BMI-6016 project proposal

Spring 2024, Group 4

Basic Info

Project title: Social determinants of health in Utah

Participants:

- Logan Correa (u1094034, logan.correa@utah.edu)
- Joshua Choi (u1492647, joshua.choi@utah.edu)
- Lauren Cutler (u6037420, lauren.cutler@hsc.utah.edu)
- Ashok Vengala (u1419414, u1419414@utah.edu)

GitHub repository: https://github.com/js-choi/sdh-data-wrangling

Background and motivation

Health is determined not only by genetic factors or encounters with the healthcare system but also by social factors from their family, culture, daily environment, and community. These social determinants of health (SDH) have been robustly shown to have significant effects on multiple health outcomes. Thus, SDH have become prominent in health predictive models, risk stratification, and value-based care as data that give more realistic descriptions of patient populations. These are especially needed for the state of Utah. Utah faces numerous long-existing health disparities between different economic, racial, and ethnic groups, leading to the One Utah Summit Roadmap released by Governor Spencer Cox to prioritize health equity.

Although SDH data are plentiful in various open datasets, these often require complex querying and extensive processing to create data usable by research, particularly for Utah-specific research. There therefore is a need to create a granular database of community-level SDH in Utah.

Project objectives

We will investigate whether it is possible, using open datasets, to create a useful and high-quality database of community-level SDH data for the Salt Lake City metropolitan area. We will use open datasets due to their low cost for routine research, the ability to assess their data quality freely, and the ability to mine their data programmatically at scale for predictive models or other population-level datasets. We anticipate that the database would be useful to health systems and researchers evaluating SDH in this important geographic area.

Data

We will utilize openly available data from the following government agency.

- United States Environmental Protection Agency 2019 AirToxScreen Assessment
    - https://www.epa.gov/AirToxScreen/2019-airtoxscreen-assessment-results#state
- USDA Food Access Research Atlas 2019
    - https://www.ers.usda.gov/data-products/food-access-research-atlas/
- Social Vulnerability Index 2020
    - https://data.cdc.gov/Health-Statistics/CDC-Social-Vulnerability-Index-SVI-/u6k2-rtt3/about_data
- American Community Survey
    - https://data.census.gov/

Each of these sources provides the data in a csv file broken down by census tract and or county. Combining this data we will be able to look at demographics information and social determinants of health data.

Data processing

According to “Facets: using open data to measure community social determinants of health" all of these data sources are “analysis-ready” so we do not expect substantial data cleanup. We plan to derive similar data quantities as the Facets paper including, demographics data race, ethnicity, income, level of education, and insurance status. The specific social determinants of health quantities we plan to derive include air quality hazard index, access to healthy food, and social vulnerability index. The data will be processed using python.

Design

For demographic data and other numerical indicators (like income levels, educational attainment, etc.), bar charts and histograms will be used to communicate the distribution of these factors within the population. If enough historical data is available, line graphs showing trends over time for certain SDH could provide insights into whether disparities are widening or narrowing. A comprehensive view of SDH factors will be shown using radar charts, which will allow for a compact, comparative view of multiple health determinants simultaneously, facilitating quick assessments of community health profiles.

An alternate design for data visualization will include SDH maps. If enough geospatial data is available, SDH maps will be created to show geographical disparities and how determinants are distributed across the region of interest. Maps provide an intuitive spatial understanding of how social determinants of health vary across different parts of the metropolitan area, highlighting areas of need.

Must-have features

- The project must integrate datasets from multiple sources including the United States Environmental Protection Agency, USDA, CDC, and American Community Survey into a unified database.
- Utilization of Python programming for processing and cleaning the integrated datasets to derive relevant demographic and social determinant of health (SDH) quantities, such as demographics data (race, ethnicity, income, level of education, insurance status), air quality hazard index, access to healthy food, and social vulnerability index.
- A thorough data quality assessment process must be conducted to ensure the integrity, accuracy, and completeness of the processed datasets.
- Implementation of various visualization techniques including but not limited to bar charts, histograms, line graphs, radar charts, and maps to effectively communicate the distribution of demographic factors, trends over time for certain SDH, and geographical disparities in SDH across the Salt Lake City metropolitan area.

Optional features

- Implementation of interactive data visualization tools or dashboards can greatly enhance user engagement and understanding of the SDH data. Users can explore the data dynamically, filter information based on their interests, and gain deeper insights by interacting with various visualization elements. This feature provides a more intuitive and personalized experience for stakeholders, facilitating better decision-making and action planning.
- Integration of advanced analytics techniques such as predictive modeling or clustering analysis can uncover hidden patterns, correlations, and insights within the SDH data. By leveraging machine learning algorithms, the project can identify complex relationships and make predictions about future health outcomes or intervention effectiveness. This feature enables more sophisticated analysis and decision support, empowering stakeholders with actionable insights to address health disparities effectively.

Project schedule

- Role A (Logan Correa, Lauren Cutler): Look at datasets' descriptions, perform data quality assessment, assess temporal/spatial granularity, arranging the presentation
- Role B (Joshua Choi, Ashok Vengala): Download data, store data in CHCP, processing the data, arranging the presentation

Monday, February 12	Submit project proposal
Tuesday, February 20	Team meeting Role A: Locate data sources. Role B: Set up storage, set up presentation.

Monday, February 12 - Monday, February 19	Meet with instructor
Tuesday, February 27	Team meeting Role A: Start data-quality assessment. Role B: Download data, continue presentation work.
Tuesday, March 4	Team meeting Project update submission
Tuesday, March 11	Team meeting
Monday, March 18	Intermediate work presentation
Tuesday, March 19	Team meeting
Tuesday, March 25	Team meeting
Monday, April 1 - Monday, April 8	Meet with instructor
Tuesday, April 2	Team meeting
Monday, April 8	Intermediate work presentation
Tuesday, April 9	Team meeting
Tuesday, April 16	Team meeting
Tuesday, April 23	Team meeting
Monday, April 29	Final project presentation
Monday, May 6th	Final project submission

References

Cantor MN, Chandras R, Pulgarin C. FACETS: using open data to measure community social determinants of health. *J Am Med Inform Assoc*. 2018;25(4):419-422. doi:10.1093/jamia/ocx117