An Interactive Application for Identifying and Explaining Mental Health Provider Shortages by Population

CSE6242 - Team 03

Mohit Aggarwal, James Boyle, Nathan Cook, Andrew Doss, Suzi Pike

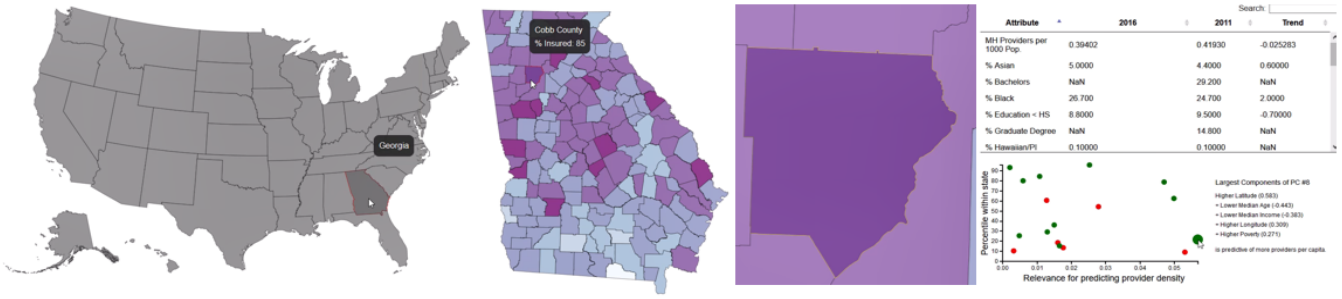

Figure 1: *Selected views from the dynamic visualization web application. Navigation proceeds through nested geography levels of state, county, and census tract. Second from Left: All features can be viewed as layers on choropleth heat maps. Right: Individual counties and tracts can be studied in detail using a descriptive data table and results from a predictive model indicating which combinations of features are distinctive to the area and relevant to predicting the supply of mental health providers in the county or state.*

1 INTRODUCTION – MOTIVATION

Studies [6,10] have shown that shortages of mental health providers limit patient access to required treatment. In addition to provider supply, sociodemographic factors such as age, race, income and health insurance, are associated with access to care and health outcomes [2,11].

2 PROBLEM DEFINITION

Public health stakeholders use numerous data sets to derive insights into the prevalence of health care providers and the impact of social determinants of health within their geographic service areas. The volume of available data can make it difficult for these stakeholders to determine how to best reach underserved populations [20]. Our application integrates sociodemographic and provider location datasets and uses advanced data analysis and visualization techniques to show the variation of provider supply over different geographical areas along with a summary of the most distinctive and relevant sociodemographic features that are predictive of the variation in the selected area.

3 LITERATURE SURVEY

Our team studied nineteen medical and scholarly publications to understand critical levers driving access to healthcare. Our research highlights mental health needs, the prevalence of, and reasons for, provider shortages, potential solutions [17] and the estimated costs [16].

Current State

During a 12-month period from 2005-2006, mental health issues were reported in 26.2% of the U.S. adult population [12], yet only one-third [6] of the population was treated. Primary reasons for the lack of care include a shortage of mental health providers [5,13,14], particularly in rural areas [2,9,15], and sociodemographic barriers such as age, race, and ability to pay [2,11].

The Underserved

The U.S. government defines mental health provider shortages areas as a ratio of 1 provider per 9,000 population [21]. Rural populations account for 19.3% of the total U.S. population, or 59.5 million people as of the 2010 census [19]. As such, the rural population is 4.7 times more likely to experience a mental health provider shortage than their urban counterparts [14].

Access to mental health care is also affected by sociodemographic factors. Utilization of mental health providers is lowest among adolescents, minorities and low-income populations. Over 80% of children aged 6–17 years old who were defined as needing mental health services, did not receive care [10]. In addition, characteristics including race, insurance, income and wealth affect access to care [2].

Solution

Many solutions have been proposed to increase mental healthcare access for the underserved [1,4,7,11,13,17]. Using public provider data and sociodemographic data, Team03 applied advanced data analysis and visualization techniques to derive and explain healthcare provider shortage areas for stakeholders [20] working to solve mental health provider shortages.

4 INTUITION

Proposed Methods

Our platform makes understanding provider supply, in combination with other social determinants of health, easier to research and assess. Existing provider supply evaluations tend to focus on geographical distribution alone, relate to single features like rurality [9,14,15], and/or use regressions to derive conclusions. There are a variety of tools to explore physician supply and underserved areas or to see heat maps of various census and community survey data, but they show the data in a disparate way that masks underlying correlations. We improve upon this by integrating multiple data sources to explore high-dimensional feature spaces for associations between a broad set of features. Our method looks at many measures, finding relevant patterns and makes the findings available to stakeholders leading population health efforts in their geographies.

Description of Approaches

We integrated provider locations from the National Plan and Provider Enumeration System (NPPES) with urbanicity data from the United States Department of Agriculture Economic Research Service (ERS) and sociodemographic features from the United States Census Bureau American Community Survey. We used geocoding services [22] to map the NPPES provider addresses to census geographies and individual latitude and longitude coordinates. This enables computation of provider density per area and capita for geographies defined in the census data. We stored the integrated datasets in a relational database with capacity for aggregating features such as means, differences, and ratios. The original datasets, before reduction to the features and providers of interest within the relational database, summed to several gigabytes. In total, the final database provides dozens of attributes for over 70,000 census tracts across fifty states, plus latitude and longitude coordinates for over 200,000 mental health providers. To study the vast number of samples and features, we used analytical techniques ranging from regression to supervised and unsupervised machine learning.

We used D3.js to build an interactive visualization application that caters to community needs assessment planners seeking to better understand social determinants of health associated with provide shortage areas.

The choropleth on the left side of the application allows the user to graphically select a state, and then view a choropleth heat map of the attributed selected in the drop down. The user can use the buttons at the top to toggle between county and tract, and to view data for 2016 (the most current publicly available), 2011, or view the trend from 2011 to 2016.

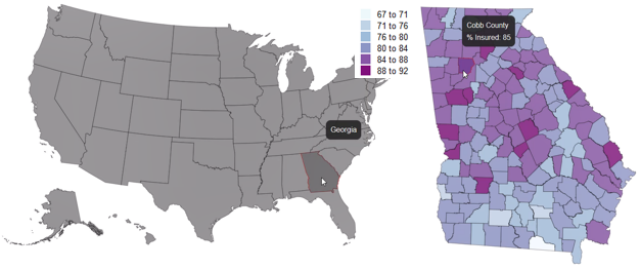

Figure 4.1: *States can be selected and then choropleth layers can be selected to display provider density or any of the socioeconomic features by county or census tract.*

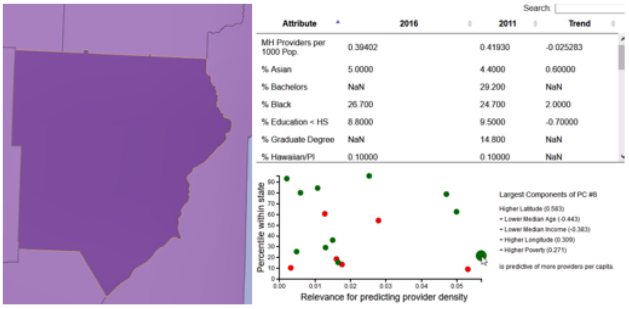

Figure 4.2: *By clicking on a county or tract, the map zooms to that geography and two new visualizations appear on the right. Top right: a table showing all of the available attributes (the same as those available in the drop down) for 2016, 2011, or the trend from 2011-2016 . Bottom right: a scatterplot ranking distinctiveness and predictive power of various combinations of socioeconomic features.*

In the lower right, the scatterplot shows the machine learning-generated interactive summary of patterns for the selected county or tract. We have transformed all the attributes using principal component analysis (PCA), and each circle within the scatterplot represents one of the principal components. PCA was performed for each state at the county and tract levels. By hovering over a circle, one can see which attributes are most correlated with that principal component. The green circles represent principal components predictive of a higher provider density, and the red circles of lower provider density. The x-axis shows how relevant each principal component is for predicting provider density. The direction of effect and relative predictive power of each principal component was determined

using the coefficients from a regularized linear model tuned using 10-folds cross-validation. The y-axis shows for the selected county/tract, the percentile ranking for each principal component relative to all the other counties/tracts in the state. The scatterplot helps the user to understand which principal components are most distinctive to the area and most predictive of provider shortages. Generally, a user will explore principal components to the far right and either near the top or bottom of the plotting area. For example, a red principal component near the top or green principal component near the bottom indicates a combination of features that may be particularly predictive of low provider density in a selected area. Figure 4.3 shows one such example where lower education, lower Asian population, lower income, and higher male population is well above the 90th percentile for the state and strongly predictive of a provider shortage.

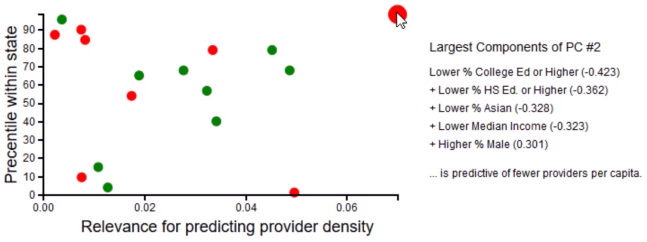

Figure 4.3: *A scatterplot shows percentile of a principal component relative to the state of the selected tract/county and relative predictive power of principal components. On mouseover, a text description displays to explain what the principal components means in terms of the original features.*

5 EXPERIMENTAL DESIGN & EVALUATION

Our experiments fell under five sequential categories:

Exploratory Data Analysis

We evaluated the distributions of all features at a national level and conditionally on sample state, county, and tract geographies. This included reviewing overlaid histograms, side-by-side boxplots, correlation in numerical statics and visual scatterplot form, and principal components analysis (PCA) to evaluate multicollinearity. The primary questions in these experiments explored the appropriate levels of geography (e.g. national, conditional on state, etc.) and combinations of features possessing interesting structures alone and in relation to provider densities.

In using NPPES data to understand provider shortage areas, we found high variance among geographically proximate census tracts. There are many residential tracts with no providers next to mixed-use tracts with high provider density. As our primary use case is for community needs assessments (which are typically done at the county level), we used provider density at the county level for our predictive modeling. We also found that county-level provider density follows an approximately exponential distribution, so we decided to log-transform the provider density for the later predictive models.

The sociodemographic census data (ACS) contained an extensive variety of attributes, with many permutations of age, race, income and health insurance. We selected a subset that we felt would be most relevant to community needs planners while trying to maintain selections from all attribute categories. Correlation matrices of these attributes helped us identify collinearity and potentially redundant features such as poverty rate and median income.

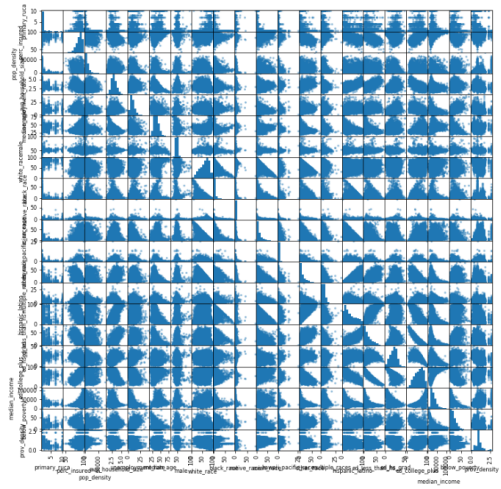

Figure 5.1: *Scatterplot review was of some benefit, but the figure above indicates how quickly the exponential growth of potential feature interactions becomes less interpretable. This motivated follow-on feature selection and feature transformation experiments.*

Feature Selection

To understand patterns associated with provider density, we performed experiments with random forest models, multiple linear regression with lasso regularization, multiple linear regression with ridge regularization, and neural networks to determine which subset of features possessed the greatest predictive power of provider density. While predictive power may only indicate correlation, this step built on our exploratory data analysis to refine our understanding of the relevant feature space. We found that different models provided very different rankings of feature importance and determined that the attempted methods could not reliability determine which of the many interacting and correlated features had predictive power. For example, below poverty and low median income might be predictive of fewer providers, but a model fitted with both might arbitrarily select only one as important due to the significant mutual information. The conclusion that we could not reliably estimate predictive power of the original features in the presence of multicollinearity motivated the next set of feature transformation experiments.

Feature transformation

To address the issue of multicollinearity, we applied Principal Components Analysis (PCA). Figure 5.2 shows the distribution of variance explained per principal component. The green bars indicate variance explained from the actual dataset, while the orange bards indicate the expected result on a similar, but perfectly uncorrelated dataset. The significant gap between the two curves confirms that significant multicollinearity is present in the socioeconomic data.

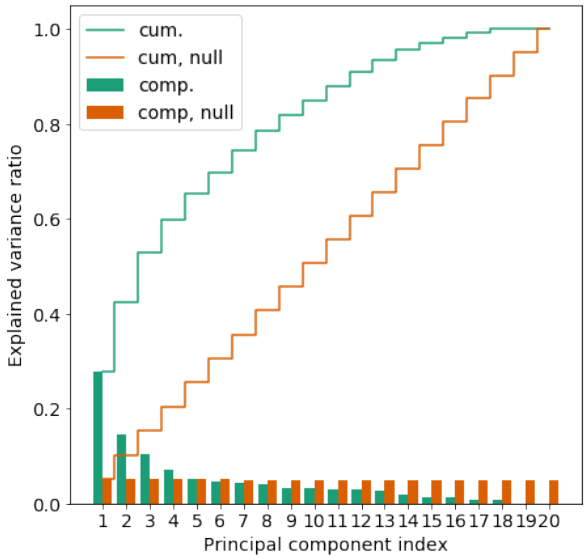

***Figure 5.2**: Principal component gap analysis shows that PCA explains significantly more variance per early feature than would be expected in an uncorrelated dataset. While all principal components were retained for later predictive modeling, this analysis still confirms the effectiveness of PCA in combining the correlated features together.*

Fortunately, PCA also provided a means of addressing multicollinearity. By changing the basis of the dataset to the principal components, a transformed dataset is produced with perfect orthogonality between features. Figure 5.3 provides a visualization of the data for all U.S. tracts projected onto the first two principal components. Projections of some of the features are also shown. As can be seen, some of the features are clearly highly correlated, such as college education and higher median income. The principal components form features that are composites of components of the original features and effectively merges the redundant components into single features in PCA space. Therefore, PCA transformed the original features into new features that could then be reliability assessed for independent predictive power using various predictive models. It can also be seen that the first two principal components do begin to differentiate the shortage and non-shortage areas, but there is still significant overlap in the projection.

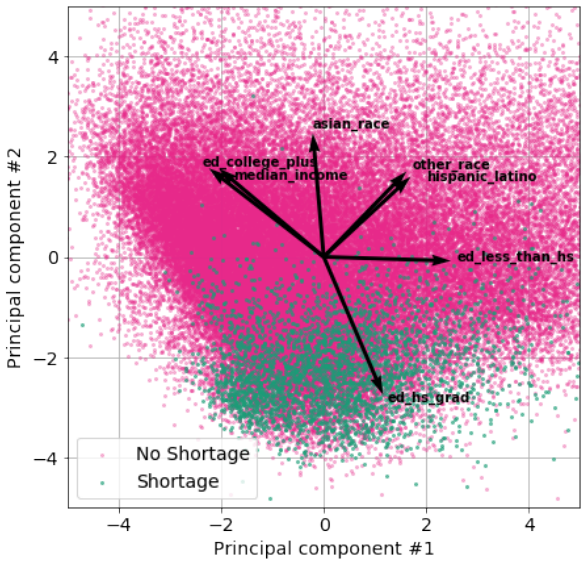

***Figure 5.3**: Projecting the data and some feature axes onto the first two principal components illustrates how the correlated components of features can be merged together in PCA space.*

We also performed independent components analysis to assess whether out dataset could be viewed as a mixture of underlying latent features. We did not find any significant results relative to what would be expected on a random dataset, indicating that such features did not exist.

Clustering

Insights into the structure of the data can be identified by clustering, particularly if there are natural clusters that fit the models being applied. We applied k-means and expectation maximization with gaussian mixtures of varying covariance to the dataset. Clustering was evaluated using standard measures of significance such as elbow plots of variance explanation and gap statistics. Our hypothesis was that we might find some natural sub-groups with varying provider density characteristics. If these natural sub-groups existed, we intended to use them in the end visualization to explain some of the structure between the provider supply and socioeconomic features. Unfortunately, the evaluation metrics indicated that no significant natural clusters existed at a county or tract level relative to what would be seen by clustering uniformly randomly distributed noise (i.e. a "null" dataset for clustering).

Prediction

The conclusions from the analyses discussed above were that PCA was useful for addressing multicollinearity and that a transformed dataset should be used for fitting various regression models that could then reliability rank the principal components in terms of predictive power. We explored the capability of random forest regressors, neural network

regressors, and various regularized linear models to predict provider densities by fitting each model to all counties and tracts within each of a few test states. All model hyperparameters were tuned to optimal values using 10-folds cross-validation grid searches. We determined that the regularized linear models, particularly ridge regression, provided the minimum cross-validation mean square error and maximum error reduction relative to a baseline "mean only" intercept model. While random forest models and neural networks are powerful function approximators, our analysis concluded that there simply is not enough data within each state to train non-linear models. Figure 5.4 provides a summary of baseline mean error as well as train and cross-validation error for random forest and three regularized linear models. The random forest model is overfitting, and it performs about the same as assuming the mean response as a constant function.

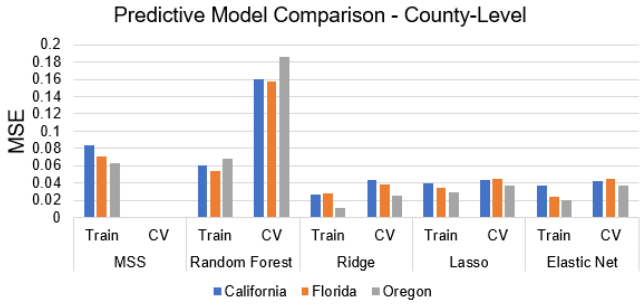

Figure 5.4: *Regularized linear models, specifically ridge regression, outperformed random forest due to the relatively small sample sizes of tracts/counties within a single state.*

We then set up an analysis pipeline to automatically fit the best ridge regression model to the county and tract data within each state and store the resulting predictive power ranks for each principal component using the coefficients of the fitted models. The signs of the coefficients were used to determine the direction of the predictive effect on provider density and the magnitudes were used to determine the relative amount of predictive power.

While PCA resolved the issues with reliability ranking features, it came at the expense of less interpretable features. This is because each principal component is a combination of the original features. We explored various visualizations such as heatmaps, shown in Figure 5.5, but ultimately decided on the scatterplot with pop-up text description of each feature as shown in Figure 4.3. The heatmap relays good information, but it is too much to be easily interpreted by a lay person exploring patterns between provider density and socioeconomic features.

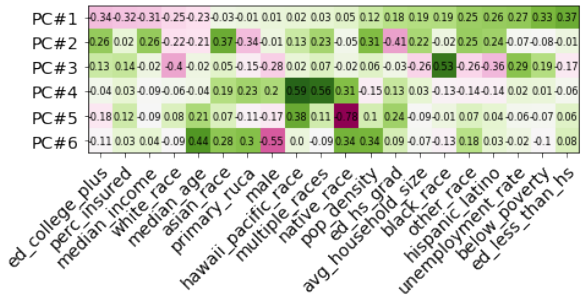

Figure 5.5: *Heatmaps conveyed useful information about the composition of principal components but are not intuitive for a lay person using the application to explore associations.*

6 CONCLUSION & DISCUSSION

We present an interactive application that empowers healthcare stakeholders in community needs assessment and planning. By better understanding the social determinants of health in their geographies and where providers are needed the most, stakeholders can prioritize interventions to address their underserved areas and resolve behavioral health provider shortages. Our platform makes it easier for them to focus on the metrics that matter the most, so they can spend less time sifting through data and more time focused on solutions for their specific geographies.

Our feature selection, feature transformation, and predictive modeling techniques will support further data integrations and the infrastructure is designed to be extensible to accommodate additional features and other provider specialties. In the future we hope to include claims data from Medicare, Medicaid, and commercial insurance organizations to better enable healthcare stakeholders in determining optimal health coverage for their communities.

7 STATEMENT OF TEAM EFFORT

All team members have contributed similar amount of effort.

REFERENCES

1. Alegría, M., Canino, G., Ríos, R., Vera, M., Calderón, J., Rusch, D., & Ortega, A. (2002). Inequalities in Use Of Specialty Mental Health Services Among Latinos, African Americans, And Non-Latino Whites. *Psychiatric Services, 53*(12), 1547-1555.

2. Arcury, T. A., Gesler, W. M., Preisser, J. S., Sherman, J., Spencer, J., & Perin, J. (2005). The Effects of Geography And Spatial Behavior On Health Care Utilization Among The Residents Of A Rural Region. *Health Services Research, 40*(1), 135–156.

3. Beardsley, K., Wish, E. D., Fitzelle, D. B., Ogrady, K., & Arria, A. M. (2003). Distance Traveled to Outpatient Drug Treatment and Client Retention. *Journal of Substance Abuse Treatment,25*(4), 279-285. doi:10.1016/s0740-5472(03)00188-0

4. Bindman, A. (2013). Using the National Provider Identifier for Health Care Workforce Evaluation. *Medicare & Medicaid Research Review,3*(3). doi:10.5600/mmrr.003.03.b03

5. Carrie B. Oser Ph.D., Elizabeth P. Biebel Ph.D., Erin Pullen M.A. & Kathi L. H. Harp M.A. (2013) Causes, Consequences, and Prevention of Burnout Among Substance Abuse Treatment Counselors: A Rural Versus Urban Comparison*, Journal of Psychoactive Drugs, 45*:1, 17-27, DOI: 10.1080/02791072.2013.763558

6. Cunningham, P. J. (2009). Beyond Parity: Primary Care Physicians' Perspectives on Access to Mental Health Care. *Health Affairs,28*(3), 490-501. doi:10.1377/hlthaff.28.3.w490

7. Donabedian, A. (1966). Evaluating the Quality of Medical Care. *The Milbank Memorial Fund Quarterly, 44*(3), 166-206. doi:10.2307/3348969

8. Garfield, R. L., Zuvekas, S. H., Lave, J. R., & Donohue, J. M. (2011). The Impact of National Health Care Reform on Adults with Severe Mental Disorders. *American Journal of Psychiatry, 168*(5), 486-494. doi:10.1176/appi.ajp.2010.10060792

9. Kaplan, L., Skillman, S. M., Fordyce, M. A., Mcmenamin, P. D., & Doescher, M. P. (2012). Understanding APRN Distribution in the United States Using NPI Data. *The Journal for Nurse Practitioners,8*(8), 626-635. doi:10.1016/j.nurpra.2012.05.022

10. Kataoka, S.H., Zhang, L.& Wells, K.B (2002) Unmet Need for Mental Health Care Among U.S. Children: Variation by Ethnicity and Insurance Status. American Journal of Psychiatry 2002 159:9, 1548-1555. https://www.ncbi.nlm.nih.gov/pubmed/12202276

11. Kathol, R. G., MD, Butler, M., PhD, Mcalpine, D. D., PhD, & Kane, R. L., MD. (2010). Barriers to Physical and Mental Condition Integrated Service Delivery. *Psychosomatic Medicine,72*(6), 511-518. doi:10.1097/psy.0b013e3181e2c4a0

12. Kessler RC, Chiu WT, Demler O, Walters EE.(2005) Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. Archives of General Psychiatry. 62(6):617–627. www.ncbi.nlm.nih.gov/books/NBK32788/

13. McAlpine, D. D., & Mechanic, D. (2000). Utilization Of Specialty Mental Health Care Among Persons With Severe Mental Illness: The Roles Of Demographics, Need, Insurance, And Risk. *Health Services Research*, *35*(1 Pt 2), 277–292. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1089101/

14. Merwin, E., Hinton, I., Dembling, B., & Stern, S. (2003). Shortages of rural mental health professionals. *Archives of Psychiatric Nursing,17*(1), 42-51. doi:10.1053/apnu.2003.1

15. Miller, Benjamin F., et al. Primary Care, Behavioral Health, Provider Colocation, and Rurality. *The Journal of the American Board of Family Medicine 27*.3 (2014): 367-374. doi:10.3122/jabfm.2014.03.130260

16. OConnell, M. E., Boat, T. F., & Warner, K. E. (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, D.C.: National Academies Press.

17. Patton, G. C., Sawyer, S. M., Santelli, J. S., Ross, D. A., Afifi, R., Allen, N. B., . . . Viner, R. M. (2016). Our future: A Lancet commission on adolescent health and wellbeing. *The Lancet,387*(10036), 2423-2478. doi:10.1016/s0140-6736(16)00579-1

18. Rana, Y., Pfrommer, K., Adamson, D. M., Brown, R. A., & Miyashiro, L. (2015). *Access to Behavioral Health Care for Geographically Remote Service Members and Dependents in the U.S*. Rand Corporation. https://www.rand.org/pubs/research_reports/RR578.html

19. Health Resources and Services Administration (2017). Defining Rural Population. https://www.hrsa.gov/rural-health/about-us/definition/index.html

20. Community Health Needs Assessment PeaceHealth (2016). Sacred Heart Medical Center. https://www.peacehealth.org/about-peacehealth/Pages/community-health-needs-assessment

21. Designated Health Professional Shortage Areas Statistics. Bureau of Health Workforce Health Resources and Services Administration (HRSA) U.S. Department of Health & Human Services. https://ersrs.hrsa.gov/ReportServer?/HGDW_Reports/BCD_HPSA/BCD_HPSA_SCR50_Qtr_Smry_HTML&rc:Toolbar=false

22. Texas A&M Geoservices. Texas A&M University 2013. http://geoservices.tamu.edu/