0006**
outc20_ecig	0.5128	(0.1456)	0.0740	**0.4315**	**(0.2600)**	**0.0536**
outc21_copd	0.3648	(0.0861)	0.0389	0.0000	(0.0000)	0.0000
outc22_dentins	0.5012	(0.0466)	0.0709	0.0000	(0.0000)	0.0000
outc23_ace	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc24_heart_disease	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc25_prediab_diag	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc26_health_cover	0.5472	(0.0751)	0.0834	**0.2365**	**(0.1260)**	**0.0167**
outc27_doctor_cost	0.3237	(0.0542)	0.0309	**0.1361**	**(0.0666)**	**0.0056**
outc28_tobacco_user	0.5007	(0.0622)	0.0708	0.0000	(0.0000)	0.0000
outc29_rt_checkup	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc30_tdap_vacc	0.2303	(0.0975)	0.0159	0.0000	(0.0000)	0.0000

	Demographic v variables only			Demographic + area variables		
Outcome	**st. dev. of area effects**	**(std. error.)**	**ICC**	**st. dev. of area effects**	**(std. error.)**	**ICC**
outc31_shingles_vacc	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc32_hiv_test	0.3069	(0.0515)	0.0278	**0.1551**	**(0.0588)**	**0.0073**
outc33_fvday	0.1324	(0.0527)	0.0053	0.0000	(0.0000)	0.0000
outc34_pneum_vacc	0.0000	(0.0000)	0.0000	0.0000	(0.0000)	0.0000
outc35_tetanus_vacc	0.0987	(0.0843)	0.0029	0.0000	(0.0000)	0.0000
outc36_flu_vacc	0.1439	(0.0656)	0.0063	0.0000	(0.0000)	0.0000

Note: *, p < 0.05; **, p < 0.01; ***, p < 0.001

5.3 Estimates and Confidence Intervals

The third subsection provides the numeric summaries for each county and independent city, grouped by health districts on each page. These summaries include the direct estimate (i.e., the estimate obtained from the survey data alone using BRFSS weights), its confidence intervals (CI), and the composite estimate with its confidence interval. The confidence intervals are Wilson confidence intervals. Dean and Pagano (2015[8]) found those to be among the most accurately and robustly performing ones in their comparison of seven different methods for proportion confidence intervals in complex surveys. For the proportion estimate $\hat{p}$ and effective sample size n^*, the Wilson confidence is found by solving for p the coverage equation

$$(\hat{p} - p)^2 \leq z^2_{1-\frac{\alpha}{2}} \frac{p(1-p)}{n^*}$$

which produces the confidence interval of the form

$$\frac{\hat{p} + \frac{z^2_{1-\frac{\alpha}{2}}}{2n^*} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n^*} + \frac{z^2_{1-\frac{\alpha}{2}}}{n^*}}}{1 + \frac{\left(z^2_{1-\frac{\alpha}{2}}\right)}{n^*}}$$

[8] Dean, N. and M. Pagano (2015). Evaluating Confidence Interval Methods for Binomial Proportions in Clustered Surveys. J Surv Stat Methodology, 3 (4): 484-503. doi:10.1093/jssam/smv024

Wilson confidence intervals are always contained between 0 and 1 (unlike Wald confidence intervals $\hat{p} \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n^*}}$), and are asymmetric near zero or one. For both the direct estimate and the composite estimate, the effective sample size is calculated by reversing the variance of an i.i.d. statistic formulae, namely:

$$n^* = \frac{\hat{p}(1-\hat{p})}{(s.e.[\hat{p}])^2}$$

where the standard error is either the design-corrected standard error for the direct estimate, or the composite standard error for the composite estimate.

5.4 Graphical Representation of the Results

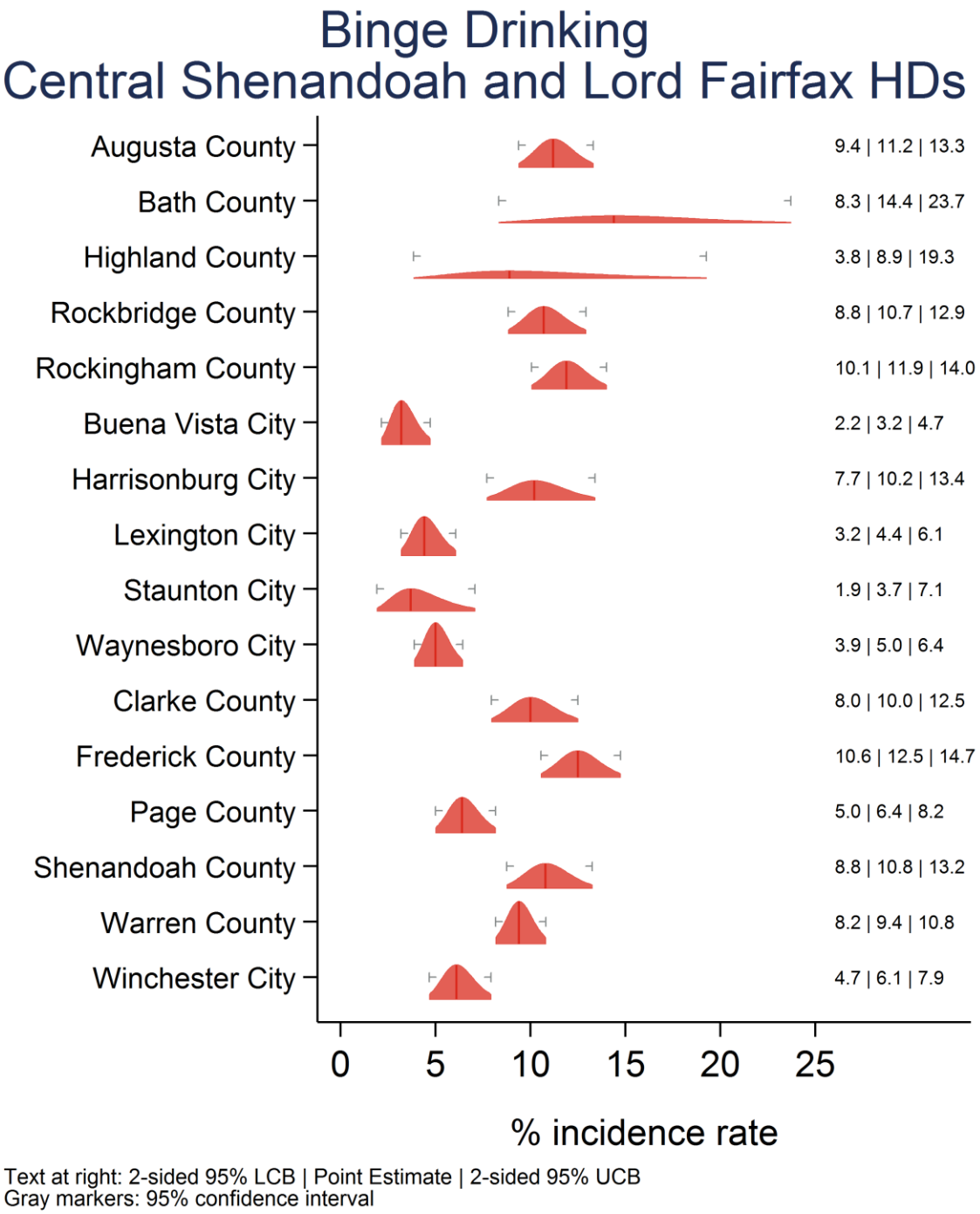
In the fourth subsection for each outcome, SAE results are visualized using inchworm plots (Rhoda, 2016; see https://github.com/BiostatGlobalConsulting/inchworm-plots-stata). The plots are designed to visually convey some of the properties of the confidence intervals for proportions that are not necessarily obvious in the tabular representation of the estimates, their standard errors, and confidence intervals. First, confidence intervals are based on a density, with the highest, most likely values near the estimate, and the least likely values further away from it. Second, asymmetry of confidence intervals for estimates near zero or one and low values of $\hat{p}n^*, (1-\hat{p})n^*$ is, likewise, not immediately obvious from the numeric summaries. Inchworm plots make these features more prominent.

On each plot, the orange curve and the shaded area beneath it represent an approximation to the posterior distribution of the estimate for the given area-and-outcome-specific value of the effective sample size n^*. The grey bookend markers represent Wilson confidence intervals. Both the estimates and the CIs are additionally reported in the numeric output panel on the right[9].

Consider the results for the first outcome, binge drinking, in Central Shenandoah and Lord Fairfax health districts.

Exhibit 4: Binge Drinking

[9] There are small discrepancies in the width and endpoints of the confidence intervals between those reported on the plots and those reported in the tables. We have been able to identify the differences as having to do with the different definitions of the effective sample sizes used in our code and in inchworm plots code.



Most results are found to be in the 5% to 15% range. Several results are of note. First, one can note that low estimates are associated with asymmetric confidence intervals, such as those for Buena Vista City, Lexington City, Page County, and especially Staunton City. Most other confidence intervals, however, are (approximately) symmetric, as sufficiently large (effective) sample sizes are available for these other counties and cities. The second

important observation is that the lack of information is reflected in very wide confidence intervals for Bath County and Highland County. Also, since each curve represents an approximation to the distribution of the estimate with an area of 1, the wider confidence intervals for those two areas are associated with lower heights of the curves.

5.5 Additional Diagnostics

Some additional diagnostics regarding the performance of SAE models in comparison to direct estimates are collected in an additional diagnostic Excel file. Along with the county name, County FIPS, Direct estimate, Direct 95% CI, SAE composite, and SAE 95% CI (same as reported in the Word file), the following information is included:

- **Direct Point vs. SAE CI**
  Results: "Check" if the direct estimate is inside the SAE CI; "Fail" if the direct estimate is outside of the SAE CI.
  Rationale: It is reasonable to expect that the direct estimates and composite SAE estimates will be close. However, since the SAE and direct estimates are not independent, it is difficult to say whether this event should happen. For most outcomes, the direct estimates are outside the SAE CI about half of the time.

- **SAE Point vs. Direct CI**
  Results: "Check" if the SAE estimate is inside the direct CI; "Fail" if the SAE estimate is outside of the direct CI.
  Rationale: It is reasonable to expect that the SAE estimate will be within the direct interval. We should expect this to happen quite often, as direct CIs are quite wide, and this diagnostic check is a very low bar to pass. If the model is successful in producing highly accurate estimates with low bias and low variance, then direct estimates, being unbiased, should contain the SAE estimates approximately 95% of the time.

- **SAE CI vs. Direct CI**
  Results: "Fully within" if the SAE CI is fully contained within the direct CI; "Some overlap" if SAE and direct CIs overlap, but each has portions not covered by the other; and "No overlap" if SAE and direct CIs do not overlap at all.
  Rationale: This is probably the most reliable check on the relation between the SAE and the direct estimates. When the SAE CI is fully within the direct CI, we can say that SAE helped zoom in on the range where the true value is likely to be. Lack of overlap should be particularly troubling.

- **SAE S.E. Less Than Direct S.E.**
  Results: "Check" if SAE standard error (s.e.) is less than or equal to that of the direct standard error; "Fail" if SAE standard error is greater that the (nonzero) direct standard error; "Zero direct s.e." if the direct standard error is zero (so the SAE is bound to be greater than the direct standard error).
  Rationale: We should reasonably expect the standard errors to go down in SAEs as we incorporate the regression model into the estimates.

- Percent variance due to ACS sampling error.

- Percent variance due to the BRFSS model coefficients sampling error (this is the dominant component for most areas and outcomes).
- Percent variance due to random effect variance (unmodeled area effects)
- Percent variance due to model uncertainty.

A summary of these results is also provided in the fifth section of the outcome-specific reports.