Final Project

Soomin Kim

2022-12-14

Introduction

The All of Us Research Program by the NIH aims to gather data from one million or more participants in the United States to foster research and conversation in how to improve health and build targeted interventions. Many types of data are collected including surveys, electronic health records, genomic data, and blood and urine tests. In this study, we aim to look at the part of the All of Us cohort that responded to the social determinants of health survey, which has self-reported scores of stress (derived from Cohen's Perceived Stress Scale), loneliness (derived from UCLA Loneliness Scale), social support (derived from RAND MOS Social Support Survey Instrument), social cohesion (Social Cohesion Neighborhood Scale), and neighborhood disorder (Ross-Mirowsky Perceived Neighborhood Disorder Scale). It is a cross-sectional view of total of 57,365 respondents. We hypothesize that self-reported scores of loneliness will be associated with those of stress. We will explore the strength of that association in comparison to the other scores. The other covariates accounted for are age, sex at birth, sexual orientation, race/ethnicity, education, income, birthplace, home ownership, marital status, and health insurance status. As we want to account for confounding in the relationship between loneliness score and stress score, we will conduct correlation analyses and linear regression analysis.

Methods

We fitted linear regression models to investigate the relationship between the participant's self-reported loneliness score and stress score. The primary exposure loneliness score has range 0-4, and higher score means greater loneliness. The primary outcome stress score has range 0-4, and higher score means greater stress. NA values take up less than 10% of each variable except for income which has 13.9% NA values. We carried out nonparametric tests for univariate correlation analyses, as our outcome stress score is not normally distributed. Stress score has median of 1.2, mean of 1.313, is right-skewed with skewness of 0.49, and has a kurtosis of 2.8. All predictors were significantly associated with stress score at p-value of 0.05 (Table 1). We checked the linearity assumption for continuous predictors loneliness score, social support score, neighborhood disorder score, social cohesion score, and age by looking at the scatter plot of each vs. stress score. For subsequent linear regression modeling, we excluded NA values which yielded a total of 45,114 complete cases.

Although the outcome stress score is not normally distributed, our sample size is large enough for the central limit theorem to hold. Because we have many predictors, a total of 18, we first carried out univariate regression analyses to exclude any variable that is not significant at p-value 0.05 and decrease the problem of over-fitting. With the significant variables, we used automated forward selection algorithm to select variables for their significance at p-value 0.05 level. To account for the fact that the automated method would include only the significant confounders, we additionally checked the automatically removed variables and included them back in if they changed the loneliness score effect size by more than 10%.

Results

Extensive data wrangling was necessary to make a final tidy dataframe. In the beginning there were three different datasets, including the Demographics (age, sex, etc.), the Basics (income, education, birthplace, etc.), and the Social Determinants (loneliness score, stress score, etc.). Each respondent had multiple rows for each question they answered. I pivoted wide the dataframe so that each column is a question and the value is the answer to the question. I collapsed rows so that each respondent had one row. I selected the questions I wanted and then renamed the columns. I joined the three different datasets together by the respondent IDs. In order to carry out a regression analysis, I changed character values into 0 and 1 for binary variables or 0-3+ factors for multi-categorical variables (Figure 1).

Regression results showed that having a 1 score higher in the loneliness score was associated with a 0.555 higher score in the stress score, with all other predictors held constant (Table 2). The effect estimate was significant at p-value level of 0.05. To note, in our regression model, the race predictor was excluded by forward selection as it was not significantly associated with the stress score after adjusting for the other predictors. Our exposure of interest, loneliness score, turned out to have the biggest effect in association with the stress score, bigger in magnitude than any other scores. Not being married or partnered was associated with a 0.186 lower score in the stress score compared to those who are married or partnered, with all other predictors held constant, which was the biggest effect size for a predictor that is associated with lower levels of stress.

Conclusion

We explored how does the level of loneliness relate to the level of stress. We used correlation analyses and regression analysis to answer this question. After adjusting for social support score, social cohesion score, neighborhood disorder score, age, sex at birth, sexual orientation, race/ethnicity, education, income, birthplace, home ownership, marital status, and health insurance status, having 1 score higher loneliness score was significantly associated with 0.555 higher stress score. This magnitude is substantial given that the stress score range is 0-4. The analysis was successful considering that the p-value significance was very high. There are limitations such as the amount of missing values causing a significant drop in sample size for the regression model and that this survey was cross-sectional from which we cannot make causal inferences. Our next step would be to follow up on the respondents to make a longitudinal analysis and calculate causal risk for mental health outcomes such as depression or anxiety for varying levels of loneliness.

References

1. All of Us Research Program. All of Us Researcher Workbench. National Institute of Health. Accessed December 13, 2022. https://workbench.researchallofus.org/.
2. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav. 1983 Dec;24(4):385-96. PMID: 6668417.
3. Russell D, Peplau LA, Ferguson ML. Developing a measure of loneliness. J Pers Assess. 1978 Jun;42(3):290-4. doi: 10.1207/s15327752jpa4203_11. PMID: 660402.
4. Sherbourne CD, Stewart AL. The MOS social support survey. Soc Sci Med. 1991;32(6):705-14. doi: 10.1016/0277-9536(91)90150-b. PMID: 2035047.
5. Bateman LB, Fouad MN, Hawk B, Osborne T, Bae S, Eady S, Thompson J, Brantley W, Crawford L, Heider L, Schoenberger YM. Examining Neighborhood Social Cohesion in the Context of Community-based Participatory Research: Descriptive Findings from an Academic-Community Partnership. Ethn Dis. 2017 Nov 9;27(Suppl 1):329-336. doi: 10.18865/ed.27.S1.329. PMID: 29158658; PMCID: PMC5684777.
6. Ross, C. E., & Mirowsky, J. (1999). Disorder and Decay: The Concept and Measurement of Perceived Neighborhood Disorder. Urban Affairs Review, 34(3), 412–432. https://doi.org/10.1177/107808749903400304

Appendix

Predictors			N (%)	Median Stress Score	Correlation Statistics	p-value
Continuous Variables						
Loneliness Score	Range (0.375 – 4)		57321 (99.9)		Spearman r 0.528	< 0.001*
	NA		44 (< 0.01)			
Social Support Score	Range (0.25 – 5)		57312 (99.9)		Spearman r -0.291	< 0.001*
	NA		53 (< 0.01)			
Neighborhood Disorder Score	Range (0.308 – 4)		57349 (99.9)		Spearman r 0.290	< 0.001*
	NA		< 20** (<0.01)			
Social Cohesion Score	Range (0.25 – 5)		57065 (99.5)		Spearman r -0.283	< 0.001*
	NA		300 (0.523)			
Age	Range (18 – 103)		57365 (100)		Spearman r -0.377	< 0.001*
Categorical Variables						
Sex at birth	Female		34688 (60.5)	1.33	Wilcoxon	< 0.001*
	Male		19451 (33.9)	1.00		
	NA (including other)		3226 (5.62)			
Sexual orientation	Not heterosexual		5297 (9.23)	1.70	Wilcoxon	< 0.001*
	Heterosexual		48554 (84.6)	1.20		
	NA		3514 (6.13)			
Race/Ethnicity	Hispanic or Latino		3589 (6.26)	1.50	Kruskal-Wallis chi-squared 582.88	< 0.001*
	Non-Hispanic or Latino	Black	3129 (5.45)	1.44		
		Asian	1213 (2.11)	1.40		
		White	44284 (77.2)	1.20		
	NA (including other)		5150 (8.98)			
Education	Less than college		5133 (8.95)	1.50	Wilcoxon	< 0.001*
	College or more		48913 (85.3)	1.20		
	NA		3319 (5.79)			
Income	Less than 35k		9434 (16.4)	1.6	Kruskal-Wallis chi-squared 2239.1	< 0.001*
	35k – 75k		13091 (22.8)	1.3		
	75k – 150k		16509 (28.8)	1.1		
	More than 150k		10373 (18.1)	1.1		
	NA		7958 (13.9)			
Birthplace	Other		4539 (7.91)	1.2	Wilcoxon	< 0.001*
	US		49682 (86.6)	1.3		
	NA		3144 (5.48)			
Home ownership	No		14955 (26.1)	1.6	Wilcoxon	< 0.001*
	Yes		38879 (67.8)	1.1		
	NA		3531 (6.16)			
Marital status	Not married/partnered		19349 (33.7)	1.4	Wilcoxon	< 0.001*
	Married/partnered		34681 (60.5)	1.2		
	NA		3335 (5.81)			
Health insurance status	No		1091 (1.90)	1.67	Wilcoxon	< 0.001*
	Yes		52883 (92.2)	1.20		
	NA		3391 (5.91)			

Table 1. Stress score univariate correlation test results
**Any aggregates below 20 is suppressed following All of Us Research Privacy Policy

person_id	survey_datetime	survey	question_concept_id	question	answer_concept_id	answer	survey_version_concept_id	survey_version_name
<dbl>	<chr>	<chr>	<dbl>	<chr>	<dbl>	<chr>	<lgl>	<lgl>

person_id	survey_datetime	education	employment	birthplace	home_ownership	income	marital_status	insurance	sexual_orientation	date_of_birth	gender	<
<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	

Figure 1. Before and after data wrangling
** Individual level data is suppressed following All of Us Research Privacy Policy

	Univariate			Forward Selection		
Predictors	**EE**	**95% CI**	**p-value**	**EE**	**95% CI**	**p-value**
Loneliness Score	0.665	(0.657, 0.673)	< 0.001*	0.555	(0.544, 0.566)	< 0.001*
Social Support Score	-0.219	(-0.225, -0.214)	< 0.001*	-0.015	(-0.022, -0.008)	< 0.001*
Neighborhood Disorder Score	0.450	(0.438, 0.462)	< 0.001*	0.120	(0.107, 0.133)	< 0.001*
Social Cohesion Score	-0.327	(-0.335, -0.318)	< 0.001*	-0.044	(-0.053, -0.035)	< 0.001*
Age	-0.0188	(-0.0192, -0.0185)	< 0.001*	-0.013	(-0.013, -0.012)	< 0.001*
Sex at birth (Female)	0.267	(0.254, 0.281)	< 0.001*	0.165	(0.154, 0.179)	< 0.001*
Sexual orientation (Not heterosexual)	0.400	(0.378, 0.422)	< 0.001*	0.069	(0.051, 0.088)	< 0.001*
Race/Ethnicity (Hispanic or Latino)	0.269	(0.242, 0.295)	< 0.001*			
Race/Ethnicity (Black)	0.169	(0.141, 0.197)	< 0.001*			
Race/Ethnicity (Asian)	0.170	(0.126, 0.214)	< 0.001*			
Education (Less than college)	0.227	(0.205, 0.249)	< 0.001*	0.054	(0.034, 0.075)	< 0.001*
Income (75k-150k)	0.07	(0.051, 0.089)	< 0.001*	0.025	(0.009, 0.040)	0.001*
Income (35k-75k)	0.202	(0.182, 0.221)	< 0.001*	0.074	(0.057, 0.090)	< 0.001*
Income (Less than 35k)	0.502	(0.481, 0.523)	< 0.001*	0.160	(0.139, 0.181)	< 0.001*
Birthplace (Other than US)	0.049	(0.026, 0.073)	< 0.001*	-0.043	(-0.064, -0.022)	< 0.001*
Home ownership (No home)	0.446	(0.432, 0.460)	< 0.001*	0.062	(0.047, 0.077)	< 0.001*
Marital status (Not married/partnered)	0.212	(0.198, 0.225)	< 0.001*	-0.186	(-0.200, -0.173)	< 0.001*
Health insurance status (None)	0.356	(0.310, 0.402)	< 0.001*	0.042	(0.0005, 0.0832)	0.047*

Table 2. Stress score linear regression modeling results