Group 5 Project Proposal

Basic Information

Project Title

Social Determinants of Health

Team Members

Name	Email	UID
Jonathan Himes	jonathan.himes@utah.edu	U1367169
Michelle Kubicki	michelle.kubicki@utah.edu	U1370752
Yueqin Yang	yueqin.yang@hsc.utah.edu	U0737683
Jackie Bearnson	u0763123@umail.utah.edu	U0763123
Lucia Ranallo	U1427138@umail.utah.edu	U1427138

Project Repository

https://github.com/Jonathan-Himes/SDOH-GroupProject.git

Background and Motivation

The World Health Organization (WHO) defines social determinants of health as nonmedical factors that influence health outcomes, including the conditions in which people live and the broader set of influences creating conditions of daily life. Examples include economic policies and systems, political systems, development, societal norms and policies, racism, and climate change. Knowledge of these factors and their incorporation into research, community, and medical practices could lead to improvements in health outcomes, as previous research has indicated that social determinants of health are associated with differences in health risks[1], life expectancy [2], and disease outcomes [3].

It is common that social determinants of health are not readily available on an individual level in electronic health records, making them difficult to incorporate. This information is, however, collected on a large scale via publicly available tools such as censuses. The ability to identify these factors and their downstream effects could play an important role in improving health and reducing health disparities in disadvantaged populations by incorporating them in methods like policy change and community outreach.

Project Objectives

In order to successfully incorporate data associated with social determinants of health into change-generating practices in research, healthcare, community work, and policy, it should first be extracted from available resources such as censuses for specific geographical regions. Data extraction from available sources should be followed by data cleanup as well as quality assessment, including spatial and temporal quality. Lastly, the data will be assessed for application in health research in order to determine how to best incorporate it into future practices that will be beneficial to disadvantaged populations.

Data extraction will be preceded by identification of variables of interest based on previous research identifying factors impacting health outcomes. It will be essential to identify sources from which appropriate data can be extracted. Next, extracted data will be assessed for quality and if needed, cleaned and transformed to make it suitable for analysis. This may include identification and treatment of missing data and outliers. Data analysis will identify social determinants of health that correlate most with health outcomes of interest, and visualizations will present this data in a digestible format. This project aims to generate social determinant of health data from census data for the state of Utah that can be suitable for use in future practices, such as further research aims and implementation of outcomes with the goal of alleviating the public health disadvantages present in communities today.

Data

Most of our data comes from the paper “FACETS: using open data to measure community social determinants of health” by Michael N Cantor, Rajan Chandras, and Claudia Pulgarin This paper can be found at the following link:

https://academic.oup.com/jamia/article/25/4/419/4569610

The authors of the paper compiled a dataset of social determinants of health which is available at the following link.

https://github.com/mcantor2/FACETS

The design section of this proposal includes a table. Listed next to each item is a source of where the authors of the paper sourced their data. The authors provide the following links to where they sourced the data.

- The majority of the data was compiled from the 2015 American Community Survey. As of now, this data has not been downloaded and a primary task of this project is going to be tracking down this data. The link sited in the paper is not functional, but Group 5 feels confident they can track this data down through US Census website or find this data on a data sharing site such as Kaggle.
- https://www.epa.gov/national-air-toxics-assessment/2011-nata-assessment-results
- https://www.ers.usda.gov/data-products/food-access-research-atlas.aspx
- Wen M , Zhang X, Harris CD, Holt JB, Croft JB. Spatial disparities in the distribution of parks and green spaces in the USA. Ann Behav Med.2013;45 (Suppl 1):S18–27
- https://beh.columbia.edu/
- https://health.data.ny.gov/Health/Active-Retail-Tobacco-Vendors/9ma3-vsuk
- https://data.cityofnewyork.us/Public-Safety/NYPD-7-Major-Felony-Incidents/hyij-8hr7
- http://coredata.nyc/
- www.census.gov/topics/income-poverty/income-inequality.html
- http://svi.cdc.gov/

Data Processing

As mentioned above, tracking down the data is going to be a substantial task. Many of the datasets, specifically those coming from the US Census Bureau, are very large in size and may require the use of the CHPC.

This project is mostly a data harmonization project. We will be matching the required dimensions to the correct geographical region.

Some of the data is in the form of .accdb and will be transformed into a more usable .csv format.

Every source dataset will be truncated to only include Utah data before any harmonization will take place.

Missing or null data values will be checked for to ensure quality.

Design

With the goal of facilitating easy access to conglomerated data on social determinants of health to groups such as clinics and government agencies, visualizations of the data should provide understanding of the datasets at a glance. As we work on processing and analyzing the data, new insights will allow us to create graphs and other representations to communicate our findings to public health stakeholders and allow them to make inferences about affected populations.

There will be a focus on the geographic and temporal qualities of the data, such as demographics of defined regions and dates of data collection. Special care will be provided when combining disparate data sets along these axes to maintain data quality. Ultimately the goal is to have an aggregated data set similar to the FACETS project applied to Utah.

Below is a table copied from the metadata tab of the data set we are emulating. We plan to produce a table that has data for each of these items for every census tract in the state of Utah. The first column will include all the census tracts for the state of Utah. The items listed in the table below will each be a column and contain the appropriate data.

Item	**Description**	**Source**
FIPS	US Census 11 digit FIPS code for tract	US Census
Urban	Urban/Rural Flag	USDA Food Access Research Atlas
Total_population	Total Population	ACS 2015 Estimates
P_WH	% White	ACS 2015 Estimates
P_AA	% African-American	ACS 2015 Estimates
P_AI	% American Indian	ACS 2015 Estimates
P_AS	% Asian	ACS 2015 Estimates
P_NH	% Native Hawaiian/Pacific Islander	ACS 2015 Estimates
P_OR	% Other Race	ACS 2015 Estimates
P_2R	% 2 or more Races	ACS 2015 Estimates

P_Latino	% Latino/Hispanic Ethnicity	ACS 2015 Estimates
P_native	% Native Born in US	ACS 2015 Estimates
P_FB	% Foreign Born	ACS 2015 Estimates
P_citizen	% US Citizen	ACS 2015 Estimates
P_non-citizen	% Non-citizen	ACS 2015 Estimates
P_NoSchool	% No schooling	ACS 2015 Estimates
P_HS_no_degree	% completed high school, no degree	ACS 2015 Estimates
P_HS_or_GED	% High school or GED degree	ACS 2015 Estimates
P_some_college	% Some college, no degree	ACS 2015 Estimates
P_college_degree	% AA or BA	ACS 2015 Estimates
P_Masters_prof_doc	% Masters, professional, doctorate	ACS 2015 Estimates
P_Other	% Other level of schooling (< High school)	ACS 2015 Estimates
P_limited_eng	% Limited English proficiency	ACS 2015 Estimates
Poverty_rate	% in poverty	ACS 2015 Estimates
MED_HH_income	Median Household Income	ACS 2015 Estimates
UE_rate	Unemployment rate	ACS 2015 Estimates
P_UI	% Uninsured (health)	ACS 2015 Estimates
P_Insured	% Insured	ACS 2015 Estimates
P_UI_under_18	% Uninsured, <18 yo	ACS 2015 Estimates
P_UI_18-64	% Uninsured, age 18-64	ACS 2015 Estimates
P_UI_65_over	% Uninsured, over 65	ACS 2015 Estimates
P_any_private_ins	% Any private insurance, all ages	ACS 2015 Estimates
P_any_public_ins	% Any public insurance, all ages	ACS 2015 Estimates
P_Medicare_alone	% Medicare only	ACS 2015 Estimates
P_Medicaid_alone	% Medicaid only	ACS 2015 Estimates
Resp_HI	Respiratory Hazad Index	EPA National Air Toxics Assessment, 2011
Low_access	Low access to healthy food (1/2 mile)	USDA Food Access Research Atlas
Park_distance	Population-weighted distance to closest 7 parks	CDC
Walkscore	Neighborhood walkability scale	Rundle- Columbia BEH
Walkscore_percentile	Percentile of walkabiliity (higher better)	Rundle- Columbia BEH
Tob_retailer_per_1000	Tobacco retailers/1000 population	NYS open data- active tobacco retailers
Crime_per_1000	# of 7 serious crimes /1000 population	NYC open data nypd all felony incidents geocoded
GINI	GINI inequality index	US Census

SVI_themes_total	Social Vulnerability index- total themes percentile- higher is more vulnerable	CDC SVI
SVI_flags	Social Vulnerability index- score for flags	CDC SVI
Housing_violations_per_1000	Housing violations per 1000 rental units	Furman center
Voter Turnout	Voter Turnout	BOE (by state assembly district 2014)
Turnout quartile	Turnout Quartile	BOE calculated

An alternative to this design would be to list the data on a county level.

Must-Have Features

1. A clear and concise description of the SDOH variables used in the analysis.

2. SDOH variables for every census tract in the state of Utah

3. A cleaned and transformed dataset that is suitable for analysis.

4. Identification and treatment of missing data and outliers.

5. Exploratory Data Analysis of the Data to understand distributions and skewness

Optional Features

1.Use of machine learning techniques to predict health outcomes based on SDOH variables.

2. Incorporation of additional data sources, such as electronic health records or data from public health agencies.

3. Comparison of health outcomes and SDOH variables across different geographic areas, such as states or counties.

4. Use of interactive visualizations that allow users to explore the data and findings.

5. Collaboration with community-based organizations or other stakeholders to identify additional variables or insights to include in the analysis.

6. Literature Review to understand the wide-scale effects of different SDOH factors and if certain factors are targeted towards certain populations

Project Schedule

Week 1 - Ending 2/18:

Notable Deadlines:

- Project Proposal due 2/16
- Self-assessment due 2/16

Tasks:

- Come up with a direction the project takes on. (Everyone)
- Create a GitHub repository (Jonathan)
- Complete the project proposal on the teams (everyone)
- Upload completed proposal to GitHub repository (Jonathan)
- Complete self-assessment on Canvas (everyone)

Week 2 – Ending 2/25:

Tasks:

- Download all required data and store on GitHub repository (Everyone)
- Identify and gather additional data sources, if necessary (Everyone)
- Define the scope and objectives of the project.

Week 3 – Ending 3/4:

Notable Deadlines:

- Meeting with Instructors and Feedback due 3/2

Tasks:

- Adjust the scope and objectives of the project if necessary.
- Establish a plan for data cleaning and transformation.

Week 4 – Ending 3/11:

Tasks:

- Clean and transform the SDOH data to a format suitable for analysis.
- Identify and address missing data and outliers.
- Determine all the geographic codes that will be used for Utah data

Week 5 – Ending 3/18:

Notable Deadlines:

- Project Update Submission due 3/16
- Self-assessment due 3/16

Tasks:

- Complete and upload the project update (Everyone)
- Complete self-assessment on Canvas (everyone)

- Perform statistical analysis to identify correlations between SDOH and health outcomes.
- Perform EDA to understand distributions of variables and how they could affect statistical calculations

Week 6 – Ending 3/25:

Notable Deadlines:

- Intermediate work presentation due 3/22
- Self-assessment due 3/16

Tasks:

- Develop visualizations to communicate the findings.
- Determine if building a model is achievable in the time frame
- Work to determine if certain factors are more influential than others

Week 7, Week 8: – Ending 4/8:

Notable Deadlines:

- Meet with instructors for feedback 4/6

Tasks:

- Refine the analysis and visualizations based on feedback and input from instructors
- Create any additional aspects (charts, calculations, models) that are recommended

Week 9, Week10 – Ending 4/28:

Tasks:

- Prepare the final report and presentation of the findings.
- Review the report and presentation with Instructors for feedback.

Week 11 – Ending 5/6:

Notable Deadlines:

- Final Project Due 5/4
- Self-assessment due 5/4

Tasks:

- Make final adjustments to the report and presentation as necessary.
- Present the final report and findings

References

1. Bieler, G., Paroz, S., Faouzi, M., Trueb, L., Vaucher, P., Althaus, F., Daeppen, J.-B. and Bodenmann, P. (2012), Social and Medical Vulnerability Factors of Emergency Department Frequent Users in a Universal Health Insurance System. Academic Emergency Medicine, 19: 63-68. https://doi.org/10.1111/j.1553-2712.2011.01246.x
2. Chetty R, Stepner M, Abraham S, et al. The Association Between Income and Life Expectancy in the United States, 2001-2014. JAMA. 2016;315(16):1750–1766. doi:10.1001/jama.2016.4226
3. Walker, R.J., Gebregziabher, M., Martin-Harris, B. et al. Relationship between social determinants of health and processes and outcomes in adults with type 2 diabetes: validation of a conceptual framework. BMC Endocr Disord 14, 82 (2014). https://doi.org/10.1186/1472-6823-14-82