Transgender Healthcare in the U.S.

Zhen Fang, Candus Shi, Jiyeon Song, Yuancheng Cao
Diversity in Data Science @ University of California, San Diego

May 18, 2023

1 Introduction

explain about health care + anti-trans laws involving gender-affirming care bans, etc.

2 Dataset & EDA

The dataset was obtained from the public-use data of the 2016-2018 TransPop Study, available on ICPSR and conducted by researchers at the Williams Institute at UCLA. TransPop is "the first national probability sample of transgender individuals in the United States" consisting of three surveys: TransPop 1 (April - August 2016), TransPop 2 (June 2017 - December 2018), and a cisgender national probability sample for the purpose of comparison (February 19 - 23, 2018 and November 12 - December 10, 2018). In total, there are 274 transgenders and 1,162 cisgenders in the dataset. The combined dataset has a dimension of 1,436 rows and 613 columns. Below is a snapshot of the dataset.

Table 1: a sample row

	1
X	0
STUDYID	151768927
WEIGHT_CISGENDER_TRANSPOP	0.02203922
WEIGHT_CISGENDER	NA
WEIGHT_TRANSPOP	0.9861429
GMETHOD_TYPE	
SURVEYCOMPLETED	0
GRESPONDENT_DATE	26-APR-2016
GCENREG	1
RACE	6
RACE_RECODE	1
RACE_RECODE_CAT5	1
SEXUALID	1
SEXMINID	0
HINC	11
HINC_I	11
PINC	4
PINC_I	4
GEDUC1	2
GEDUC2	2
GANN_INC	7
GANN_INC2	NA
GD74	1
GD75	1
GD76	NA
GEDUCATION	4
GEMPLOYMENT2010	3
GMSANAME	MASKED BY ICPSR
SEX	2
GENDER_IDENTITY	4
TRANS	2
TRANS_CIS	1

Table 1 shows the transpose of the first row in the combined dataset, limited to the first 32 columns. The variable or parameter of interest in our models is TRANS_CIS.

We conduct some exploratory data analysis on the demographics of the survey respondents, grouped by transgenders and cisgenders.

3 Methods

The high-level hypothesis we propose is that transgenders have worse health care and health care access than cisgenders. To test this, we split health care into different subsets based on the questions available in the survey and conduct hypothesis tests on the coefficient of the transgender-vs-cisgender variable in each regression model. To account for multiple testing issues such as family-wise error rate (FWER) and false discovery rate (FDR), we implement the Benjamini-Hochberg Procedure in our decision rules for each test.

Model-specific methods. Due to the large number of columns in the dataset, we will selectively choose certain columns to include as covariates in our linear models. This is done subjectively and based on our judgement of whether a certain demographic or question is important or relevant to the question being predicted. Then, we use stepwise subset selection and lasso regression to systematically choose features for the final models. Multicollinearity and two-way interaction terms are considered and considered in each feature selection method. All models are trained on the entire dataset, based on the missingness of all variables (explanatory and prediction).

4 Models

4.1 Cost

4.2 HIV

4.3 Mental Health

4.4 Insurance

5 Inference & Results

6 Conclusion

Acknowledgements

We thank Dr. Haojian Jin, Dr. Daniel Ries, Yuyao Wang, and Suraj Rampure for extensive discussions and guidance.

References