# Predicting the Unobserved: Improving Sexual Identity Measures in Health Disparity Studies with Machine Learning and Resampling

Rona Fang-Yu Hu [1,2,3], Brady T. West [1]
[1]Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104
[2]SSRS, 1 Braxton Way Suite 125, Glen Mills, PA 19342
[3]Michigan Program in Survey and Data Science, University of Michigan, Ann Arbor, MI 48104

**Abstract**
Survey research on sexual identity often categorizes respondents as heterosexual, homosexual, and bisexual, but may miss more nuanced identities. Recent Federal recommendations regarding best practices for the measurement of sexual identity have called for the inclusion of "something else" response options. Prior research has suggested that estimates of health disparities between sexual identity subgroups can be affected if a "something else" response option is not provided, given that respondents who do not identify as heterosexual, homosexual, or bisexual may be forced to select one of these options even if they do not see them as being relevant. Unfortunately, some surveys lack this option. We propose an innovative method using random forests to predict four-category sexual identity based on surveys that provide these options, followed by the prediction of four-category sexual identity responses in surveys that do not include these categories but do include common covariates to construct the "unobserved" identities retrospectively. We utilize bootstrap resampling to capture uncertainty with this prediction process. Leveraging a split-ballot experiment in the 2015-2019 National Survey of Family Growth, we first fit benchmark models of interest to selected health outcomes as a function of sexual identity in the half-sample including "something else" as a response category. Using this as a training data set, we then develop a classifier for four-category sexual identity, and use the half-sample excluding "something else" as a test data set, where we develop predictions of four-category sexual identity and then use these predictions to estimate the models of interest. We repeat this process for each of several hundred bootstrap samples, and evaluate the ability of this methodology to recover the model of interest based on four sexual identity categories in the data set that did not offer them.

**Key Words:** Sexual Identity Measurement; Machine Learning; Health Disparity Estimates; National Survey of Family Growth (NSFG); Bootstrap Resampling

## 1. Introduction

Sexual identity is commonly measured with no more than three response categories in surveys. These categories are straight/heterosexual, gay/lesbian/homosexual, and bisexual. Including a "something else" response category captures the intricate nature and diversity in how people with different sexual orientations represent themselves. As transgender, intersex, and nonbinary populations gain visibility, researchers have started to acknowledge that binary measurements (e.g., heterosexual, homosexual) may not adequately capture how individuals experience and identify their sexual orientation. By offering a "something else" option, surveys provide respondents with the flexibility to express their identity in ways that do not fit within conventional labels, reducing the risk of misrepresentation and misclassification. This inclusion is particularly important for

enumerating marginalized populations, ensuring appropriate healthcare, and improving the validity of scientific research. Including "something else" facilitates more inclusive data collection and a clearer understanding of sexual identity (NASEM, 2022).

Recent work has suggested that adding a fourth response category, "something else," can improve the survey measurement of sexual identity (NASEM, 2022; Office of the Chief Statistician of the United States, 2023). Furthermore, empirical studies have shown that surveys that do not include "something else" as a response category may produce biased estimates of health disparities by sexual identity, with significant public health implications for sexual minorities (West and McCabe, 2021; West et al., 2024; Engstrom et al., 2024). Hence, the ideal survey measurement of sexual identity in studies on health disparities should include "something else" in addition to the conventional three-category responses.

This exclusion of a "something else" response category in the survey measurement of sexual identity may be viewed as a missing data problem or an instance of omitted-variable bias. That is, respondents whose "true" sexual identity is better represented by this category and would have chosen "something else" if this option existed are compelled to choose among predefined categories—heterosexual, homosexual, or bisexual—when answering surveys. Our objective is to retrospectively impute the missing/omitted "something else" identity for these respondents by using innovative machine learning techniques.

We draw upon a split-ballot experiment from the 2015-2019 National Survey of Family Growth (NSFG) to achieve this objective. In these years of the NSFG, respondents were randomly assigned to one of two treatment groups. The first group, TG1, was presented with a sexual identity question offering three response options (heterosexual, homosexual, bisexual). The other group, TG2, was given essentially the same question but with the additional option of "something else." The split-ballot design ensured that TG1 and TG2 were statistically identical on all other variables (including but not restricted to key socio-demographic characteristics), as our prior analyses have confirmed (West et al., 2024).

Conventional machine learning approaches typically rely on a random division of a single data set into training and test subsets. Our machine learning framework is distinct in that we use TG2, the group exposed to the "something else" category, as the training data set, and TG1, the group without this option, as the test data set.

We first train the model on TG2, which contains the "something else" response option, and then apply this model to TG1 to predict endorsement of the fourth sexual identity category. Using these predictions, we calculate revised estimates of health disparities. We further refine our estimates by applying bootstrap resampling techniques, generating an empirical distribution of health disparity estimates. Finally, we compare these estimates to the original health disparity calculations derived from TG2's four-response category data to assess the impact of including an imputed "something else" response category.

More specifically, using TG2 as the training data set, we apply the random forest algorithm of reinforcement learning to build models to predict the probability of choosing each of the four categories of sexual identity. The specific set of covariates can include the entire NSFG data set. In future efforts to impute the "something else" category of sexual identity in other data sets, we can restrict the inclusion to specific covariates of theoretical interest or accommodate specific survey data availability. The covariates in this study encompass demographic, behavioral, and health-related variables, such as age, education, marital status, sexual behaviors and risks, substance use, and prevalence of STDs.

The aims of this study are thus: (1) evaluate a framework for retrospectively constructing the "something else" category of sexual identity in existing data sets, (2) build models to predict "something else" based on TG2, and see how the models behave on TG1 with the newly predicted

sexual identity data, as compared to the original TG2, and (3) assess the effectiveness of using newly imputed values of sexual identity in TG1 to improve estimates of health disparities based on TG2.

## 2. Methods

### 2.1 The Split-Ballot Design of 2015-2019 National Survey of Family Growth (NSFG)

In the 2015-2019 NSFG, respondents were randomly assigned to two treatment groups. Half of the sample received the questionnaire that includes the sexual identity measurement of three response categories, and we call the first treatment group TG1. The other half of the sample received a questionnaire that included "something else" in addition to the common three categories; we call this half-sample TG2. These two groups should be otherwise equivalent, based on the split-ballot design. Our prior work shows that these two groups are otherwise identical on key socio-demographic measures (West et al., 2024).

### 2.2 Data and Variables

Data come from the 2015-2019 years of the NSFG, which includes a stratified multistage cluster sample of households. The NSFG conducts a personal interview of a randomly selected individual aged 15-49 in a sampled household to gather information about family structure, fertility, and reproductive health (Lepkowski et al., 2013).

In this study, we only focused on one sex (males) and one health outcome measured in the NSFG; future work will continue applying this methodology to other health outcomes for males and for the female subsample. The public-use NSFG data we used includes a pooled national probability sample of 9,746 males (4,884 in TG1, 4,733 in TG2, with another 129 surveyed respondents not answering the sexual identity question). About 2.5% of TG2 respondents chose "something else."

We select covariates not only for their scientific relevance to family formation and reproductive health but also in accordance with the established associations with sexual identity or interactions involving sexual identity when sexual identity serves as a predictor, as in prior studies (e.g., West et al., 2024; Engstrom et al., 2024). The family-related variables include current marital status (married, not married) and how many times the respondent had married (Kerridge et al., 2017); household size; and a dummy variable indicating whether the respondent was currently living with at least one child under the age of 18 (Weber, 2008). The reproductive health covariates include dummy variables for current sexual activity without contraceptive use (lifetime, and last sexual activity) and the use of various types of contraceptives (lifetime, and last sexual activity) (Charlton et al., 2013); current sexual activity without contraceptive use (Charlton et al., 2013); and whether the respondent wanted to have a child/another child in the future (Shenkman and Abramovitch, 2021). Other health behavior covariates include current substance use, including past-month binge drinking; cigarette smoking, including at the rate of a pack per day; marijuana use; and other illicit drug use (e.g., cocaine, crack, and crystal meth) (Drabble et al., 2021; Klare et al., 2021); the number of current sex partners, and risky sexual behaviors, including anal sex for men who have sex with men (Parmenter et al., 2020), and measures of sexual health, including sexually transmitted diseases (Everett et al., 2013). We also include the following demographic variables: age (18-19, 20-24, 25-34, and 35-49), education (less than high school, high school, or greater than high school), total family income ($0-$19,999, $20K- $34,999, $35K-$69,999, and $70K+), race (White, Black, or other), and a dummy variable of whether the respondent has Hispanic ethnicity.

The specific set of covariates can include the entire NSFG data set in theory. In future efforts to impute the "something else" category of sexual identity in other data sets, we can restrict the inclusion to specific covariates of theoretical interest.

### 2.3 Model Building and Verification of Imputed Sexual Identity

In this section, we discuss the rationale for building a model and how we verify whether our method is effective. We break down the process into seven steps, as follows.

**(1) Fit the substantive model of interest.**

We verify our model building efforts using the association between sexual identity and illicit drug use in the past year, estimated using the TG2 data, as a benchmark. First, we fit the substantive model of interest (a logistic regression model predicting illicit drug use in the past year as a function of sexual identity) to the training set, TG2, to generate the benchmark set of estimates. Based on prior work, the estimated association of sexual identity with illicit use of drugs in the past year among males varies depending on how sexual identity is measured (West and McCabe, 2021; West et al., 2023).

**(2) Draw 500 bootstrap samples of the training set (TG2).**

Following step (1), we draw 500 bootstrap samples from the training data set to ensure robustness and capture the variability of the estimates.

**(3) Build a random forest to predict four-category sexual identity in each bootstrap sample.**

For each bootstrap sample, we utilize the aforementioned covariates to construct a random forest that can be used to predict the four-category sexual identity of individuals (Breiman, 2001).

**(4) Refer the four predicted probabilities for each case in the test set (TG1), based on the random forest constructed using TG2, to a random Uniform(0,1) draw to impute four-category sexual identity.**

The predicted probabilities for each case in the test set (TG1) are then referenced against a random Uniform(0,1) draw, facilitating the imputation of the four-category sexual identity for each individual in the test data set.

**(5) Fit the substantive model of interest to that imputed data set and save the estimates.**

We next apply the primary substantive model to this imputed data set, and record the estimates of the coefficients in the logistic regression model.

**(6) Repeat steps (3)-(5) for each of the 500 bootstrap samples.**

We repeat this process of building a random forest, referencing the predicted probabilities, and fitting the substantive model to the imputed data for each of the 500 bootstrap samples. This comprehensive reiteration ensures a thorough evaluation of the model across multiple simulated data sets.

**(7) Based on the distribution of estimates of the substantive model coefficients across the bootstrap samples, examine whether we would make the same inference about the relationship of sexual identity to the outcome as in the original training set (TG2).**

Finally, we analyze the distribution of the estimates for the substantive model coefficients across all bootstrap samples to determine if the inferences about the relationship between sexual identity and illicit drug use remain consistent with those derived from the original training data set (TG2). This methodology not only validates the initial findings but also enhances the credibility of the results through rigorous resampling and predictive modeling techniques.

## 3. Results

Table 1 shows evidence of the substantial changes in estimated disparities introduced by the different sexual identity measurement styles since the samples randomized to TG2 and TG1 are otherwise identical. Based on prior work, the estimated association of sexual identity with illicit use of drugs in the past year among males varies significantly depending on how sexual identity is measured (West and McCabe, 2021; West et al., 2023).

**Table 1:** Design-adjusted estimates of logit models based on TG2 and original TG1

| | TG2: Est. Coefficient (SE) | TG1: Est. Coefficient (SE) |
|---|---|---|
| **Intercept** | -2.83 (0.10)* | -2.95 (0.09)* |
| **Gay** | 0.14 (0.37) | 0.91 (0.37) |
| **Bisexual** | -0.02 (0.40) | 1.35 (0.32)* |
| **Something Else** | 0.87 (0.35)* | N/A |

*$p \leq .05$.

In Table 2, adding the new estimates of health disparities between the original estimates based on TG1 and TG2 helps to gauge the performance of our machine learning modeling exercise. The new "intercept" coefficient, which indicates the expected log-odds of illicit drug use for straight or heterosexual males, is the same as the original TG1 intercept coefficient, but it has a smaller standard error of .02, compared to .09 in the original estimates.

**Table 2:** Design-adjusted estimates of logit models based on original TG1, imputed TG1, and TG2

| | TG1: Est. Coef | TG1: SE | New: Est. Coef (Mean) | New: Est. SE (SD) | TG2: Est. Coef | TG2: SE |
|---|---|---|---|---|---|---|
| **Intercept** | -2.95 | 0.09 | -2.95 | 0.02 | -2.83 | 0.10 |
| **Gay** | 0.91 | 0.37 | 0.84 | 0.21 | 0.14 | 0.37 |
| **Bisexual** | 1.35 | 0.32 | 1.20 | 0.34 | -0.02 | 0.40 |
| **Something Else** | N/A | N/A | 0.64 | 0.51 | 0.87 | 0.35 |

The estimated differences between the "something else" males and the heterosexual males in the new TG1 and in TG2 are similar, with the estimated coefficients being .64 versus .87. These differences in coefficients, given their standard errors, would not be statistically significant.

The coefficients for gay and bisexual both move in the "right" direction. That is, adding something else in TG1 reduces the original differences in health disparities between the original TG1 and TG2. The gay coefficient moves from .91 to .84, and the bisexual coefficient moves from 1.35 to 1.20. Though the discrepancies remain large, meaning there is still room for improvement in our machine learning model, these results are encouraging.

Table 3 below shows how the sexual identities would be predicted to shift in TG1 if the respondents were given the "something else" option. This is based on one randomly selected bootstrap sample. We find that an estimated 2.3% of heterosexuals, 4.2% of gays, and 8.3% of bisexuals would move to "something else." A fair number of identities are changed, with 124 to "something else" (2.5%).

**Table 3:** Predicted changes in sexual identities between the original TG1 and the imputed TG1, including row percentages

| Observed (original TG1) | Predicted (imputed TG1) | | | |
|---|---|---|---|---|
| | **Straight** | **Gay** | **Bisexual** | **Something Else** |
| **Heterosexual** | 4,438 (95.8%) | 15 (0.3%) | 71 (1.5%) | 108 (2.3%) |
| **Gay** | 25 (20.8%) | 79 (65.8%) | 11 (9.2%) | 5 (4.2%) |
| **Bisexual** | 78 (59.1%) | 21 (15.9%) | 22 (16.7%) | 11 (8.3%) |

We also examined the random forests out-of-bag (OOB) error rates to check the accuracy of our predictions. Based on the first few bootstrap samples, the OOB error rates ranged from 2.96% to 3.11%, indicating high levels of predictive accuracy.

## 4. Discussion

In this study, we evaluate a novel approach to imputing the "something else" sexual identity category in surveys where it is not explicitly measured. Utilizing data from a split-ballot experiment in the 2015-2019 NSFG, we apply a random forest machine learning model to predict four-category sexual identity based on respondents' demographic, behavioral, and health-related characteristics. The major innovation of our paper is methodological. Conventional machine learning typically relies on the random partition of a single data set into the training and test subsets. In contrast, our machine learning approach is distinct and innovative in that we rely on the original survey design, which featured a split-ballot experiment, to partition the data set into two subsets—TG2, the group with the "something else" category, as the training data set and TG1, the group without, as the test data set. This innovative approach allows us to maximize the number of cases in the reinforcement learning exercise because we would otherwise only use half of the sample (TG2) as the training data set and split the TG2 half-sample into the training data and test data subsets. By integrating the machine learning approach into a resampling framework, we can then calibrate the uncertainty (or the performance) of our random-forest algorithmic model.

The findings demonstrate that including the imputed "something else" category in TG1 improves health disparity estimates, bringing them closer to the "benchmark" estimates based on TG2, which was the half-sample with the category "something else" directly measured. This approach enhances the accuracy and inclusivity of sexual identity measurement in health disparity research when "something else" is not a response category.

There were 129 respondents who didn't report their sexual identity. We used listwise deletion to deal with these missing data. We acknowledge that these might be respondents who felt that the available response categories did not accurately describe their true sexual identity. Item nonresponse to sexual identity, as we interpret it, essentially means "none of the above" (three categories of straight, gay, and bisexual). Thus, one may view such a response as equivalent to reporting "something else." We suggest that we should consider these respondents' sexual identity as the *de facto* "something else," and we are currently evaluating the implications of this approach.

Despite decades of scholarship in sexual orientation studies, it is not until recently the measurement of sexual identity has received its deserved share of attention (NASEM, 2022). We cannot retrospectively redesign the previous survey questions and identify the health disparities by sexual identity in the past. Therefore, there is presumably no way to accurately document the trends in health disparities for sexual minorities. Even with increasing awareness of how to better measure sexual identity, there is no guarantee that all survey designers will be willing to include the fourth category of "something else" in questionnaires. Our approach provides a potential solution to these limitations. We suggest that the methodology we delineate in this study can be applied as a

framework for imputing additional sexual identity categories in surveys that do not measure them. Such exercises may lead us to see a different picture of how individuals with an unusual sexual identity fare regarding not just health but also a variety of other measures of well-being. While we have focused on the applications of our approach in health disparity studies in this paper, we believe that improved measurement of sexual identity will help improve empirical studies with a much broader scope of sexual minority well-being.

Furthermore, this approach will allow us to "bridge" different data sets and reveal what existed but was unobserved, thereby reducing omitted variable bias and improving the quality of survey data. Not only can we expand the empirical applications to other outcome measures of health, we can also expand the imputation to other measures aside from sexual identity, such as race and ethnicity. Although the federal government has its mandated versions of questions for inquiring about race and ethnicity, given the increasing complexity of racial and ethnic identities (Harris and Sim, 2002; Liebler, et al., 2017), we can reexamine the racial and ethnic disparities in health and well-being by imputing unobserved racial and ethnic identities.

Future research endeavors should focus on additional factors that could significantly alter estimates of disparities in the anticipated direction when applying the procedure evaluated here. Subsequent studies will strive to enhance the model's accuracy and examine its broader applicability across various health and well-being contexts. This will provide a more profound comprehension of the distinctive encounters and requirements of diverse sexual minority populations.

## Acknowledgements

This work was funded by NIH Grant number 1R03HD107236-01A1 (PI: West). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Eunice Kennedy Shriver National Institute of Child Health and Human Development.

## References

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Drabble, L. A., Mericle, A. A., Gómez, W., Klinger, J. L., Trocki, K. F., & Karriker-Jaffe, K. J. (2021). Differential effects of state policy environments on substance use by sexual identity: Findings from the 2000–2015 National Alcohol Surveys. *Annals of LGBTQ public and population health*, *2*(1), 53.

Engstrom, C.W., West, B.T., Schepis, T.S., and McCabe, S.E. (2024). Does the approach used to measure sexual identity in surveys affect estimates of identity-based health disparities differently by race? A randomized experiment from the National Survey of Family Growth. *Social Science & Medicine*. DOI: https://doi.org/10.1016/j.socscimed.2024.116887.

Everett, B. G. (2013). Sexual orientation disparities in sexually transmitted infections: examining the intersection between sexual identity and sexual behavior. *Archives of sexual behavior*, *42*, 225-236.

Harris, D.R. and Sim, J.J. (2002). Who Is Multiracial? Assessing the Complexity of Lived Race. *American Sociological Review*, 67(4), 614–627. doi: 10.2307/3088948.

Kerridge, B. T., Pickering, R. P., Saha, T. D., Ruan, W. J., Chou, S. P., Zhang, H., ... & Hasin, D. S. (2017). Prevalence, sociodemographic correlates and DSM-5 substance use disorders and other psychiatric disorders among sexual minorities in the United States. *Drug and alcohol dependence*, 170, 82-92.

Klare, D. L., McCabe, S. E., Ford, J. A., & Schepis, T. S. (2021). Prescription drug misuse, other substance use, and sexual identity: The significance of educational status and psychological distress in US young adults. *Substance abuse*, *42*(3), 377-387.

Lepkowski, J. M., Mosher, W. D., Groves, R. M., West, B. T., Wagner, J., & Gu, H. (2013). Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth.

Liebler, C.A., Porter, S.R., Fernandez, L.E., Noon, J.M. and Ennis, S.R. (2017). America's Churning Races: Race and Ethnicity Response Changes Between Census 2000 and the 2010 Census. *Demography*, 54(1), 259–284.

McCabe, S. E., Hughes, T. L., Bostwick, W. B., West, B. T., and Boyd, C. J. (2009). Sexual orientation, substance use behaviors and substance dependence in the United States. *Addiction*, 104(8), 1333-1345.

National Academies of Sciences, Engineering, and Medicine (NASEM). (2022). Measuring Sex, Gender Identity, and Sexual Orientation. Washington, DC: The National Academies Press.

Office of the Chief Statistician of the United States (2023). Recommendations on the Best Practices for the Collection of Sexual Orientation and Gender Identity Data on Federal Statistical Surveys. Office of Management and Budget (OMB).

Parmenter, J. G., Crowell, K. A., & Galliher, R. V. (2020). Subjective importance of masculinity as a factor in understanding risky sexual attitudes and behaviors among sexual minority men. *Sex Roles*, *82*(7), 463-472.

Shenkman, G., & Abramovitch, M. (2021). Estimated likelihood of parenthood and its association with psychological well-being among sexual minorities and heterosexual counterparts. *Sexuality Research and Social Policy*, *18*, 221-232.

Weber, S. (2008). Parenting, family life, and well-being among sexual minorities: Nursing policy and practice implications. *Issues in Mental Health Nursing*, *29*(6), 601-618.

West, B.T., Engstrom, C.W., McCabe, S.E., Schepis, T.S., and Tani, I.J. (2024). How a "Something Else" Response Option for Sexual Identity Affects National Survey Estimates of Associations Between Sexual Identity, Reproductive Health, and Substance Use. *Archives of Sexual Behavior*. DOI: https://doi.org/10.1007/s10508-023-02710-7.

West, B.T. and McCabe, S. E. (2021). Choices matter: How response options for survey questions about sexual identity affect population estimates of its association with alcohol, tobacco, and other drug use. *Field Methods*, 33(4), 335–354.