Methods for Calculating State and National Prevalence Estimates: An Application of Estimates of Sexual Orientation and Gender Identity

Ronaldo Iachan[1], Yangyang Deng[1], Kristie Healey[1]
[1]ICF, 530 Gaither Road, Suite 500, Rockville, MD 20850

Abstract
This research focuses on the methods for modeling estimates at the state level when data are available from a subset of states. We used the Sexual Orientation and Gender Identity (SOGI) optional module questions from the Behavioral Risk Factor Surveillance System (BRFSS) for 2014 to 2016 to develop models and provide estimates for all states. Models are validated against direct estimates where available. SOGI questions represent the most vigorous test of such a model in that limited proportion of the sample who identify as transgender, bisexual and/or gay/lesbian. The process presented also provides a mechanism for imputation of responses where non-substantive answers are given (i.e. “do not know” or refusal to answer). The methodology is adaptable to other BRFSS optional models used by subsets of the states annually.

Key Words: BRFSS, Sexual Orientation, Gender Identity, Statistical Modeling, Imputation

1. Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about US residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and participating territories. BRFSS completes more than 400,000 adult interviews each year nationally. The BRFSS sample is drawn by individual states, rather than being drawn as a single, national sample. The BRFSS is comprised of a core set of questions, which are adopted in standard form for all states. States may also select from standardized optional module on a number of health topics (Centers for Disease Control and Prevention 2017).

Data from the BRFSS have been used to model for Small Area Estimates (SAEs) in many studies (Guo et al. 2013, X. Zhang et al. 2014, Zhang et al. 2011). A recent publication reported a method of deriving county-level estimates from BRFSS state-level data (Pierannunzi et al. 2016). In most instances the large sample included in the BRFSS supports methods for creating sub-state prevalence estimates using state data, or aggregating the BRFSS to a nationwide sample (Khalil and Crawford 2015) and then modeling sub-state areas (Song 2016, Li W 2009).

The literature provides instances where researchers have modeled from direct data from one geographic location that has sufficient direct observations to geographic areas where data are missing (National Cancer Institute 2017). Similar methods may be used to calculate estimates at the state level for questions within optional modules that have been asked only in a few states.

Since 2014, the BRFSS has used an optional model on sexual orientation and gender identity that states may choose to append to the core portion of the survey. Questions on sexual orientation and gender identity (SOGI) were included so researchers could use the data to compare responses from persons who identify as gay, lesbian, bisexual, and/or transgender with those of persons who do not identify themselves in these categories (Pierannunzi et al. 2017). The questions themselves are administered in two parts (Centers for Disease Control and Prevention 2017) as follows:

1. Do you consider yourself to be:
    1 Straight

2 Lesbian or gay
3 Bisexual
4 Other
7 Don’t know/Not sure
9 Refused

2. Do you consider yourself to be transgender?
1 Yes, Transgender, male-to-female
2 Yes, Transgender, female to male
3 Yes, Transgender, gender nonconforming
4 No
7 Don’t know/not sure
9 Refused

SOGI questions used in the BRFSS optional module were developed by a group of survey professionals within the US Department of Health and Human Services (Institute of Medicine 2013). The questions are similar to those proposed by the Williams Group (Herman 2014). A number of other SOGI question formats have been proposed and are in used on other surveys (Federal Interagency Working Group 2016). A total of 19 states participated in the optional module in 2014; 22 states used the module in 2015; 25 states participated in 2016, 28 states participated in 2017 and 2018. The list of states participating in the module by year is provided in Table 1.

Table 1: States participating in the SOGI optional module by year

Year	**Participating States**
2014	Delaware, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maryland, Minnesota, Montana, Nevada, New York, Ohio, Pennsylvania, Vermont, Virginia, Wisconsin, Wyoming
2015	**Colorado**, **Connecticut**, Delaware, **Georgia**, Hawaii, Idaho, **Illinois**, Indiana, Iowa, Kansas, Maryland, **Massachusetts** , Minnesota, **Missouri** , Nevada, New York , Ohio, Pennsylvania, **Texas**, Virginia, **West Virginia**, Wisconsin
2016	**California**, Connecticut, Delaware, Georgia, Hawaii, Idaho, Illinois, Indiana, Iowa, **Kentucky**, **Louisiana**, Massachusetts, Minnesota, **Mississippi**, Missouri, Nevada, New York, Ohio, Pennsylvania, **Rhode Island**, Texas, **Vermont**, Virginia, **Washington**, Wisconsin
2017	California, Connecticut, Delaware, **Florida**, Georgia, **Guam**, Hawaii, Illinois, Indiana, Iowa, Louisiana, Massachusetts, Minnesota, Mississippi, **Montana**, Nevada, New York, **North Carolina**, Ohio, **Oklahoma**, Pennsylvania, Rhode Island, **South Carolina**, Texas, Vermont, Virginia, Washington, Wisconsin
2018	**Arizona**, Connecticut, Florida, Guam, Hawaii, **Idaho**, Illinois, **Kansas**, Louisiana, **Maine**, Maryland, Minnesota, Mississippi, **Missouri**, Montana, Nevada, New York, North Carolina, Ohio, Oklahoma, Pennsylvania, Rhode Island, South Carolina, **Tennessee**, Texas, Vermont, Washington, **West Virginia**, Wisconsin

***States in bold are those participating in the SOGI module for the first time.**

The state-level sample allows for direct estimates for each state participating in the model. Direct state prevalence estimates, however, cannot be calculated for the states that did not participate in any given year. A method is needed, therefore, to model prevalence estimates where no data were collected. The model-based methods described in this article build on published methods to model estimates from one geographic area with sufficient direct observations to other areas. Such models are usually applied to achieve small-area estimates. In this instance, however, state estimates from direct observations are used to model estimates in other states. In the process, we are also able to generate national estimates of the prevalence of SOGI.

Overall, here are three main research questions this paper aim to address:

1. Can we estimate prevalence in those states which have not opted for the SOGI module?
2. What characteristics best predict LGBT estimates at the state level?
3. Do state laws and policies influence LGBT identification?

2. Methods

Our model-based estimates are based on multilevel logistic regression models for each of the dichotomous outcomes produced by the SOGI module in the BRFSS. Model-based estimates for states not using the SOGI module are produced by predicting the outcome for each respondent in the state using a wide range of predictors. As a result, national estimates are also made possible. Data from the 2014 – 2016 BRFSS were combined resulting in a sample size of 1,392,423. Weighted multilevel logistic regression models were built to identify characteristics associated with LGBT identification. The multilevel component allowed for generalization of linear models that vary at more than one level. Specific to these models, variance could be observed at the individual or state level. Additionally, the multilevel approach allowed for the accurate variance calculation relative to the clustering of data by state.

During the analysis, we had to contend with Don't Know (DK) responses and refusals to answer the questions for each of the outcomes. Our framework considered that the prevalence of each outcome in these categories (DK's or Refusals) was greater than for the population as a whole. We confirmed this premise by profiling the respondents in each of the DK/Refusal categories for the three outcomes along the dimensions defined by the predictors. A comparison with the profiles of respondents in each outcome group (e.g., the gay/lesbian group) confirmed that DK's and refusals were much more similar to these groups than to other respondents or to the population as a whole. As a result of this comparative analysis, we imputed responses in the DK-Refused at higher rates than random imputation would suggest. Specifically, we imputed 5% of the DK-Refusals to each of the gay/lesbian and bisexual categories. For the transsexual question, 2% of the DK-refusal responses were imputed as transsexual.

Preliminary bivariate analyses identified a set of individual level demographics as potential predictors of LGBT identification. These predictors were tested along with state level indices summarizing the acceptance of LGBT according to state laws.

Weighted multilvel logistic regression models were fit using SAS PROC GLIMMIX procedure. These models are an extension of logistical models which are appropriate for data organized at more than one level (i.e., nested data). In multilevel models, state-level indices related to sexual minority and transgender laws were computed and used, to incorporate state effects on LGBT prevalence. We defined two different indices to reflect state-level laws in the general sexual minority and in the more specific transgender domains.

The sexual minority index is created in each state as the sum of four separate (0-1) indicators defined as follows for the different laws in this area:

1. States don't have hate crime laws specifically protecting LGBT;
2. States don't have non-discrimination employment laws protecting LGBT;
3. States don't have laws prohibiting establishments from discriminating against LGBT customers;
4. States don't have laws making same-sex marriage legal.

The sexual minority law index is the sum of these four indicators (Indicators 1 to 4). Any indicator equals to 1 suggests a law that is harmful to sexual minorities, so that larger values of the overall sum index suggest a state environment that is negative towards the LGBT population.[1]

[1] 1. Source New York Times: https://www.nytimes.com/interactive/2017/08/25/opinion/sunday/worst-and-best-places-to-be-gay.html?action=click&pgtype=Homepage&clickSource=story-heading&module=opinion-c-col-right-region&region=opinion-c-col-right-region&WT.nav=opinion-c-col-right-region&_r=0

The transgender law index is computed similarly for each state as the sum of four different indicators for relevant laws:

5. States don’t have laws protecting youths from conversion therapy;
6. States don’t have explicit bans on excluding trans individuals from receiving health insurance coverage;
7. States don’t have laws for gender-neutral single-occupancy restrooms;
8. States have laws prohibiting transgender people from receiving documents reflecting their gender identity. [2]

A third index was created to reflect the state’s overall unwelcomeness of LGBT identification by summing the two defined indices.

We developed a total of four models using different dependent variables, the binary indicators of lesbian/gay identification, bisexual identification, transgender identification, and overall LGBT identification. The indices of state laws were included as random effects, while demographics and health indicators were entered as fixed effects. Similar to the sexual minority index, larger values of the overall sum index suggest a state environment that is negative towards the transgender population. We separated the data with available SOGI modules into training and validation datasets; we used the training data to build models and then used the validation dataset to evaluate our model performance.

Our fixed effect predictors are demographics and health behavior variables from BRFSS data, which includes alcohol use, if children present in home, education level, employment status, ethnicity, if ever take HIV test, marital status, mental health status, obesity, race, tenure, gender and smoker stauts. They are all dichotomous for this analysis.

3. Results

We developed weighted logistic multilevel models which incorporate the state-level effects associated with the laws in the LBGT domain. Figures 1 – 4 display odds ratios for fixed effects for each of the four models. Marital status was shown to be a strong predictor of lesbian/gay or bisexual identification. Presence of a college degree was shown to be a strong predictor of transgender identification, as well as LGBT status overall. Other predictors include ever having had a HIV test (gay/lesbian, bisexual) mental health (bisexual), and race (transgender, LGBT).

[2] Source: Wikipedia
https://en.wikipedia.org/wiki/Same-sex_marriage_in_the_United_States

Figure 1: Gay/Lesbian

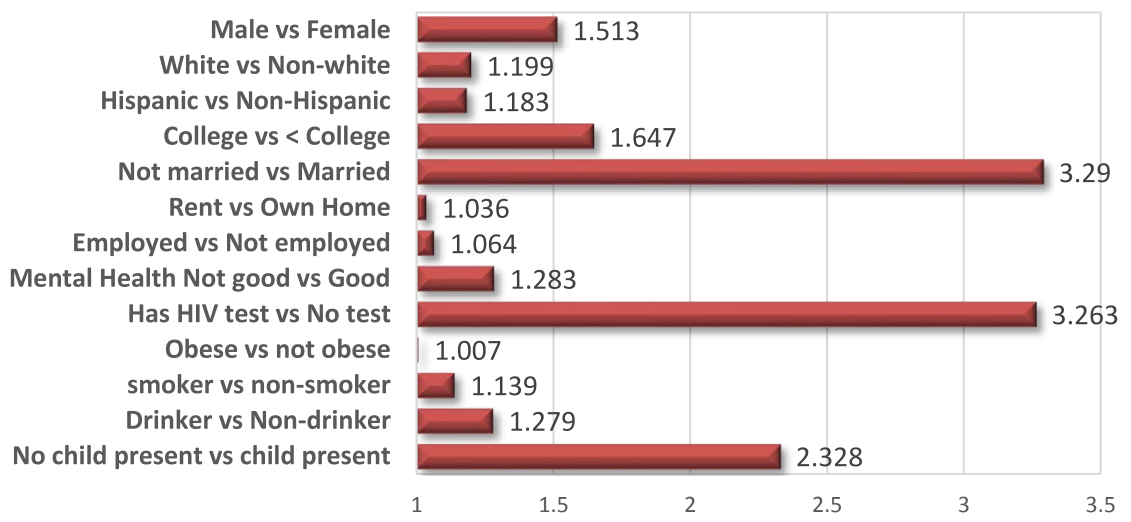


Figure 2: Bisexual

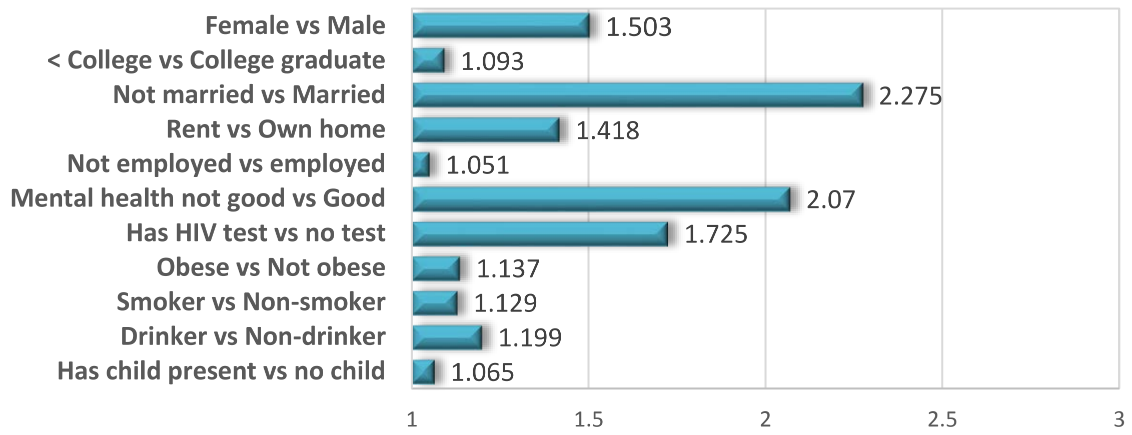
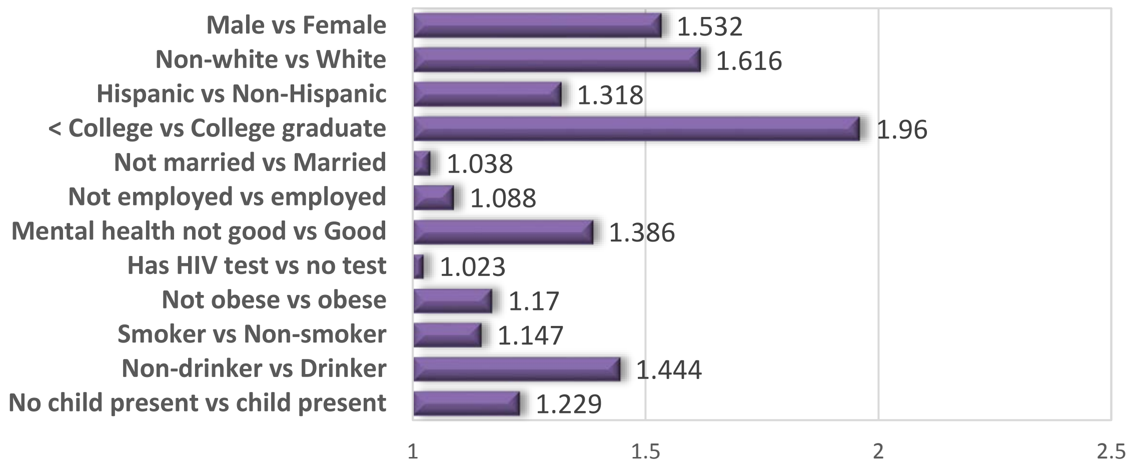


Figure 3: Transgender



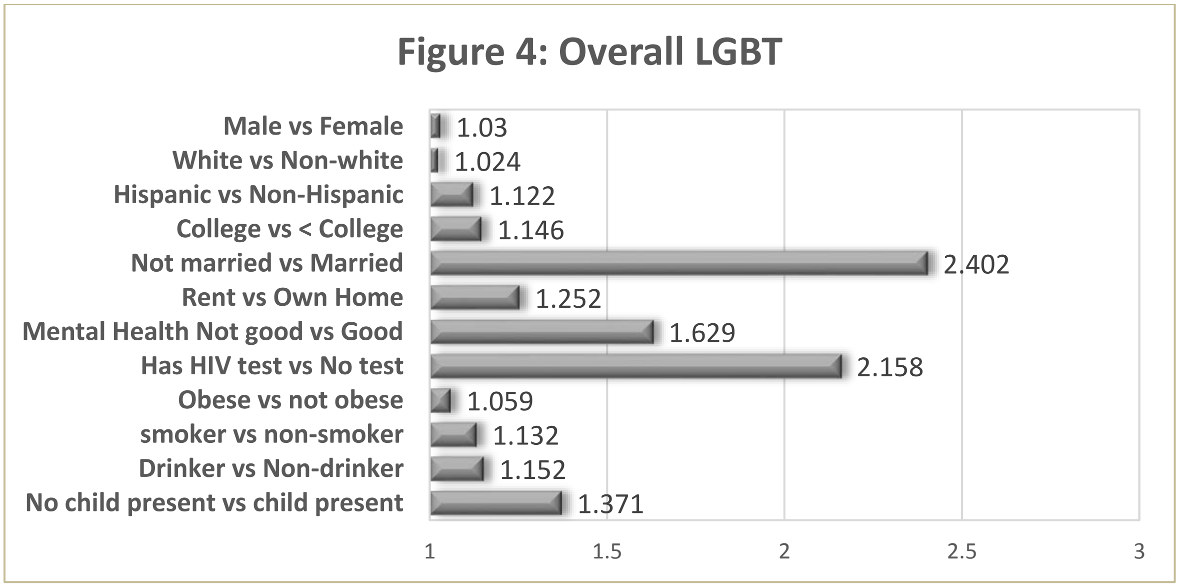


Figure 4: Overall LGBT

In our multilevel random effects models, state-level law indices are fit as random effects. The results presented in Table 2 for gay/lesbian and bisexual show that the more unwelcome same-sex laws, the lower the LGB prevalence. For transgender, however, the results are reversed and harder to interpret except for instability of the models for very low prevalence estimates.

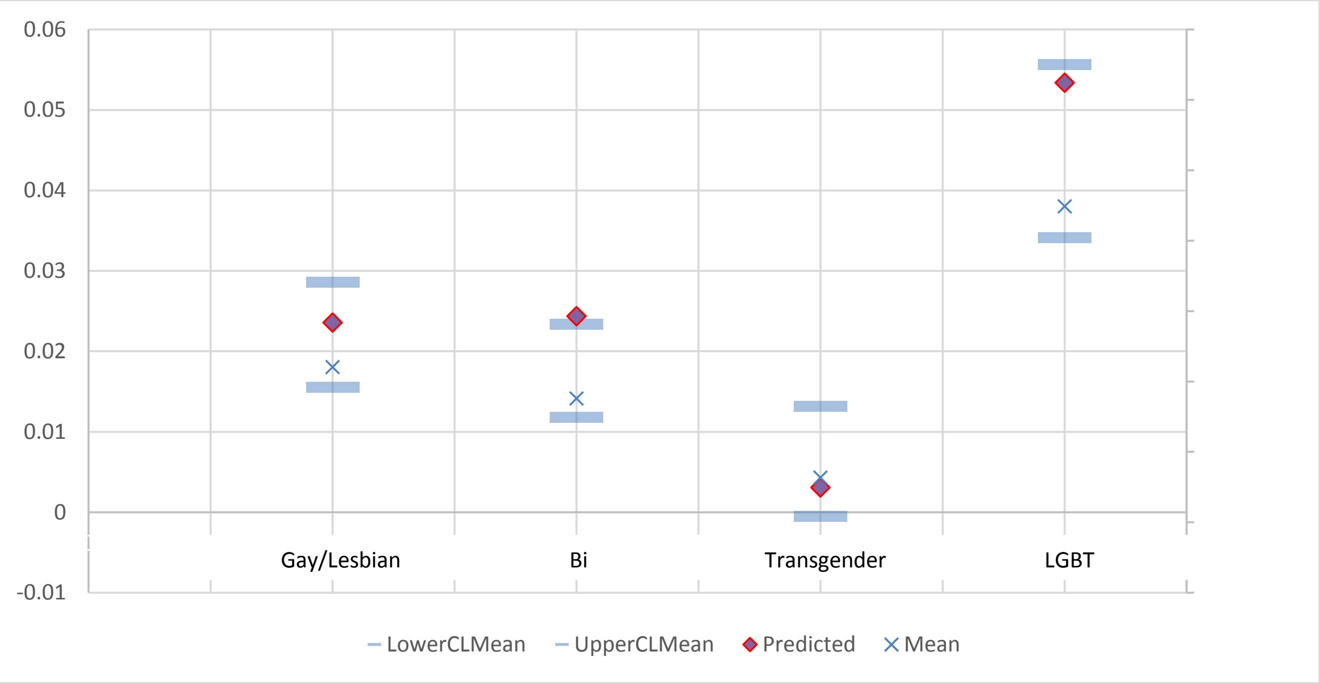
Table 2: Solution for Random Effects

	Estimate	**Std Err Pred**	**Pr > |t|**
Gay/Lesbian			
Same-Sex Unwelcome Index = 0	**0.1002**	**0.05155**	**0.0519**
Same-Sex Unwelcome Index = 1-3	-0.0292	0.05155	0.5713
Same-Sex Unwelcome Index = 4	-0.071	0.05156	0.1683
Bisexual			
Same-Sex Unwelcome Index = 0	**0.08237**	**0.0414**	**0.0467**
Same-Sex Unwelcome Index = 1-3	-0.0339	0.04141	0.4127
Same-Sex Unwelcome Index = 4	-0.0485	0.04141	0.242
Transgender			
Transgender Unwelcome Index= 0	**-0.1708**	**0.07283**	**0.019**
Transgender Unwelcome Index= 1	-0.0439	0.07281	0.5463
Transgender Unwelcome Index= 2	-0.0392	0.07279	0.5906
Transgender Unwelcome Index= 3	-0.0163	0.07279	0.823
Transgender Unwelcome Index= 4	**0.2702**	**0.07281**	**0.0002**
Overall LGBT			
LGBT Unwelcome Index = 0	**0.08688**	**0.02361**	**0.0002**
LGBT Unwelcome Index = 1	0.03041	0.02361	0.1978
LGBT Unwelcome Index = 2	**0.09899**	**0.0236**	**<.0001**
LGBT Unwelcome Index = 3	**-0.0548**	**0.02363**	**0.0203**
LGBT Unwelcome Index = 4	**-0.0839**	**0.02361**	**0.0004**
LGBT Unwelcome Index = 5	-0.0084	0.0236	0.722

LGBT Unwelcome Index = 6	**-0.0484**	**0.02362**	**0.0404**
LGBT Unwelcome Index = 7	-0.0208	0.02361	0.379

For model validation, we use the validation dataset to produce the weighted survey state-level estimates, then compare them with the state-level estimates produced from our models. When these estimates are close and the confidence interval for the direct survey estimate contains the model-based prediction, the models provide accurate predictions of the LGBT prevalence. Figure 5 shows one example of our validation concept. The figure show that the model estimates are within the confidence intervals for the direct survey estimates for gay/lesbian, transgender and overall LGBT. For gay/lesbian, there are 77% states with model predictions fall into the Cis of survey estimates, and for bisexual, transgender and overall LGBT, there are 84%, 90% and 87% states with model predictions fall into the Cis of survey estimates.

Figure 5: Model Validation for California



In the final stage of our analysis, we apply our models to all the states in the US to produce state-level predictions. Figures 6-9 show our final predictions. In the maps, darker orange hues indicate higher prevalence; blue the lowest prevalence and green the next lowest. The maps show that DC and CA have the highest prevalence for gay/lesbian, bisexual and overall LGBT. The transgender model-based prevalence estimates tend to be unstable at the state level.

Figure 6: Final Model Prediction for Gay/lesbian (%)

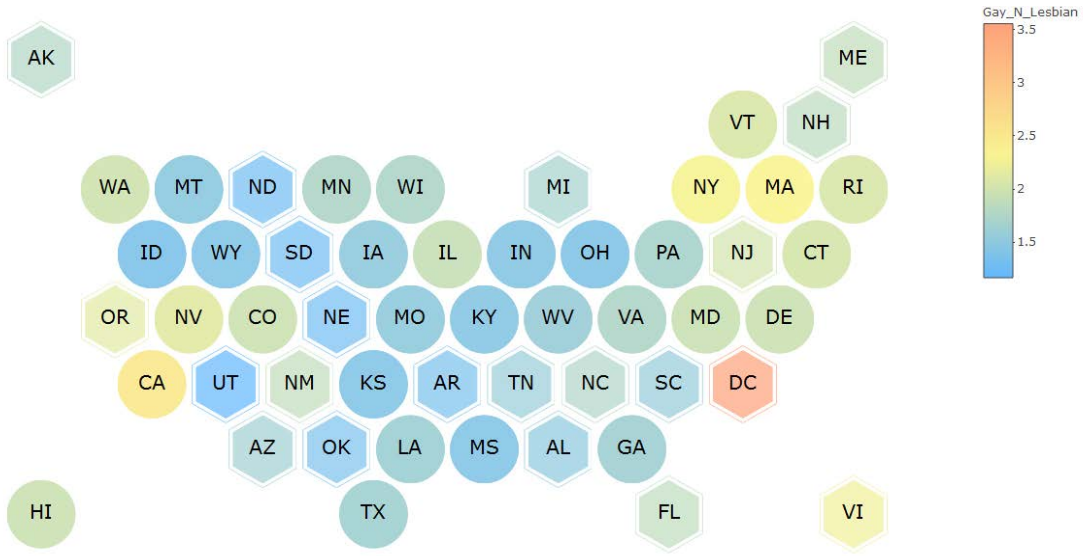


Figure 7: Final Model Prediction for Bisexual (%)

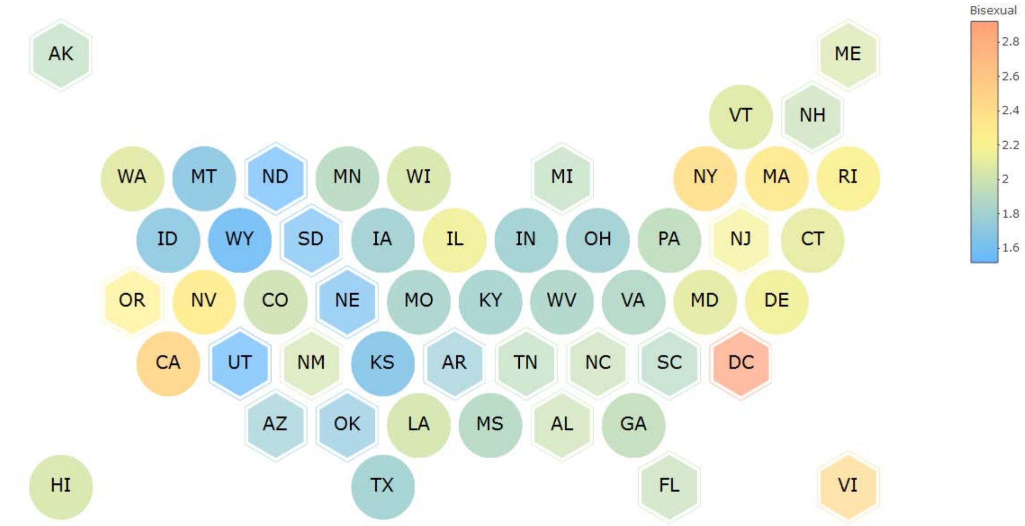


Figure 8: Final Model Prediction for Transgender (%)

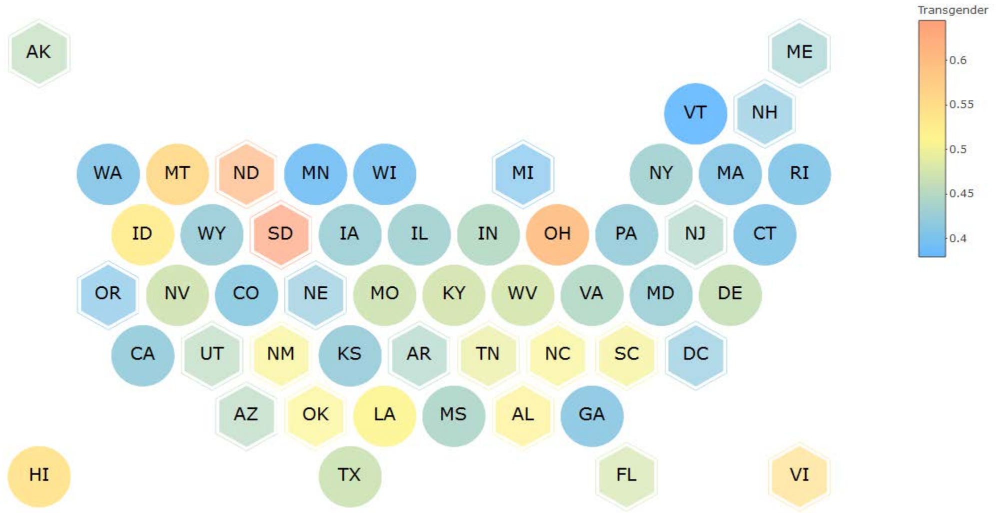
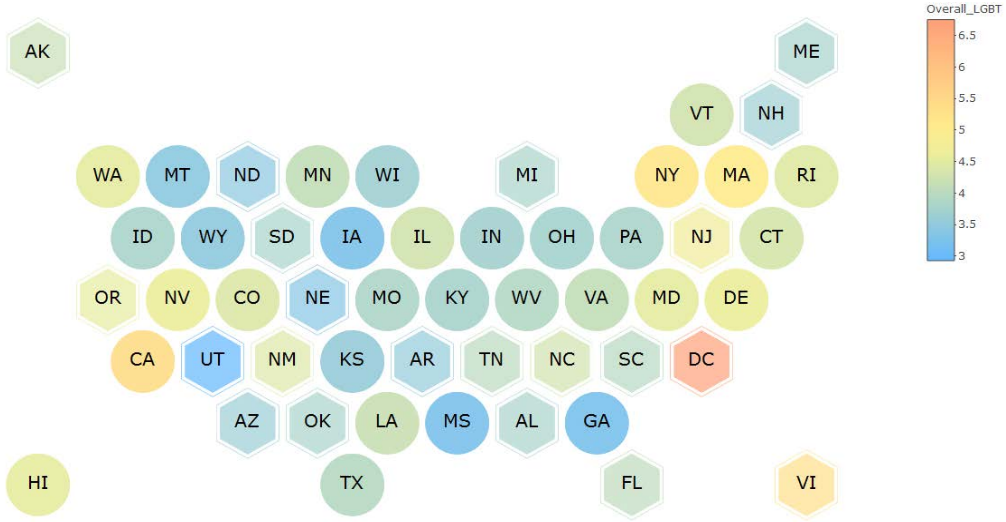


Figure 9: Final Model Prediction for Overall LGBT (%)



4. Conclusions and Discussion

This study showed the feasibility of developing multivariate models to generate state estimates that borrow estimation power from states with module data. The estimation methods were validated by comparing with states with direct survey estimates and sufficient observations. The methodology also supported the computation of national estimates based on the incomplete mosaic of states with module data. While developed in the context of the BRFSS data for SOGI outcomes, the approach can be used for other BRFSS topics and/or for other national surveys based on state samples.

The model-based methodology developed for this study can be applied to any BRFSS modules that are used in a subset of states as long as the number of states exceeds a minimum (15–16 states) in order to provide sufficient observations. Although an exact number of observations is not specified herein, researchers will have to take care when the number of states is low and/or the number of observations is a substantial portion of the total number of observations. Researchers are urged to review the application of this method to other variables where the total number of persons who report the variable of interest is low. Our research was conducted with a variable in which less than 1% of the total number of observations responded that they were transgender. It is unlikely, therefore, that researchers will apply the method to an indicator with lower prevalence in the state-level population; however, as a general rule, researchers applying the method must ensure that the demographic and/or risk groups are of sufficient size and scope to represent the other states.

As with all research, we found some limitations in our approach. Given that we began with a demographic that represented a small portion of the population, we believe that some of the variability of our approach resulted from the low prevalence estimates of persons who are transgender. This may not be a factor if the modeling approach were to be used for estimates of Body Mass Index (BMI), diabetes, or other health indicators found in the BRFSS. We also acknowledge that the treatment of persons who refuse to answer and/or answered “do not know” to any of the questions included in the analyses is subjective. In future research we intend to dive deeper into these responses. It may be that respondents in fact “do not know” their sexual orientation and/or gender identity, or it may be that some respondents do not understand the questions themselves. Another potential response bias is that the surveys are based on self-reports, which may cause different understandings among different respondents regarding to the question description. In addition, we found that there is a relatively high proportion of DNK and refusal answers among non-English speaking respondents, which may be caused by the difference interpretation when the questionnaire is translation into another language such as Spanish.

Future research will consider alternative imputation procedures for missing values as well as DNK and refusal answers. We will also consider adding 2017-2018 BRFSS data into our further analysis. In addition, we will update the state law index to match the changes in laws.

References

Centers for Disease Control and Prevention. 2017. *Behavioral Risk Factor Surveillance System.* April 18. http://www.cdc.gov/brfss.

Guo, Jing, Thomas Land, Jean M Zotter, Xingyou Zhang, Erica Marshall, and Wenjun Li. 2013. "Small-area Estimation on Current Asthma Prevalence Among Adults in Massachusetts Using BRFSS Survey Data." *American Public Health Association.* Boston, Massachusetts: American Public Health Association. 283510. https://apha.confex.com/apha/141am/webprogram/Paper283510.html.

Institute of Medicine. 2013. "Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records: Workshop Summary." *Board on the Health of Select Populations.* Washington DC: National Academies Press. Accessed April 18, 2017. doi:https://doi.org/10.17226/18260.

J.L. Herman, editor. 2014. *Best Practices for asking questions to identify transgender and other gender minority respondents on population surveys.* Los Angeles: The Williams Institute.

Khalil GM, Crawford CAG. 2015. "A Bibliometric Analysis of U.S.-Based Research on the Behavioral Risk Factor Surveillance System." *American Journal of Preventive Medicine* 48 (1): 50-57. doi:10.1016/j.amepre.2014.08.021.

Li W, Kelsey JL, Zhang Z, Lemon SC, Mezgebu S, Boddie-Willia C, Reed GW. 2009. "Small-Area Estimation and Prioritizing Communities for Obesity Control in Massachusetts." *American Journal of Public Health* 99 (3): 511-519. doi:10.2105/AJPH.2008.137364.

National Cancer Institute. 2017. *Cancer Prevalence Statistics: Approaches to Estimation Using Cancer Registry Data.* Accessed April 21, 2017. https://surveillance.cancer.gov/prevalence/approaches.html.

National Cener for Biotechnology Information. 2017. *Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records: Workshop Summary.* April 18. https://www.ncbi.nlm.nih.gov/books/NBK154077/.

Pierannunzi, C, F Xu, R C Wallace, W Garvin, K J Greenlund, W Bartoli, D Ford, P Eke, and G M Town. 2016. "A methodological approach to small area estimation for the Behavioral Risk Factor Surveillance System." *Preventing Chronic Disease.*

Song L, Mercer L, Wakefield J, Laurent A, Solet D. 2016. "Using Small-Area Estimation to Calculate the Prevalence of Smoking by Subcounty Geographic Areas in King County, Washington, Behavioral Risk Factor Surveillance System, 2009–2013." *Preventing Chronic Disease* 13.

Zhang, X, J B Holt, H Lu, A G Wheaton, E S Ford, K J Greenlund, and J B Croft. 2014. "Multilevel regression and post stratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the Behavioral Risk Factor Surveillance System." *American Journal of Epidemiology* 179 (8): 1025-33. doi:10.1093/aje/kwu018.

Zhang, Xingyou, James B Holt, Ann G Wheaton, Earl S Ford, Kurt J Greenlund, and Janet B Croft. 2014. "Multilevel Regression and Poststratification for Small Area Estimation of Population Health Outcomes: A Case Study of Chronic Pulmonary Disease Prevalence Using the Behavioral Risk Factor Surveillance System." *American Journal of Epidemiology* 179 (8): 1025-1033. doi:10.1093/aje/kwu018.

Zhang, Z, L Zhang, A Penman, and W May. 2011. "Using small-area estimation to calculate county-level prevalence of obesity in Mississippi 2007-2009." *Preventing Chronic Disease* 8 (4): 1-11.