Predicting Psychological Distress in a Sample of U.S. Adults Identifying as Sexual Minorities

Lauren Berny

EDLD 654, Fall 2023

GitHub Repo Link

Predicting Psychological Distress in a Sample of U.S. Adults Identifying as Sexual Minorities

Prior research indicates that the LGBTQ+ (Lesbian, Gay, Bisexual, Transgender, Queer/Questioning, and others) population is at elevated risk for psychiatric distress. A recent meta-analysis on sexual minority status and mental health found that lesbian, gay, and bisexual individuals had significantly higher odds of having one or more psychological disorders than heterosexual individuals (Wittgens et al., 2022). Another meta-analysis examining the prevalence of non-suicidal self-harm among both sexual and gender minorities found that sexual and gender minorities are more likely to self-harm than heterosexuals, but transgender and bisexual individuals were at the highest risk (Liu et al., 2019). The minority stress theory (Meyer, 2003) is a well-accepted framework for understanding why higher rates of psychiatric distress occur in the LGBTQ+ population, postulating that increased stress stemmings from their minority status in society contributes to these adverse mental health outcomes. In other words, constant exposure to societal hostility, prejudice, victimization, expectations of rejection, and internalized homophobia lead to heightened psychological distress.

Research Problem

This project aims to predict psychological distress in a sample of LGBTQ+ adults residing in the United States. This is important at the individual level given the connection between poor mental health and low quality of life, but also at the societal level given that poor mental health contributes to strained healthcare systems and reduced workforce productivity (Kessler et al., 2008; McKenna, 2011). Predicting psychological distress in this population has multiple beneficial implications. Identifying who is at greater risk of psychological distress would allow for better-timed intervention and support, thereby potentially preventing severe mental health outcomes and helping allocate healthcare resources. The model could also help develop tailored interventions by identifying the factors that are most important in the predictive model. Moreover, having a deeper understanding of the factors influencing mental health in the

LGBTQ+ community can help improve public health policies related to inclusivity and gender-affirming care.

Methods

Dataset

Data from a national study of individuals who identify as sexual minorities were utilized for this project. Generations was a multi-year study that examined health and well-being across three generations of lesbians, gay men, and bisexual individuals who came of age in different time periods (Meyer, 2023). The study assessed the ways in which their identities were viewed, stress was experienced, and patterns of resilience varied across these different generations, particularly between the younger and older cohorts. Another aim of the study was to examine whether the effects of stress on mental health and well-being differed between these generations. A survey consulting agency recruited a nationally representative sample of U.S. sexual minorities ages 18-25, 34-42, and 48-55 using a dual sampling framework.[1] Participants completed a battery of measures online or through the mail across three different waves, each roughly one year apart, from 2016 to 2018. An important note is that the study's inclusion criteria was limited to sexual minorities who did not identify as transgender, and as such, transgender individuals were not represented in the study.

Sample

Only data collected during the first wave (i.e., baseline) of the study were used for this project, yielding a sample of 1,518 individuals. The younger cohort comprised 44.1% of the sample, whereas the middle and older cohort comprised 24.5% and 31.4% of the sample, respectively. The average age was 36.5 years (*SD* = 14.7). The race/ethnicity of the sample was 61.3% White, 15.2% Multiple Races, 11.9% Black, 10.4% Hispanic/Latinx, 1.3% other races (Asian, Native Hawaiian/Pacific Islander, American Indian, or Middle Eastern). The sex/gender of the sample was 49.4% cisgender women, 44.4% cisgender men, 4.1% non-binary females, and

[1] Sample weights were not included in the analysis.

2.1% non-binary males. The sexual identities were 54.9% lesbian/gay, 32.5% bisexual, 11.9% other sexual minority identity, and 0.7% heterosexual.[2] The sample was distributed across the United States, with 34.0% in the South, 28.2% in the West, 20.0% in the Northeast, and 17.9% in the Midwest. The median personal income range was $24,000 to $35,999.

Measures

The dataset included various measures across the following categories: positive health (e.g., life satisfaction, social well-being), health outcomes (e.g., chronic conditions, days of poor physical health), stressors (e.g., discrimination, victimization), social support (e.g., overall and subscales around friends, family, and significant others), identity (e.g., gender conformity, connectedness to LGBTQ community), and demographics (e.g., age, income). This project's GitHub repo also includes documentation related to the measures included under each domain. The outcome of interest was the total score on the Kessler-6, which is a popular screening tool for psychological distress with high validity and reliability (Kessler et al., 2010). Participants responded to six items about their mental health on a scale of 0 (never) to 4 (all of the time). The sum of the scores ranged from 0-24, with higher scores indicating greater psychological distress. The average score in the sample was 7.67, with a standard deviation of 5.49. A full list of the variables names and descriptions used in the analysis are included in Appendix A, which included 63 predictor variables across the aforementioned categories.

The study had a high participation rate at baseline, and most of the variables with missing data were imputed by the original study's researchers. As such, only two of the variables had missing data in approximately two percent of the cases. Two changes were made to the dataset. First, for modeling purposes, the small number of individuals identifying as Asian, Native Hawaiian/Pacific Islander, American Indian, or Middle Eastern were combined into an Other Races category. Second, the 12-level personal income variable was changed from a factor to numeric due to the number of levels and sparsity at the highest levels.

[2] Participants had to identify as a sexual minority to screen into the survey. However, in the full survey, a small number identified as heterosexual. I retained these individuals in the sample as I did not see documentation from the Generations study indicating they should be removed.

Analysis

First, a blueprint was created with two steps: (1) impute the two missing variables using bagged imputation; and (2) make dummy variables for each level of factor variables with more than two levels. Next, the dataset was randomly divided into a training set comprising 80% of the data and a testing set with the remaining 20%. The data were subsequently shuffled at random and split into 10 folds for cross-validation. Four models were estimated: a non-penalized linear regression model, lasso model, ridge model, and a bagged random forest model. Each model was trained on the same training set using 10-fold cross-validation and tested on the same testing set. All of the models used the same seed number and cross-validation folds.

The non-penalized model is a typical linear regression model, which assumes a linear relationship between the dependent and independent variables. It fits a straight line on the data by estimating the coefficients that minimize the sum of squared differences between the observed and predicted values, aiming to find the best-fitting linear relationship between the predictors and the dependent variable. Ridge and lasso models impose penalties on the linear regression model by introducing a small amount of bias in order to decrease variance. On one hand, ridge regression adds a regularization term to the linear regression model that constrains the coefficients and prevents them from growing too large. On the other hand, lasso regression adds a regularization term that penalizes by shrinking less important coefficients to zero, thereby discounting the least influential predictors. Both models are appropriate for the aim of this project because they help identify important features, attempt to strike a balance between bias and variance to increase predictive accuracy, and prevent overfitting. Thus, whereas the linear model has the assumption that variables are not highly correlated and can be prone to overfitting the data, ridge and lasso models' penalties help control model complexity and reduce variance (Kuhn & Johnson, 2013). The ridge regression model was trained over a grid of alpha (0) and lambda values ranging from 0 to 2 at intervals of .001. The lambda value associated with the lowest root-mean-square error (RMSE) was chosen as the best lambda value, which was 0.637. The trained model was then evaluated on the testing set using the RMSE, r-squared, and mean

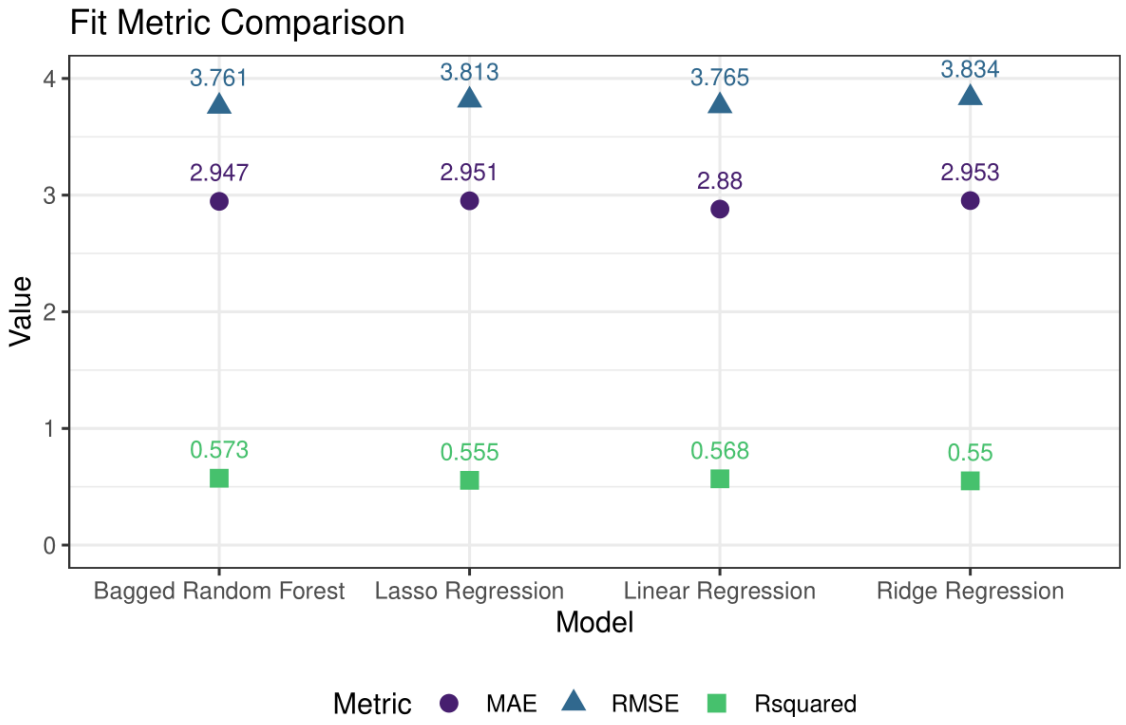
absolute error (MAE). The same process was used for the lasso model, with the exception that it was trained over a grid of alpha (1) and lambda values ranging from 0 to 2 at intervals of .001; a lambda of 0.075 was chosen based on its associated RMSE.

A random forest model is different from the other three models because it is based on recursive partitioning. It aggregates multiple decision trees trained on different subsets of the data with randomly selected predictor variables, which helps reduce overfitting and improve prediction accuracy. Bagging trains the individual decision trees on bootstrapped samples (i.e., sampling with replacement), allowing each tree to be exposed to slightly different versions of the data. I chose to use bagging because it enhances the model's robustness and generalization to unexposed data (i.e., the testing set) by combining the predictions of numerous trees trained on modified versions of the original dataset. This technique allows for more accurate predictions, making it an appropriate method for the aim of this project (Kuhn & Johnson, 2013). The bagged random forest model was tuned over a grid with the number of variables randomly sampled at each split specified from 5 to 60 at intervals of 5 and number of trees specified at 100, 250, 500, 750, and 1,000; variance was used as the split rule. The model was specified such that it sampled with replacement using a sample fraction of 0.8. The parameter combination that yielded the lowest RMSE was identified as 55 variables randomly sampled at each split and 500 trees. This trained model was subsequently evaluated on the testing set using the RMSE, r-squared, and MAE.

Results

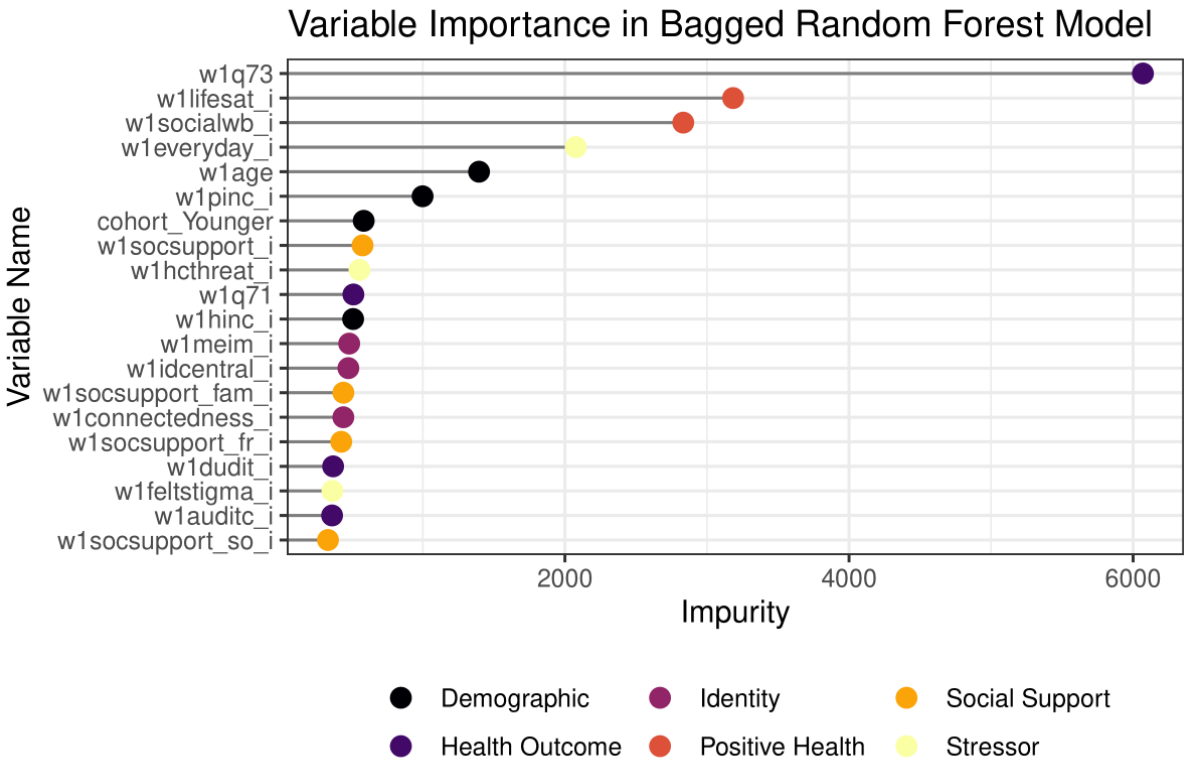
The fit metrics were compared across models to assess which of the models had the most optimal fit. As shown in the figure below, the fit of the models were roughly similar across all of the models. In terms of RMSE, the random forest model had the lowest RMSE (3.761), slightly below that of the non-penalized linear model (3.765). Similarly, the random forest model had the highest r-squared value (0.573), indicating that the model explained approximately 0.5% more of the variance in the data than the non-penalized linear model (0.568). MAE was noticeably lower in the non-penalized linear model compared to all of the other models. Broadly, the ridge model

has the worst fit of all the models, evidenced by it having the highest RMSE and MAE as well as the lowest r-squared value; the lasso model had a slightly better fit than the ridge model. Taken together, although the differences among the models are slight, I chose the random forest model because it had the best fit in two of the three metrics.



The random forest model assessed variable importance using the impurity metric, which indicates how effectively each variable decreased uncertainty and enhanced the overall purity in the ensemble trees. The figure below shows the 20 most important variables according to the model and which category the variables represent. Overall, it consisted of 4 health outcomes (number of days physical and mental health interfered with everyday life, poor physical health days, alcohol use disorder screening test, and drug use disorder screening test), 2 positive health factors (life satisfaction, social well-being), 3 stressors (everyday discrimination, healthcare stereotype threat, feelings of stigma), 4 demographics (age, personal income, being part of the younger cohort, and household income), 4 social support (overall social support, family social support, friend social support, significant other social support), and 3 identity (multi-group ethnic identity measure, connectedness to the LGBTQ+ community, and sexual minority identity

centrality). Thus, this model suggests that multiple domains are predictive of psychological distress.



Discussion

This project sought to predict psychological distress in a sample of U.S. adults who identified as sexual minorities. A non-penalized linear regression model, ridge model, lasso model, and bagged random forest model were estimated in pursuit of this aim. Using RMSE, MAE, and r-squared to evaluate fit, the random forest model was chosen as the best fitting model as it had the best performance in two of the three fit metrics. However, there were only slight differences in fit across the four models. This was the most surprising result as I expected there to be a somewhat larger difference between the three machine learning models and the non-penalized model. However, it is possible that changing the functional form of the variables in the lasso and ridge models would improve their fits. It is also possible that entering interaction terms into the model would yield help increase predictive accuracy, and it would also be theoretically justifiable given the importance of intersectionality between different identifies such as race and

sexual orientation. Although the models were tuned, finer tuning in the random forest model (e.g., minimum node size, maximum depth) could also be initiated to improve performance. As such, these results should be viewed as preliminary. A future avenue of inquiry in this area is estimating a binary model that uses score thresholds established by prior studies with the Kessler-6 a categorical indicator of psychological distress (yes/no).

The top 20 most important predictors of psychological distress were comprised of various variable categories, indicating that a variety of factors are influential in predicting psychological distress. The five most important predictors were the number of days physical and mental health interfered with everyday life, life satisfaction, social well-being, everyday discrimination, and age. I was not surprised that variables that are reflective of quality of life were the three most important predictors, nor was I surprised that everyday discrimination was the fourth important. I was surprised that age broke the top five, but I think that the presence of age and belonging to the younger cohort being in the top 10 most important predictors speaks to Generations' overall aim to explore different experiences across the three age cohorts. I was surprised that adverse childhood experiences, particularly around experiences of victimization, were not in the top 20 predictors. However, because the dataset was new to me and rather large, I did not include all of the variables in the dataset in the model. As such, it is possible that both the fit of the models and the important predictors would change as more variables are added to the blueprint.

Although these findings are preliminary, they still have some notable implications in the field of prevention science. The breadth of predictors reflected in the importance is a testament to collecting a comprehensive set of measures spanning various categories. For example, rather than only collecting data on risk factors, the Generations study collected data on positive health factors and social support, both of which were identified as important predictors. Moreover, the importance of age is made clear by this study. Although we commonly think of age as a predictor of physical health, it is less common to consider how age may influence mental health experiences. As noted earlier in the discussion, it would be interesting to see whether interactions with the cohort variable may further improve performance of the models.

References

Kessler, R. C., Green, J. G., Gruber, M. J., Sampson, N. A., Bromet, E., Cuitan, M., Furukawa, T. A., Gureje, O., Hinkov, H., Hu, C.-Y., Lara, C., Lee, S., Mneimneh, Z., Myer, L., Oakley-Browne, M., Posada-Villa, J., Sagar, R., Viana, M. C., & Zaslavsky, A. M. (2010). Screening for serious mental illness in the general population with the K6 screening scale: Results from the WHO World Mental Health (WMH) survey initiative. *International Journal of Methods in Psychiatric Research*, *19*(S1), 4–22. https://doi.org/10.1002/mpr.310

Kessler, R. C., Heeringa, S., Lakoma, M. D., Petukhova, M., Rupp, A. E., Schoenbaum, M., Wang, P. S., & Zaslavsky, A. M. (2008). Individual and societal effects of mental disorders on earnings in the united states: Results from the national comorbidity survey replication. *American Journal of Psychiatry*, *165*(6), 703–711. https://doi.org/10.1176/appi.ajp.2008.08010126

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Liu, R. T., Sheehan, A. E., Walsh, R. F. L., Sanzari, C. M., Cheek, S. M., & Hernandez, E. M. (2019). Prevalence and correlates of non-suicidal self-injury among lesbian, gay, bisexual, and transgender individuals: A systematic review and meta-analysis. *Clinical Psychology Review*, *74*, 101783. https://doi.org/10.1016/j.cpr.2019.101783

McKenna, M. (2011). The growing strain of mental health care on emergency departments: Few solutions offer promise. *Annals of Emergency Medicine*, *57*(6), A18–A20. https://doi.org/10.1016/j.annemergmed.2011.04.013

Meyer, I. H. (2003). Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. *Psychological Bulletin*, *129*(5), 674–697. https://doi.org/10.1037/0033-2909.129.5.674

Meyer, I. H. (2023). *Generations: A study of the life and health of LGB people in a changing society, United States, 2016-2019* (Version v2) [dataset]. Inter-university Consortium for Political and Social Research. https://doi.org/10.3886/ICPSR37166.V2

Wittgens, C., Fischer, M. M., Buspavanich, P., Theobald, S., Schweizer, K., & Trautmann, S. (2022). Mental health in people with minority sexual orientations: A meta-analysis of population-based studies. *Acta Psychiatrica Scandinavica*, *145*(4), 357–372. https://doi.org/10.1111/acps.13405

APPENDIX A

Variable Names and Labels

Variable Name	Variable Label
STUDYID	Study ID
W1WEIGHT_FULL	Survey weight full w1 sample
GEDUC1	Education
GRUCA_I	Rural-urban commuting area (RUCA) urbanicity score, missing imputed
GURBAN_I	Urbanicity with imputation
GCENREG	Census region
W1Q71	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?
W1Q73	During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
W1Q74_1	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Hypertension (high blood pressure)
W1Q74_2	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. High cholesterol
W1Q74_3	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Heart condition or heart disease
W1Q74_4	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Angina
W1Q74_5	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. A heart attack
W1Q74_6	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. A stroke
W1Q74_7	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Emphysema
W1Q74_8	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Asthma

W1Q74_9	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. An ulcer
W1Q74_10	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Cancer or a malignancy of any kind
W1Q74_11	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Diabetes
W1Q74_12	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Prediabetes, impaired fasting glucose, impaired glucose tolerance, borderline diabetes, or high blood sugar
W1Q74_13	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Arthritis, rheumatoid arthritis, gout, lupus, or fibromyalgia
W1Q74_14	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Blood clots in legs or lungs
W1Q74_15	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Osteoporosis or loss of bone density
W1Q74_16	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Thyroid problems
W1Q74_17	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Liver disease
W1Q74_18	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Chronic obstructive pulmonary disease (COPD)
W1Q74_19	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Crohn's disease or ulcerative colitis
W1Q74_20	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Kidney disease
W1Q74_21	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. HIV/AIDS
W1Q74_22	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Other sexually transmitted infection (not including HIV/AIDS)
W1Q74_23	Have you EVER been told by a doctor or health professional that you had any of the following? Please mark all that apply. Sleep disorder (e.g., insomnia or sleep apnea)
W1Q169	Do you have any children?

W1RACE	Race - survey
W1SAMPLE	Original or extended enrollment sample
W1SEX_GENDER	Sex and gender categories
W1AGE	Age on survey
W1SEXMINID	Sexual minority identity
W1PINC_I	Personal income with imputation
W1HINC_I	Household income with imputation
W1ACE_I	Adverse Childhood Experiences with imputation
W1ACE_EMO_I	Adverse Childhood Experiences (ACE) emotional abuse with imputation
W1ACE_INC_I	Adverse Childhood Experiences (ACE) incarceration household member with imputation
W1ACE_IPV_I	Adverse Childhood Experiences (ACE) household intimate partner violence with imputation
W1ACE_MEN_I	Adverse Childhood Experiences (ACE) household mental illness with imputation
W1ACE_PHY_I	Adverse Childhood Experiences (ACE) physical abuse with imputation
W1ACE_SEP_I	Adverse Childhood Experiences (ACE) parental separation or divorce with imputation
W1ACE_SEX_I	Adverse Childhood Experiences (ACE) sexual abuse with imputation
W1ACE_SUB_I	Adverse Childhood Experiences (ACE) household substance abuse with imputation
W1AUDITC_I	Audit-C with imputation
W1CHILDGNC_I	Childhood gender nonconformity with imputation
W1CONNECTEDNESS_I	Community connectedness with imputation
W1DUDIT_I	Drug Use Disorders Identification Test (DUDIT) with imputation
W1EVERYDAY_I	Everyday discrimination with imputation
W1FELTSTIGMA_I	Felt stigma with imputation
W1HCTHREAT_I	Healthcare stereotype threat with imputation
W1IDCENTRAL_I	Identity centrality with imputation
W1INTERNALIZED_I	Internalized homophobia with imputation
W1KESSLER6_I	Kessler 6 with imputation
W1LIFESAT_I	Satisfaction with life with imputation

W1MEIM_I	Multigroup Ethnic Identity Measure (MEIM_R) with imputation
W1SOCIALWB_I	Social well-being scale with imputations
W1SOCSUPPORT_FAM_I	Social support - family with imputation
W1SOCSUPPORT_FR_I	Social support - friends with imputation
W1SOCSUPPORT_I	Social support - full scale with imputation
W1SOCSUPPORT_SO_I	Social support - significant others with imputation