How Do Mental Health and Healthcare Access Shape Diabetes Risk? A Study of Understudied Populations Using Latent Class Analysis and Factor Analysis

Project Description and Abstract

This study examines the impact of healthcare access, health-related behaviors, and unobserved population-level traits on diabetes risk, with a specific focus on an underrepresented demographic characterized by poor mental health and limited access to healthcare services. The primary objective is to investigate how lifestyle patterns within this group contribute to adverse health outcomes and to identify subpopulations that may benefit from targeted interventions or enhanced care measures. Unlike traditional regression-based approaches, this research employs advanced unsupervised learning methods to uncover latent health profiles that may influence diabetes risk. Using data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), Latent Class Analysis (LCA) and Factor Analysis are utilized to identify underlying health subgroups and key health indicators. These techniques facilitate dimensionality reduction and enable the classification of individuals into latent classes, with a particular focus on their predictive capacity for diabetes status. To ensure model robustness and parsimony, the Bayesian Information Criterion (BIC) is used for model selection and tuning. Additionally, numerical and ordinal variables are binned into categorical groups to harmonize them with the measurement scales of other variables. This methodological approach enables a more comprehensive analysis of heterogeneous health behaviors and risk factors. The findings reveal significant associations between specific lifestyle factors, such as dietary patterns, and diabetes risk, particularly within vulnerable subpopulations. These insights provide a more nuanced understanding of how social, behavioral, and healthcare-related disparities contribute to health inequities, thereby informing the design of targeted health interventions for at-risk groups.

AI Usage Statement:

This report leveraged artificial intelligence (AI) tools to facilitate data preprocessing, grammatical refinement, and optimization of data visualizations. All AI-generated outputs were systematically reviewed to ensure accuracy, validity, and compliance with the university's academic integrity standards.

Team Members:

- Ash Sharma (ayushs11)
- Joe Li (zizhoul2)
- Samyak Pokharna (samyakp3)

Literature Review

Several studies have analyzed the predictors of diabetes using national health datasets:

Study 1:

Title: Using Latent Class Analysis to Identify Chronic Disease Risk Factors

Findings: This study utilized Latent Class Analysis (LCA) to uncover distinct groups of individuals sharing common risk factors such as hypertension, obesity, and limited healthcare access. The analysis revealed that these groups had significant correlations with diabetes prevalence, highlighting the importance of tailored interventions for at-risk populations.

URL: https://link.springer.com/article/10.1186/s12889-021-10608-z

Study 2:

Title: A Latent Class Analysis of Metabolic Syndrome Among Hispanics/Latinos Living in the United States in Relation to Cardiovascular Disease Prevalence

Findings:

The study utilized latent class analysis (LCA) to categorize individuals from a Hispanic/ Latino cohort into two primary clusters: those with relatively healthy metabolic profiles (Non-MetS cluster) and those exhibiting elevated clinical markers of metabolic syndrome (MetS cluster). The analysis confirmed that demographic factors like older age and a family history of coronary heart disease increased the likelihood of MetS classification. Membership in the MetS cluster significantly correlated with a higher prevalence of cardiovascular disease (CVD).

URL: https://www.proquest.com/openview/7ff84cf532f9b8b5d90cdd7675c507c1/1?pq-origsite=gscholar&cbl=18750

Both studies highlight the utility of grouping individuals by shared characteristics to predict diabetes outcomes. Our analysis builds on this by incorporating a wider range of variables and applying LCA with BIC for model optimization.

Data Processing and Summary Statistics

The dataset used in this study was derived from the BRFSS 2015 survey, which contains 441,456 records across 330 variables. BRFSS is a health-related telephone survey conducted annually by CDC. Researchers utilize both landline and cellular telephone surveys. They use randomly generated phone numbers with the removal of unlisted numbers to reach a broad spectrum of participants. For landlines, disproportionate stratified sampling is employed to direct more calls to high-density residential areas, making the sample representative. The survey is cross-sectional, capturing various features at a single point of time. Different weights are assigned to adjust for unequal chance of selection that could have resulted from the density or phone differences. Weights are essential in terms of compensating for non-responses, for either not answering the call or refusing to answer a survey question. The weights are adjusted to match the distributional properties in the general population, for example, younger respondents and low-income groups who have higher chances of nonresponse are weighted more heavily.

Relevant variables were selected based on research findings on most relevant diabetes risk factors, reducing the dataset to 22 variables. The processed dataset contains health indices such as blood pressure, cholesterol levels, and BMI. Health-related behaviors, including whether the individual regularly smokes, the presence of physical activities, and consumption of fruits and vegetables, are incorporated. It also covers self-reports from respondents regarding physical, mental, and general health, in which respondents report the number of days per month they are feeling unwell. Demographic information such as sex, age, and income levels are also part of the data.

As mentioned, we were interested in an underrepresented subgroup of population with poor mental health and no healthcare access. The data preprocessing are accomplished with the dplyr package in R. We filtered the dataset to include individuals who reported having 15 or more days a month feeling unwell mentally, and who have no access to healthcare. This leaves us with 1462 observations for our analysis. For clarity in the LCA and Factor Analysis, we conducted a binning process to convert certain non-categorical variables to categorical variables interpretable by these algorithms. For instance, the BMI values are dichotomized into healthy BMIs ranging from 18 to 25, and unhealthy BMIs otherwise. The general health variable was initially an ordinal variable of perceived health in general, with a scale from 1 to 5, and higher values representing poorer health. We defined levels 1 to 3 as healthy and levels 4 to 5 as unhealthy. The remaining variables, mostly demographics, are maintained in their original level structures.

We inspected key aspects of the cleaned dataset and examined how this compared to the unfiltered dataset. The dataset uses 0 for no diabetes, 1 for pre-diabetes, and 2 for diabetes. As shown in **table 1**, among the 17685 respondents who self-reported poor mental health, the mean score in diabetic status is significantly higher for individuals with no healthcare (0.4749) than those with healthcare (0.3967). For reference, the mean of this variable is 0.2969 for the general population. Other noticeable patterns are that those who have access to healthcare are more likely to be males than females (0.4166 to 0.3421), and belong to a younger age group (6.257 to 7.575). It is also crucial to understand correlational patterns before fitting models. **Figure 1** is a heat map for the correlations of all the variables in the cleaned dataset. It can be concluded from this that most variables groups will not exhibit high collinearity, and can be safely incorporated into the models. It is justifiable that most relatively high correlations are between general health and other variables, such as physical health (0.6070) and difficulty walking (0.4976).

Characteristic	Total Targeted Population (n=17685) Means	Healthcare (n=16403) Means	No Healthcare (n=1462) Means
Diabetes_012	0.4685	0.3967	0.4749
HighBP	0.5285	0.4774	0.5331

HighChol	0.5246	0.4651	0.5299
CholCheck	0.9587	0.8598	0.9675
BMI	30.1	30.07	30.11
Smoker	0.5926	0.6272	0.5895
Stroke	0.08547	0.06566	0.08724
HeartDiseaseorAttack	0.1591	0.1402	0.1608
PhysActivity	0.5812	0.5766	0.5816
Fruits	0.5379	0.4836	0.5428
Veggies	0.7398	0.699	0.7434
HvyAlcoholConsump	0.06913	0.0985	0.06651
NoDocbcCost	0.2414	0.6245	0.2072
GenHlth	3.498	3.503	3.498
MentHlth	27.36	27.55	27.35
PhysHlth	14.0	12.8	14.1
DiffWalk	0.4404	0.3748	0.4463
Sex	0.3482	0.4166	0.3421
Age	7.468	6.257	7.575
Education	4.718	4.423	4.744
Income	4.695	3.769	4.777

Table 1. Sociodemographic and behavioral characteristics of the diabetic population stratified by healthcare status. Means are presented with the columns.

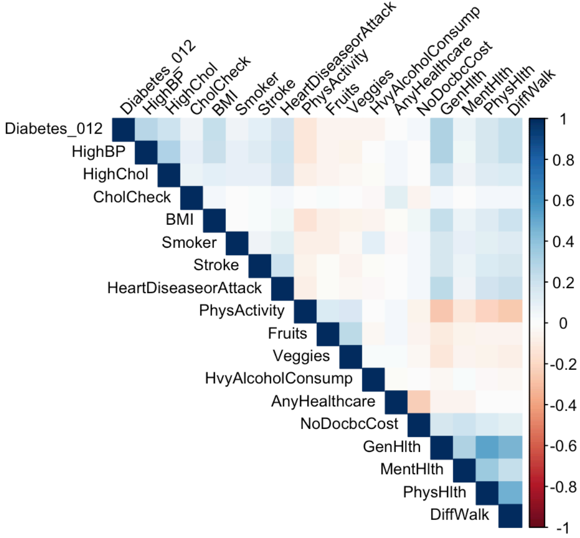

Figure 1. Heat map of Pairwise Correlations of All Variables in the Cleaned Dataset

Unsupervised Learning

The first main component of analysis involves Factor Analysis. We aimed to extract meaningful patterns about variable groupings that provide insights into further analysis using LCA. Factor Analysis is a technique used to identify underlying relationships between measured variables and to condense a large number of variables into a smaller set of latent factors. This method assumes that observable data are generated by a number of unobserved factors. Factor analysis is primarily used for data reduction purposes, or enhancing the interpretability of data by revealing the underlying structure. It helps understand which variables share common underlying dimensions and groups them accordingly. The factors are constructed to capture the maximum variance among the variables with a minimum loss of information.

Before conducting factor analysis, Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy are used to assess whether the dataset is suitable. Bartlett's Test of Sphericity checks if the correlation matrix is significantly different from the identity matrix, which would indicate that the variables are intercorrelated and therefore suitable for factor analysis. We obtained a chi-squared of 1897.804 for this test, with its p-value approximating 0, indicating significance and thus suitability for Factor Analysis. KMO evaluates the adequacy of sampling by comparing the magnitudes of observed correlation coefficients to the magnitudes of partial correlation coefficients. A KMO value closer to 1 suggests that a large portion of the variance among the variables might be explained by underlying factors. For our cleaned dataset result in a overall KMO of 0.73, and these values are all above 0.5 for the variables of interest. This signifies that factor analysis is likely to be reliable and meaningful. In addition, certain variables are reversed prior to analysis, to align in health implications with the majority variables and to increase interpretability of output. For instance, while 1 represents presence of physical activities, it is recoded as 0 to match a healthy implication in most variables such as the absence of smoking, high blood pressure, and high cholesterol, coded as a 0. Finally, a subset of variables are chosen for factor analysis and others are excluded. The excluded variables are mental health and healthcare access, which has no variability as the filtering criterion. Demographical variables such as sex, age, income, education level, are also omitted because they do not have straightforward health implications. We decided to have the binary variable “No Doctor Because of Cost”, and “Heart Disease or Attack” also excluded. While these variables do not exhibit high collinearity with others, they are excluded a posteriori due to the significant distortion to the Factor Analysis outcomes with their inclusion. Theoretically, it is justifiable to exclude them because the former is expected to be a closely associated with healthcare access, while the latter is too potent as an indicator of physical and general health.

Figure 2 represents a Parallel Analysis Scree Plot, used in factor analysis to determine the optimal number of factors to extract from the data. Parallel Analysis compares the eigenvalues from the actual data, shown by the blue line with triangle markers, with eigenvalues derived from simulated data, represented by the red dotted line, and resampled data, shown by the dashed red line. The objective of Parallel Analysis is to identify the point where the eigenvalues of the actual data fall below those of the simulated or resampled data. We selected the optimal number of factors to be 4 based on this result. It suggests that additional factors beyond four factors yield diminishing returns, and do not significantly contribute to explaining the variance in the data. The first four factors are significant, as they have eigenvalues larger than what could be expected by chance based on the simulated and resampled data. In this case, choosing four factors maximizes the explanation of the data's variance while avoiding overfitting by extracting too many factors that do not provide additional information.

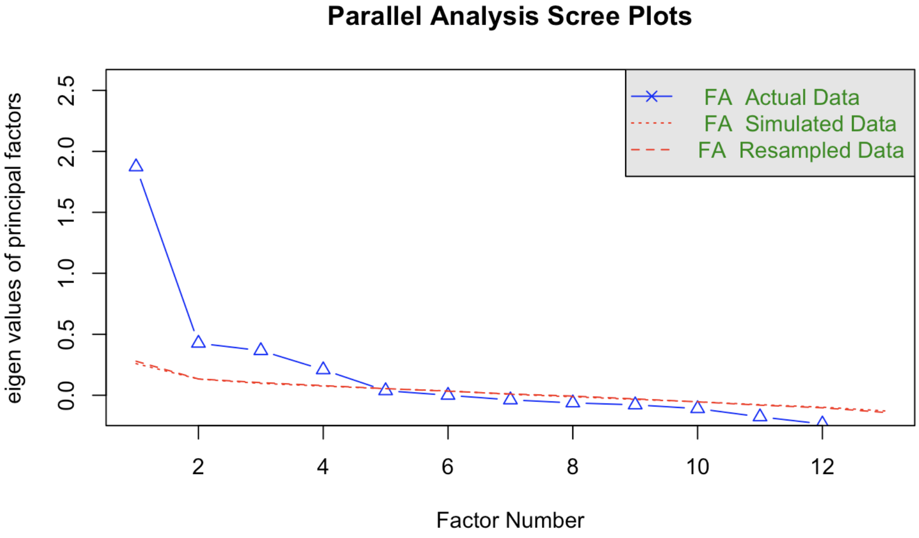

Figure 2. Scree Plot from Parallel Analysis Determines that 4 factors is Optimal for the Data.

Using this information and fitting the Factor Analysis with 4 factors, we obtained the grouped variables and factor loadings as displayed in **Figure 3**. Without knowing the variables semantically, the algorithm is capable of cluttering the predictors into meaningful groups. Factor MR1 can be interpreted as general health status. It groups physical health (0.7), difficulty walking (0.6), and general health (0.6). These relatively high loadings reflecting a high degree of shared variance among these variables, and that they all contribute significantly to the latent construct of overall health and physical functioning. Factor MR2, likely representing dietary habits, includes fruits with a high loading (0.7) and Veggies with a moderate loading (0.4). The variance explained by this factor highlights the combined effect of dietary choices. Factor MR3 focuses on measurable indices on cardiovascular risks, explained by the variables high blood pressure (0.7) and high cholesterol (0.4). These variables share common variance pertinent to cardiovascular health concerns. Lastly, Factor MR4 encompasses variables related to alcohol consumption and smoking behavior, with moderate loadings (0.3 and 0.4, respectively). There is also a moderate inverse loading for BMI (-0.3), implying an inverse relationship.

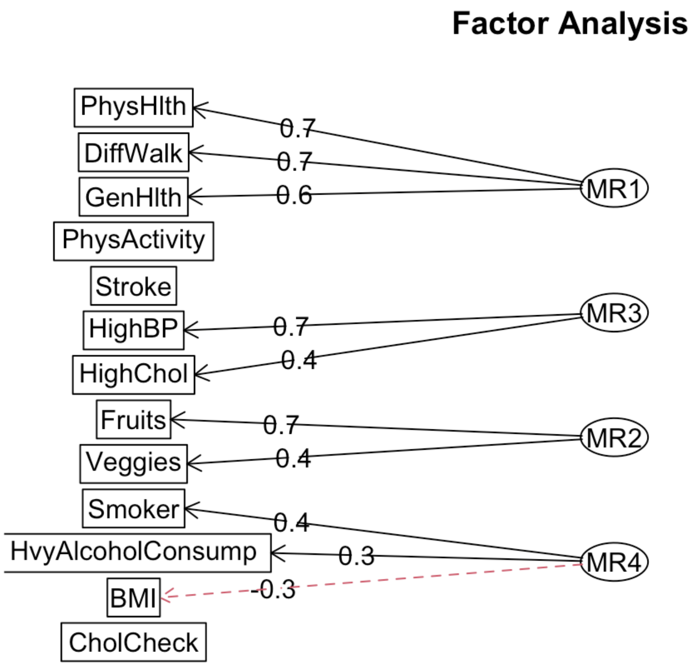

Figure 3: Factor Loadings and Groupings from Factor Analysis of Health-Related Variables.

To validate the clustering of health predictors from Factor Analysis and to create meaningful health profiles, we conducted Latent Class Analysis (LCA) as the second unsupervised learning component. In addition to the grouping of variables rendered by Factor Analysis, LCA allowed us to explore unobserved subgroups of respondents within our data, based on patterns of responses across the various health-related variables identified. This method assigns individuals to classes in a way that maximizes within-class homogeneity and optimizes between-class heterogeneity. LCA provides a rigorous statistical framework to categorize individuals into distinct classes that differ systematically in specific health behaviors and risk factors. Hence, it offers deeper insights into the underlying structure of the population's health. The LCA outputs can be crucial for tailoring targeted interventions by addressing the specific needs of each identified class.

As in Factor Analysis, binning of non-categorical variables is performed to make the data feasible for the model. We include a subset of variables in LCA identical to those in Factor Analysis. However, reverse coding of certain variables is unnecessary and thus is skipped for LCA, since the algorithm is proficient in identifying response patterns, and alignment is irrelevant.

We employed the Bayesian Information Criterion (BIC) to determine the optimal number of latent classes for our LCA model. An algorithm that resembles cross-validation iterating from 1 to 7 classes is used. As shown in **Figure 4**, The minimum BIC value is consistently obtained when the number of latent classes is set as 4.

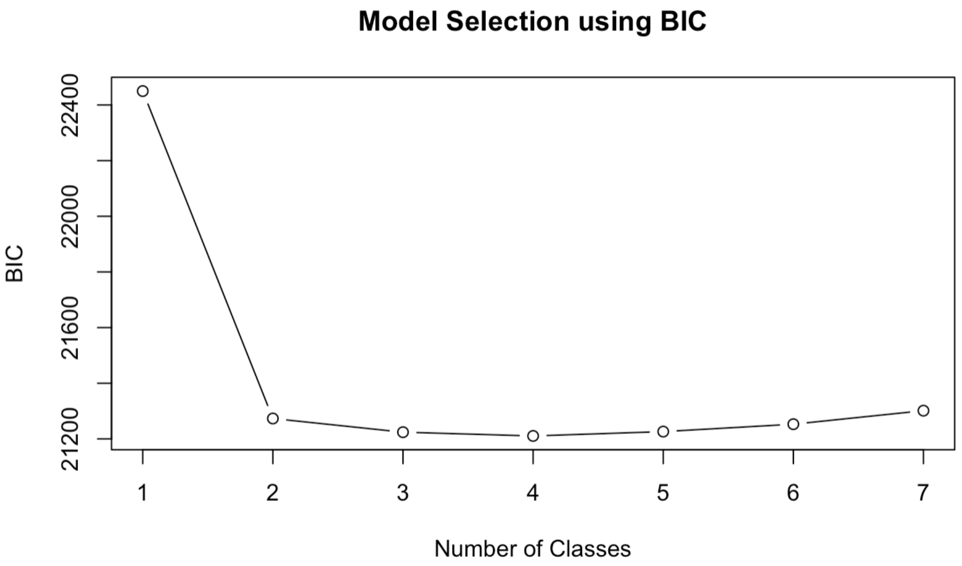

Figure 4. Modeling Selection for the Optimal Number of LCA Classes Using BIC

As illustrated in **Table 2**, LCA of health-related behaviors and conditions revealed four distinct classes within our dataset to reflect unique patterns of health profiles among participants. The specific columns to display in this table are selected to be the variable with the highest factor loading in each factor in our previous analysis. This assists us in examining whether the extracted variables of importance to each factor also varies significantly across assigned classes in LCA, but other lower-loading variables are also considered.

Class 1 can be characterized as the "Relatively Healthy Group". It comprises individuals with the lowest rate of poor general health (15.66%) and poor physical health (5.97%), and the highest engagement in physical activity (73.09%). This classify a subgroup of the population who are both confident about their health and have desirable measures in indices such as blood pressure as a confirmation. Class 2 is identified as the "Moderately Healthy Group with High Awareness", featuring the highest vegetable (90.16%) and fruits (70.07%) consumption rates, and a significant engagement in physical activities (57.32%). The lowest rate of high blood pressure (21.67%) lays in contrast with the second highest rate of self-reported poor physical health (76.35%). This can imply that Class 2 individuals are highly self-conscious, while their health conditions are not as bad in reality. We assign Class 3 as the "Unhealthy Group with Poor Awareness". This class displays markedly high rates of poor physical health (77.06%), the lowest consumption of vegetables (23.85%), and low physical activity (24.83%), with no fruit consumption (0%) reported. In application, the individuals may be specifically targeted for health-related education, and promotion of preventative treatments. Finally, Class 4, the "High-Risk Group," has notably high levels of unhealthy BMI (90.90%) and poor general health (81.06%), along with significant issues like high blood pressure (90.67%). This class is presumably the class with the worst overall health conditions, and special care should be offered. These classifications are effective in identifying patterns of diverse health behaviors and risk factors within the population.

Class	HighBP (1=Yes) (1=Yes)	PhysHlth (1=Poor) (1=Poor)	Fruits (1=Yes)	Smoker (1=Yes) (1=Yes)
1	29.69%	5.97%	49.59%	60.80%
2	21.67%	76.35%	70.07%	71.80%
3	64.53%	77.06%	0%	68.92%
4	90.67%	69.63%	64.42%	56.15%

Table 2. Probability Scores and Percentages Were Assigned to Classes and Distinct Identifiable Variables, with Key Columns Identified Through Factor Analysis

Model Prediction

Predictions of diabetic status can be made once the resulting classes from LCA is used to stratify the population. We attempt to associate these classes to the diabetic status variable and investigate their differences. **Table 3** represents the proportions of individuals with no diabetes, pre-diagnosed with diabetes and with diabetes. The digits in the parenthesis are the expected number of people in each cell, assuming that class assignment is uncorrelated with diabetic status. In terms of proportions, the expected proportions are 0.785, 0.033, and 0.182, if the classifications were trivial. It can be observed from this table that Class 1 (0.914) and Class 2 (0.827) individuals are significantly more likely to not suffer from diabetes compared to Class 3 (0.681) and Class 4 (0.557) individuals. The structured ordering of diabetic rate from Class 1 to Class 4 verifies that the risk levels we assigned to each classes in previous analysis are valid, and the patterns extracted for the critical variables are meaningful for predictions.

To further validate this, we performed a chi-squared test using the actual number of individual in each condition and class with respect to the predicted number if class had no effect. This yield a chi-squared statistic of 191 with 6 degrees of freedom, with a p-value subliminally small ($p < 2.2e-16$). An ANOVA test of mean difference across classes is also implemented, obtaining $F = 69.23$ with small p-value ($p < 2.2e-16$). These tests provide evidence that the association between class assigned by LCA and risk for diabetes is real.

Class	NoDiabetes (=0)	PreDiabetes (=1)	Diabetes (=2)
1	0.914 (647)	0.013 (9)	0.073 (52)
2	0.827 (158)	0.031 (6)	0.141 (27)
3	0.681 (162)	0.063 (15)	0.256 (61)
4	0.557 (181)	0.055 (18)	0.388 (126)

Table 3. Diabetes Rates for the Latent Classes and Individuals Within These Classes Stratified by Diabetes Type. Probability Scores are Provided With Frequencies Indicated in Parentheses.

Conclusions

The application of Latent Class Analysis (LCA) proved to be an effective method for identifying distinct lifestyle patterns within an underrepresented population characterized by limited healthcare access and poor mental health. The analytical outcomes were both interpretable and actionable, offering meaningful insights into the underlying health behaviors associated with diabetes risk. Validation through Factor Analysis further reinforced the robustness of the derived latent classes, revealing significant associations between specific lifestyle factors—such as dietary patterns—and diabetes outcomes.

These findings underscore the disproportionate burden of diabetes risk faced by this vulnerable subgroup, highlighting the potential for targeted interventions to mitigate health disparities. Moreover, the methodological framework established in this study is broadly generalizable, allowing for replication across other population subsets. Future research could apply a similar analytical approach to other demographic groups, thereby advancing the understanding of health behavior heterogeneity and informing population-specific intervention strategies.

References

1. Figner B, Weber EU. Who takes risks when and why? Determinants of risk taking. Curr Dir Psychol Sci. 2011;20(4):211–6. https://doi.org/10.1177/0963 721411415790.

2. Vanzile-Tamsen C, Testa M, Harlow LL, Livingstong JA. A measurement model of women's behavioral risk taking. Health Psychol. 2006;25(2):249–54. https://doi.org/10.1037/0278-6133.25.2.249.

3. Hingson R, Zha W, Smyth D. Magnitude and trends in heavy episodic drinking, alcohol-impaired driving, and alcohol-related mortality and overdose hospitalizations among emerging adults of college ages 18-24 in the United States, 1998-2014. J Stud Alcohol Drugs. 2017;78(4):540–8. https://doi.org/10.15288/jsad.2017.78.540.

4. DiMatteo MR. Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research. Med Care. 2004;42(3):200–9. https://doi.org/10.1097/01.mlr.0000114908.90348.f9.

5. Schwartz SJ, Weisskirch RS, Zamboanga BL, Castillo LB, Ham LS, Park HQ, Donovan R, Kim SY, Vernon M, Davis MJ, Cano MA. Dimensions of acculturation: associations with health risk behaviors among college students from immigrant families. J Couns Psychol. 2011;58(1):27–41. https://doi.org/10.1037/a0021356.

6. Josef AK, Richter D, Samanez-Larkin GR, Wagner GG, Hertwig R, Mata R. Stability and change in risk-taking propensity across the adult life span. J Pers Soc Psychol. 2016;111(3):430–50. https://doi.org/10.1037/pspp0000090.

7. Hanoch Y, Rolison JJ, Freund AM. Does medical risk perception and risk taking change with age? Risk Anal. 2018;38(5):917–23. https://doi.org/10.1111/risa.12692.

8. Courtenay W. Constructions of masculinity and their influence on men's well-being: a theory of gender and health. Soc Sci Med. 2000;50(10):1385– 401. https://doi.org/10.1016/S0277-9536(99)00390-1.

9. Wood W, Eagly AH. Two traditions of research on gender identity. Sex Roles. 2015;73(11-12):461–73. https://doi.org/10.1007/s11199-015-0480-2.

10. Wang XT, Zheng R, Xuan YH, Chen J, Li S. Not all risks are created equal: a twin study and meta-analyses of risk-taking across seven domains. J Exp Psychol: Gen. 2016;145(11):1548–60. https://doi.org/10.1037/xge0000225.

11. Kuhn DK. How do people know? Psych Sci. 2001;12(1):1–8. https://doi.org/1 0.1111/1467-9280.00302.

12. Bardi A, Schwartz SH. Values and behavior: strength and structure of relations. Personal Soc Psychol Bull. 2003;29(10):1207–20. https://doi.org/1 0.1177/0146167203254602.

13. Kim HS, Sherman DK, Updegraff JA. Fear of Ebola: the influence of collectivism on xenophobic threat responses. Psychol Sci. 2016;27(7):935–44. https://doi.org/10.1177/0956797616642596.

14. Sabogal F, Marín G, Otero-Sabogal R. Hispanic familism and acculturation: what changes and what doesn't? Hisp J Beh Sci. 1987;9(4):397–412. https://doi.org/10.1177/07399863870094003.

15. Schwartz SJ. The applicability of familism to diverse ethnic groups: a preliminary study. J Soc Psychol. 2007;147(2):101–18. https://doi.org/10.3200/SOCP.147.2.101-118.