COMMUNITY COMPANION CHALLENGE

TRANSFORMING CARE WITH DATA-DRIVEN COMPASSION

Team Code:
DA48

Table of Contents

Abstract

In our quest to understand health, we must recognize the impact of where we live, our income, education, race, age, and support systems. These social determinants of health (SDOH) are hidden keys that can unlock or hinder our well-being. They play a significant role in creating health inequalities, so-called risk factors that can affect individuals and entire communities.

This report underscores the classification of an individual's risk factors using one or more of these SDOH variables through modern tools and techniques using Artificial Intelligence, particularly Large Language Models (LLMs). The solution for classifying the risk factors as explored in this report is superior to the current traditional methods, which usually rely on surveys or self-reports, which take up time and are prone to human error. This report also aims to build upon the solution of identifying these risk factors by providing recommendations to an individual based on the individual's risk factors to mitigate the risks associated with him/her.

1 Introduction

In the evolving landscape of Large Language Models (LLMs), we wish to harness their reasoning power to create a fresh perspective on understanding social factors that impact an individual's health and other social factors. Instead of surveying individuals, we believe in a proactive approach. We aim to utilize the advanced language comprehension abilities of LLM models to build a proactive model to aid healthcare staff but also to enable direct support and guidance for individuals.

Current approaches to dealing with SDOH variables include manual screening by PNs and physicians, who look for flags and potential risks by analyzing patient inputs. However, these methods could be more effective, bringing in feasibility issues, evaluation limitations, and restricted capabilities.

To address the challenges, we propose a predictive framework that operates by combining statistical and qualitative semantics from social determinants of health (SDOH) variables and census tract data to generate a preliminary risk profile for various social determinants in the form of a scorecard for an individual profile to flag potential social and health risks by leveraging the comprehension and logical reasoning power of LLMs. Utilizing these predictive scorecards, the framework identifies potential risk domains for the individual and suggests local resources from a diverse data source. In the next phase, we optimize and adapt our framework for India's unique healthcare challenges, analyzing the distinct risks and hurdles we face while implementing our framework and working on SDOH variables in the Indian context.

Section 2 of this report presents a thorough literature review to better understand the landscape of working with LLMs and an in-depth analysis of the SDOH variable. Section 3 presents exploratory data analysis of the dataset and correlation analysis between SDOH variables across different domains of interest. In section 4, we analyze the dataset and the key insights we gained. Section 5 details the methodology and components of our proposed framework, and Section 6 examines the performance and validation of our results. Section 7 presents our approach to optimizing our framework for the Indian context and analyzing the challenges we face in this regard. Lastly, through our frameworks, we aim to provide a tool that leverages the power of machine learning to proactively address individuals and healthcare staff to provide them with guidance and support, particularly in resource-constrained environments.

Risks:

Economic Stability	Striking a balance for economic stability means juggling income, expenses, debts, medical costs, and support, all while steering clear of potential pitfalls.
Neighborhood and Physical Environment	The dynamics of one's surroundings demand careful consideration due to the potential challenges associated with housing, transportation, etc.
Education	Education presents opportunities, but only with its accompanying challenges, including literacy rate and access to educational institutes.
Food	Challenges due to certain uncertainties can impact hunger and access to healthy food.
Community and Social Context	It involves acknowledging and managing potential challenges ensuring all participants' inclusivity, understanding, and well-being.
Health Care System	The healthcare system can be uncertain, with challenges concerning coverage, provider availability, and the quality of care.

Literature Review

2.1 Synthea

Getting access to actual patient data for research can be challenging because of rules about privacy and limited access to diverse information. However, Synthea[1] solves this problem by making up fake patient information that looks real. It leverages statistical models, medical knowledge, and population health data to generate this data. It begins by simulating a diverse population of virtual patients, considering factors such as age, gender, race, and socioeconomic status to ensure representation across demographics. Then, it generates comprehensive medical histories for each virtual patient, incorporating past diagnoses, treatments, medications, surgeries, and healthcare encounters.

Researchers can then use this synthetic data to try different healthcare situations without worrying about breaking privacy rules. It's like playing with a model of the healthcare system. Because Synthea is open-source, anyone can use and improve it, and researchers can work together to improve it. This helps in studying different healthcare topics, like how policies affect patients or how to design better healthcare systems. Overall, Synthea is a helpful tool that makes it easier to study healthcare and find ways to improve patient care.

2.2 RoBERTa

RoBERTa**[2]**, short for "A Robustly Optimized BERT Pretraining Approach," is a powerful language model developed by Facebook AI in 2019. It's an improved version of BERT (Bidirectional Encoder Representations from Transformers) that tackles some limitations and achieves top-notch performance in different Natural Language Processing (NLP) tasks.
RoBERTa builds on BERT's**[3]** language masking strategy using **Dynamic Masking**, a method where the system learns to predict hidden text parts within unmarked language examples. RoBERTa uses bytes instead of Unicode characters and boosts the vocabulary size to 50K without any preprocessing or input tokenization, in contrast to BERT, which uses subword-level tokenization with a vocabulary size of 30K after preprocessing.

2.3 Generative Models

Generative models are natural language processing (NLP) models that generate text or sequences of words that mimic human language. These models capture the underlying structure and language patterns to produce coherent and contextually relevant text.

GPT[4] (Generative Pre-trained Transformer), a popular generative model, helps computers understand and write human-like language better by utilizing transformer architecture, which enables it to capture long-range dependencies and contextual information effectively. During training, GPT processes text sequences by attending to different parts of the input and generating predictions for the next word in the sequence based on the preceding context. Through numerous iterations of training on diverse text corpora, GPT learns to understand the nuances of language, including grammar, semantics, and pragmatics.

2.4 Prompting:

In Large Language Models (LLMs), prompting is a technique that provides relevant context to guide the model and generate desired outputs. It involves providing the model with specific instructions, questions, or context to guide the model in generating text that aligns with the objective. This flexibility makes LLMs highly versatile and applicable across various applications. The paper introduces Synthetic prompting[5], a method for improving large language models' reasoning abilities by generating examples through a backward and forward process, outperforming existing prompting techniques in numerical, symbolic, and algorithmic reasoning tasks.

2.5 Impact of Social Determinants on Health Prediction

A large, compelling body of evidence shows that a wide range of SDOH strongly correlates with health outcomes. A particular study found that 40% of deaths in the United States are caused by behavior patterns that can be modified by preventive interventions and suggested that only 10-15% of preventable mortality could be avoided by higher-quality medical care **[5].** Results indicate that additional measures of socioeconomic characteristics are required to predict better physical well-being, particularly among vulnerable groups, such as veterans**[6]**. It is increasingly recognized that to improve population health, health equity needs to become a priority in the health sector, and measures to reduce disparities must be integrated into health programs and services**[7].**

3 Dataset Review

Geography:

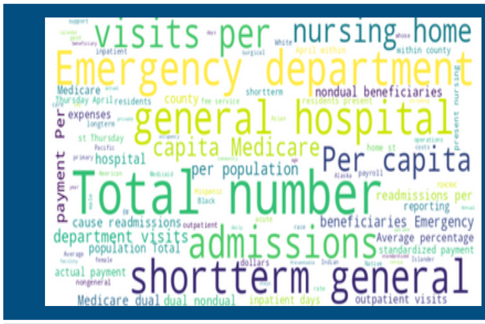
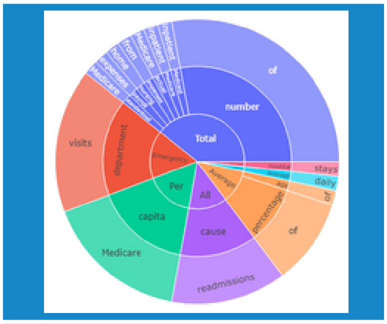
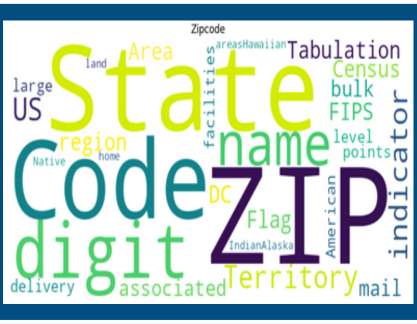
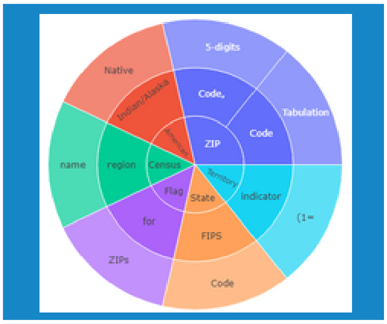
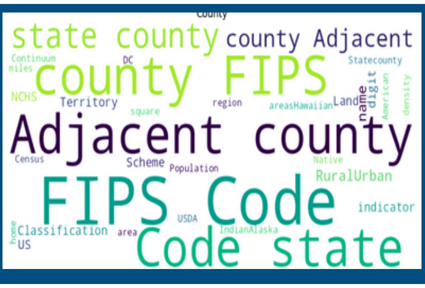
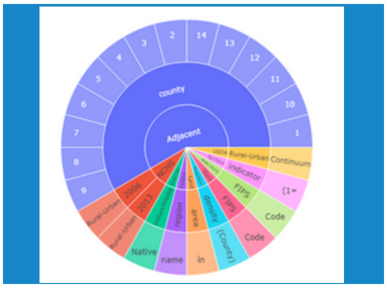
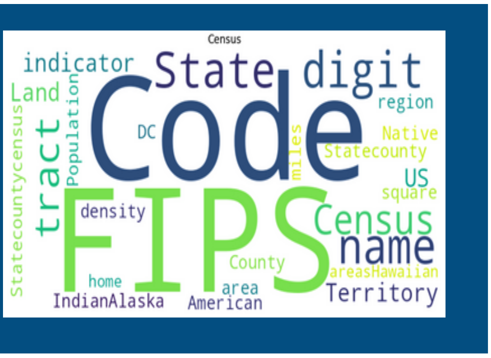
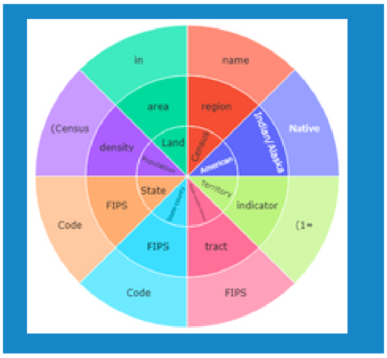
The dataset at hand offers an extensive examination of the socioeconomic aspects that have an impact on healthcare outcomes and access across various ZIP codes. This encompasses income, poverty rates, education, demographic makeup, health outcomes, and hospital utilization rates. The dataset also captures geographic diversity through regional categorization and political affiliation, along with proxies for vulnerable populations, such as Medicaid and Medicare enrollment. By analyzing health outcomes, political affiliations, and enrollment rates, policymakers and healthcare professionals can gain a better understanding of the unique challenges and opportunities to improve health outcomes and healthcare delivery within each county. By identifying correlations and patterns, the dataset enables informed decision-making and targeted interventions..

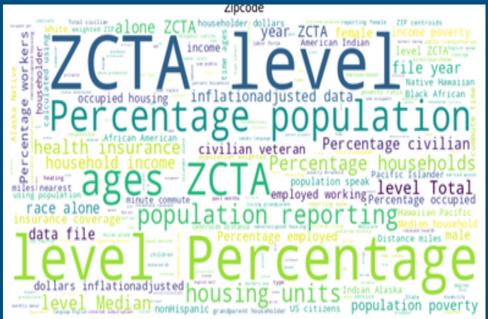
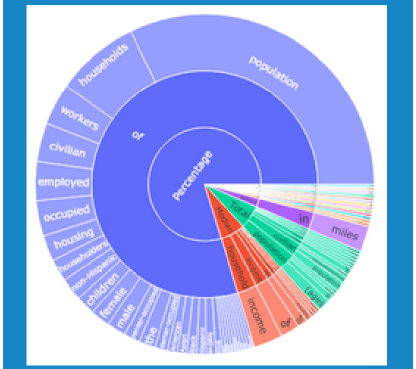
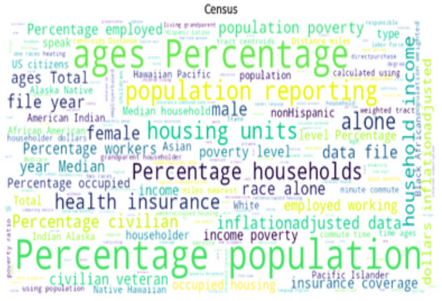
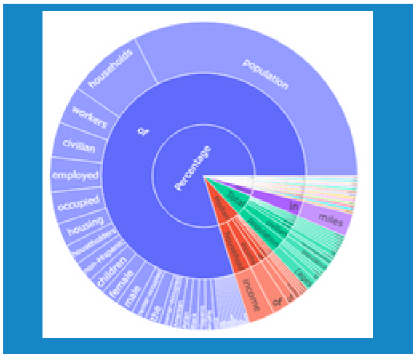
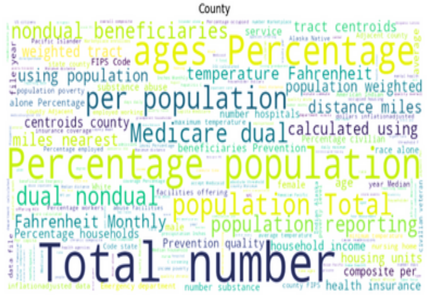
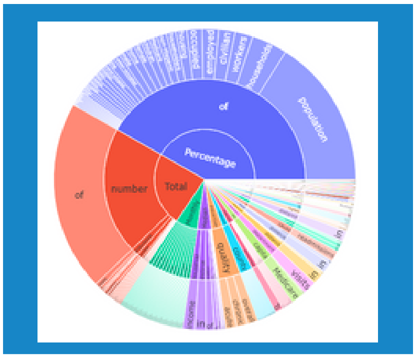
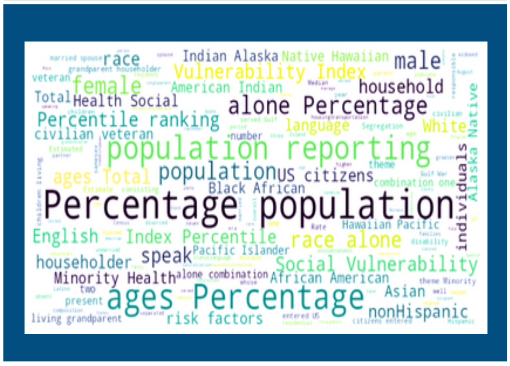
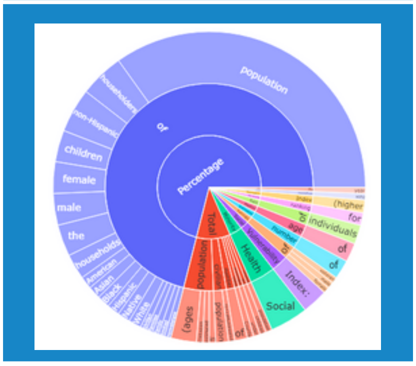
Included Features:

The dataset provides a comprehensive overview of demographic, socioeconomic, educational, housing, transportation, internet access, environmental factors, and weather events in a given area. It includes data on population composition, social vulnerability, health, language spoken, income inequality, poverty rates, employment statistics, food stamps/SNAP benefits, incarceration probabilities, economic typology codes, and more. It also includes data on educational attainment, financial data, school programs and services, safety and discipline incidents, enrollment numbers, academic performance, literacy and numeracy levels, and PIAAC data. The dataset also includes data on housing data, environmental factors, health and safety, economic indicators, and weather patterns. It also provides a comprehensive overview of healthcare coverage, professionals, hospital infrastructure, and market concentration. The dataset also includes variables such as USDA Rural-Urban Continuum Code 2013, adjacent county FIPS Codes, and ZIP flags.

Exploratory Data Analysis (EDA)

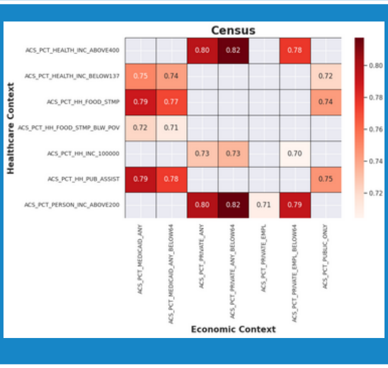
4.1 Word cloud Analysis

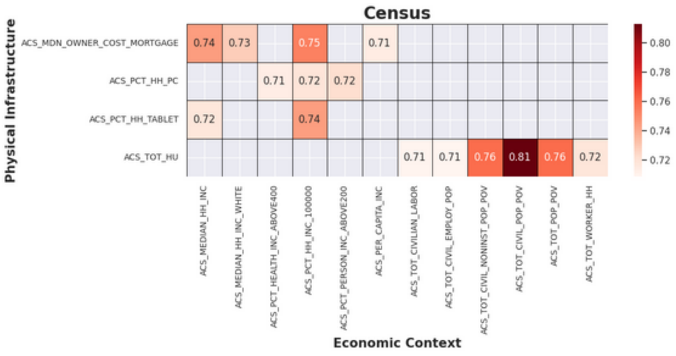
Topic	Word Clouds	Sunburst Word Clouds
Healthcare: We see 'Emergency department' being a leading determinant along with 'General Hospital' and 'nursing home', whereas we see 'beneficiaries', 'readmissions', and 'surgical' as 'minor factors'.		
Geography ZipCode: We see 'State', 'digit', and 'name' as some of the major determinants. 'Region', territory, delivery, American, and native are the smaller ones.		
Geography County: We see Adjacent, county, and FIPS Code as the significant determinants, whereas population, region, census, and territory are the smaller ones.		
Geography Census: We see code, FIPS, and State as the primary determinants and census, name, region, and indicator as the smaller ones.		

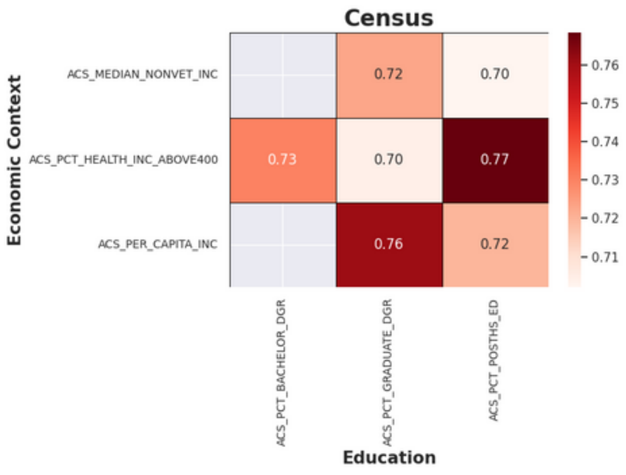
Topic	Word Clouds	Sunburst Word Clouds
Physical Infrastructure Zip Code: We see ZCTA, Percentage population, and ages as significant contributors, whereas insurance coverage, housing units, employed working, and percentage occupied are the smaller ones.		
Physical Infrastructure Census: We see Percentages of population and age as significant determinants, whereas poverty level, household dollars, and civilian veterans are smaller.		
Physical Infrastructure County: we see Medicare, age percentage, and Population total as major determinants, whereas alone percentage and health insurance are lesser.		
Social Context: The percentage of the population is the major contributor along with social vulnerability; factors like living grandparents, age total, and African American are the least contributing factors.		

4.2 Correlation Analysis

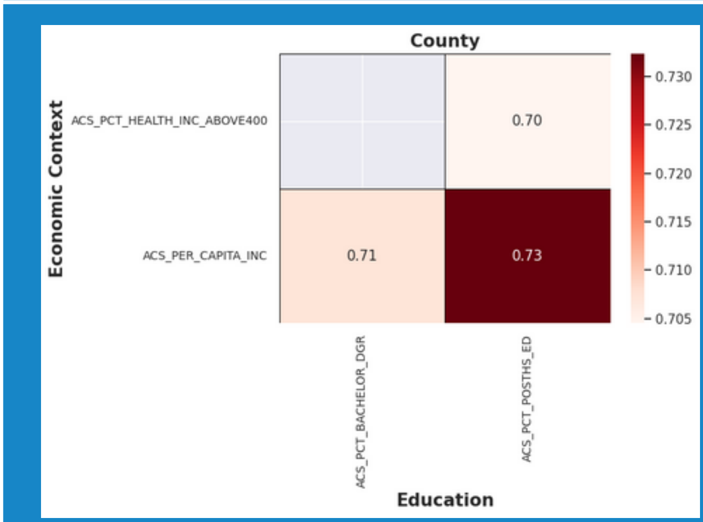
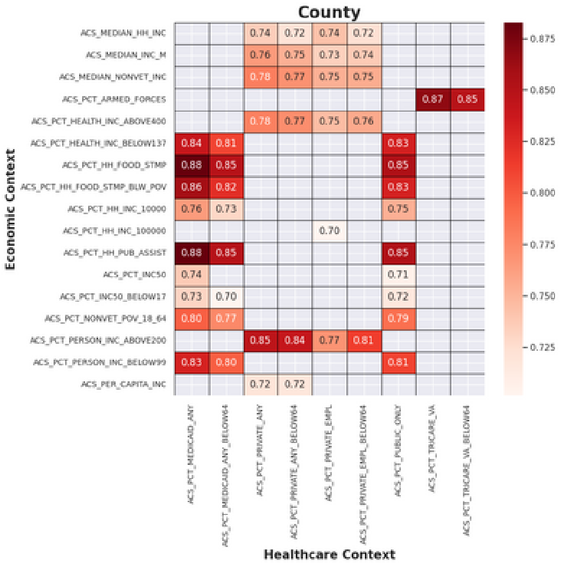
As the number of permutations of the variables for correlation analysis is vast, we have mentioned only the most significant correlations (pairs of factors that correlate greater than 0.7).

<table>
<tr><th>Topic</th><th>Correlation Analysis</th></tr>
<tr><td>Correlation between Economic and Health factors (for Census data): The dataset shows that socioeconomic factors like income, public aid, and healthcare enrolment have moderate to significant positive associations (0.70 to 0.82).</td><td>


</td></tr>
<tr><td>Correlation between Economic factors and Physical Infrastructure (for Census data): Correlations show strong links between socio-economic factors: higher median owner costs for mortgages correlate with increased median household incomes (0.744), and regions with more households owning tablets tend to have higher median incomes (0.718).</td><td>


</td></tr>
<tr><td>Correlation between Economic factors and Education (for Census data): The correlations show that higher household incomes correspond to higher educational attainment levels: households with health insurance income above $400k have higher percentages of individuals with bachelor's (0.73) and graduate degrees (0.70), while areas with higher per capita incomes tend to have higher percentages of individuals with post-high school education (0.72-0.76).</td><td>


</td></tr>
</table>

Topic	Correlation Analysis
Correlation between Economic context and Education (for County data): The study reveals a moderate positive correlation (0.71) between per capita income and the percentage of individuals with a bachelor's degree, suggesting a link between economic prosperity and educational attainment. Additionally, a moderate positive correlation (0.70) exists between health insurance income above $400,000 and the percentage of individuals with post-high school education, suggesting a link between socioeconomic status and educational attainment in wealthier areas.	
Correlation between Economic and Health factors (for County data): These correlations reveal how socioeconomic indicators are related. For example, there are strong positive correlations (around 0.80 or higher) between indicators associated with Medicaid coverage and various socioeconomic factors like households receiving public assistance, families with income below $137,000, etc. This suggests a strong association between Medicaid coverage and socioeconomic status.	

5 Methodology

5.1 Social Scorecard Prediction

Overview of proposed risk analyzing model/architecture:

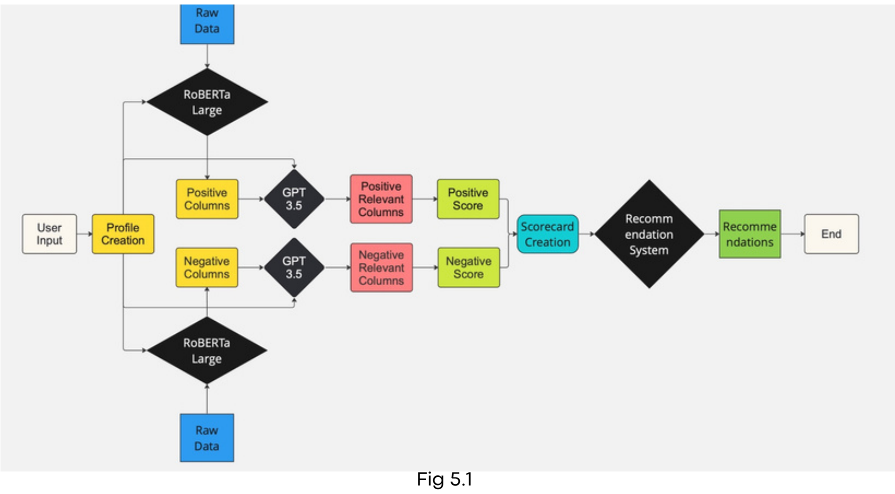


Fig 5.1

In the above-envisioned model, the scorecard for a user is generated using cutting-edge LLM architectures (GPT 3.5) to identify the features that affect the proposed social determinants based on data provided by the user. Another LLM, RoBERTa, pre-trained on sentiments, determines the nature of a particular feature, whether it affects the determinants positively or negatively, after which GPT identifies the relevance of these classified features based on user input. The model incorporates both precomputed and real-time calculated variables.

5.1.1 User Input:

User input will consist of some variables; the zip code will be provided at a minimum. The user's other information may include address, gender, age, race, income, education, and veteran status. These variables may or may not be available for each user

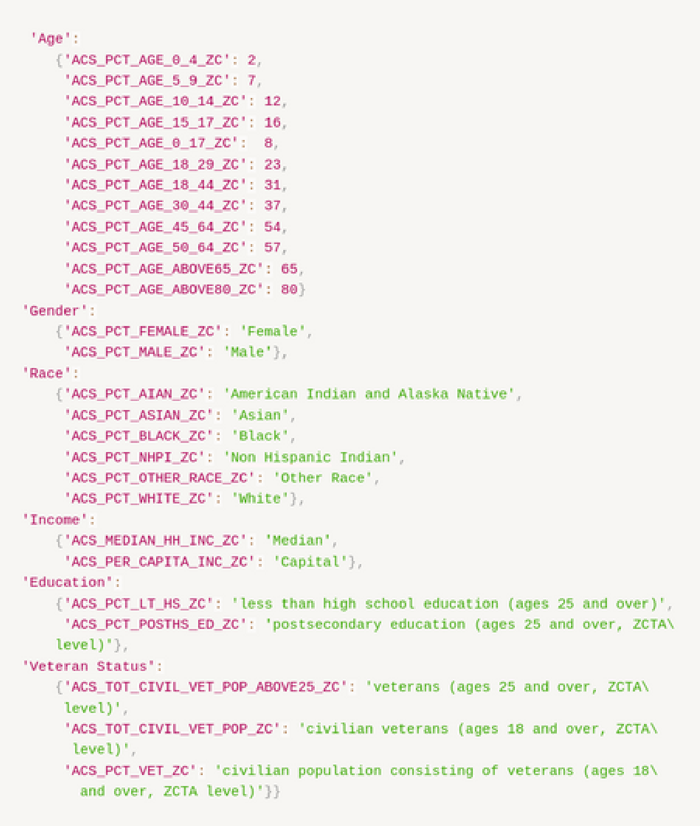
5.1.2 Preliminary computations:

1. Default profile creation per zip code:
    - Features that provide information about the majority in any case for the given population for a zipcode are identified through EDA. Their values are used for a default profile for each zip code. Thus, we expect the individual to belong to the majority of the given area. This way, the model can adapt to insufficient data. You can find the mapping of columns to determine the user profile.
2. Qualitative Classification of features with respect to Social Determinants:
    - The given data provides a description of each feature, which is used to classify them into 6 classes, which are our proposed social determinants. These sets are kept exhaustive to optimize the efficiency in the prompt provided to GPT.
3. Sentiment analysis of feature sets:
    - The qualitatively classified features of each social determinant are further passed to RoBERTa to analyze its sentiment. The features are classified positively and negatively based on how they affect the social determinant — Positively affecting features help decrease the risk of the determinant, whereas negatively affecting features increase the risk.

5.1.3 Profile creation:

The profile creation involves the utilization of our first assumption that we will always have a zipcode corresponding to each user profile. We precomputed some columns using the zip code we had to use to predict each user feature.

The user profile contains various features, so we assume that in absence of features except the zipcode corresponding to it, we extract the columns that have a majority in that region, using which we fill the other unfilled features. The user can fill in any number of features he likes and . the unfilled features are assigned a default value, which is the majority in a specific zipcode. The precomputed columns are mapped to the default values concerning a specific description. These default values are used to initialize the user profile. The list of mapping and default columns is given in the JSON file below.



```
 'Age':
    {'ACS_PCT_AGE_0_4_ZC': 2,
     'ACS_PCT_AGE_5_9_ZC': 7,
     'ACS_PCT_AGE_10_14_ZC': 12,
     'ACS_PCT_AGE_15_17_ZC': 16,
     'ACS_PCT_AGE_0_17_ZC':  8,
     'ACS_PCT_AGE_18_29_ZC': 23,
     'ACS_PCT_AGE_18_44_ZC': 31,
     'ACS_PCT_AGE_30_44_ZC': 37,
     'ACS_PCT_AGE_45_64_ZC': 54,
     'ACS_PCT_AGE_50_64_ZC': 57,
     'ACS_PCT_AGE_ABOVE65_ZC': 65,
     'ACS_PCT_AGE_ABOVE80_ZC': 80}
'Gender':
    {'ACS_PCT_FEMALE_ZC': 'Female',
     'ACS_PCT_MALE_ZC': 'Male'},
'Race':
    {'ACS_PCT_AIAN_ZC': 'American Indian and Alaska Native',
     'ACS_PCT_ASIAN_ZC': 'Asian',
     'ACS_PCT_BLACK_ZC': 'Black',
     'ACS_PCT_NHPI_ZC': 'Non Hispanic Indian',
     'ACS_PCT_OTHER_RACE_ZC': 'Other Race',
     'ACS_PCT_WHITE_ZC': 'White'},
'Income':
    {'ACS_MEDIAN_HH_INC_ZC': 'Median',
     'ACS_PER_CAPITA_INC_ZC': 'Capital'},
'Education':
    {'ACS_PCT_LT_HS_ZC': 'less than high school education (ages 25 and over)',
     'ACS_PCT_POSTHS_ED_ZC': 'postsecondary education (ages 25 and over, ZCTA\
    level)'},
'Veteran Status':
    {'ACS_TOT_CIVIL_VET_POP_ABOVE25_ZC': 'veterans (ages 25 and over, ZCTA\
     level)',
     'ACS_TOT_CIVIL_VET_POP_ZC': 'civilian veterans (ages 18 and over, ZCTA\
      level)',
     'ACS_PCT_VET_ZC': 'civilian population consisting of veterans (ages 18\
       and over, ZCTA level)'}}
```

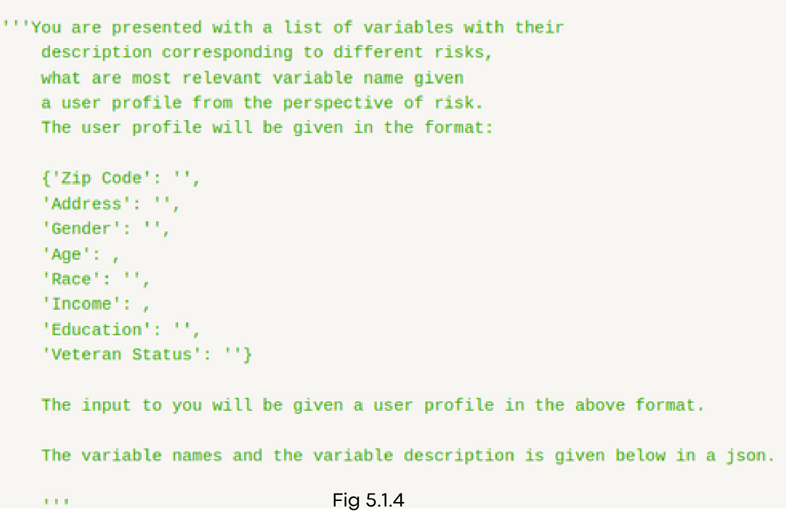
Fig 5.1.3

5.1.4 Prompt creation:

The prompt for retrieving appropriate columns involves explaining the format/schema in which the data and the user_info/ user profile will be given.

Now, corresponding to the positive sentiment columns concerning each risk are passed, and then the LLM is asked to select the top-k columns/variables that influence the user. The columns/variables are fetched in context with the user_profile.

The prompt utilizes a specific JSON format for user_profile, risk, and associated variables.

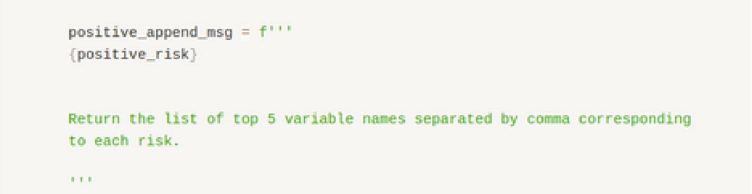


```
'''You are presented with a list of variables with their
    description corresponding to different risks,
    what are most relevant variable name given
    a user profile from the perspective of risk.
    The user profile will be given in the format:

    {'Zip Code': '',
    'Address': '',
    'Gender': '',
    'Age': ,
    'Race': '',
    'Income': ,
    'Education': '',
    'Veteran Status': ''}

    The input to you will be given a user profile in the above format.

    The variable names and the variable description is given below in a json.

    '''
```

Fig 5.1.4

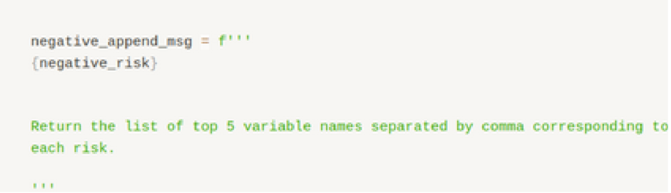
The top 20 most risk-prone variables within, both positive and negative, are pre-computed using the LLM. The top 20 variables with their description/variable_label are added to the prompt. The LLM takes the context of the variable label and user profile to output top-5 columns corresponding to each risk.

The prompting method combines action and example-based zero shot prompt, where the structure is explained to reduce/mitigate the hallucinations.



```
positive_append_msg = f'''
{positive_risk}


Return the list of top 5 variable names separated by comma corresponding
to each risk.

'''
```

Fig 5.1.4(a)



```
negative_append_msg = f'''
{negative_risk}


Return the list of top 5 variable names separated by comma corresponding to
each risk.

'''
```

Fig 5.1.4(b)

5.1.5 Identification of relevant features through GPT:

Two prompts are made to the LLM based on positively affecting and negatively affecting features. In the first prompt, the pre-defined skeleton of the prompt is coupled with the user profile and positively affecting feature bins corresponding to each determinant. Similarly, the second prompt, consisting of negatively affecting feature bins and user profiles, is made. The LLM in both the prompts is asked to identify the relevant features based on the user profile, e.g., in the case of age, through EDA, it has been identified that features targeting various age groups are present in the dataset. Thus, all the features targeting age groups where the user's age does not lie are discarded.

Finally, we have obtained all the relevant positively and negatively affecting features based on the user profile of the corresponding determinants.

5.1.6 Risk computation for scorecard:

After the normalization of each feature in both the sentimental bins, the sum of all the values of columns is found separately in both bins for a particular social determinant. The computed sums for each determinant in the positive bin are assigned a positive sign, and the sum in the negative bin is assigned a negative sign. The net of both these values is computed for the determinant, which is the risk of that social determinant. This operation is done for all 6 social determinants, and finally, we obtain the scorecard that consists of risk for each social determinant. This scorecard will further be used for suggesting appropriate facilities to the user to its aid.

5.2 Recommendation Architecture

A database was set up using public website scrapers provided in the problem statement around the domains spanning our risk set. This dataset comprises of solutions solving each risk based on the available data for the regions of A,B,C. On getting the scorecard output from our algorithm, next we have to find out the right set of recommendations for the user.

The top two social determinant risks which have the least values in the scorecard are picked. We already know the user’s zipcode and hence we couple it with the top two alarming scores and refer to the master dataset to device the perfect set of recommendations for the user as shown in the figure(abc). In cases where a particular zipcode doesn’t have available information for a particular risk, nearby zipcodes are referred to, to provide recommendations.

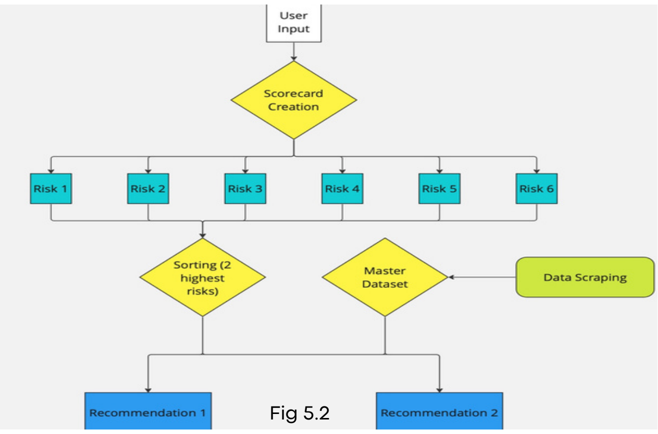


Fig 5.2

6 Results and Conclusion

Zipcode	Age	Gender	Income	Race	Education	Veteran status	Ground Truth	Health risk
2138	7	M	58418	white	'less than high school education (ages 25 and over)'	Null	Medication review due (situation)	-0.7276686250262385
2210	24	F	33245	Asian	postsecondary education (ages 25 and over, ZCTA level)'}	'civilian veterans (ages 18 and over, ZCTA level)'	Medication review due (situation)	-0.85387377008457933
2724	22	F	12290	white	postsecondary education (ages 25 and over, ZCTA level)'}	'civilian veterans (ages 18 and over, ZCTA level)	ChildhoodAsthma	-0.414211955202096
1960	71	M	35749	white	postsecondary education (ages 25 and over, ZCTA level)'}	veterans (ages 25 and over, ZCTA level)'	Received higher education (finding)	-0.5289952007247568
1453	6	M	91504	white	'less than high school education (ages 25 and over)'	Null	Medication review due (situation	-0.8980942518520187

The above table represents the profile of test data of patient data generated by the Synthea. The profile was generated as mentioned previously in methodology by replacing the values of default profile generated for the zipcode in advance by the given values. The patient is diagnosed by the “Medication review due (situation)” which is represented under Ground Truth. When this user profile was presented to our model architecture, it calculated the risk value for Health Case System to be -0.7276686250262385 along with other risk values. Thus, it can be stated that the patient is under these risks along with risks in the Health Care System. The recommender system refers to the master dataset and recommends facilities nearby the patient to reduce risks.

Through the above case study of an example patient, it is validated that our state of the architecture can precisely identify the Health Care System risks along with other risks that the user can face in the given area with zipcode. If the test data would provide other variables necessary in the user profile, our model would predict other risks more precisely.

In conclusion, a model has been created that outputs the intensity of risk of a person with respect to several contexts namely economic stability, neighbourhood and physical environment, education, food, community and social context and healthcare system. This has been done using a state of the art architecture with the LLMs at the heart of this operation. Coming to the final product, a pipeline has been implemented complete with a UI wherein the user enters his details and in return gets a scorecard where his/her risk profiles are highlighted complete with recommendations on what can be done to mitigate these said risks through methods and algorithms specified throughout the report.

In foresight, a future may be envisioned where such a risk profile of an individual can be associated with his/her Aadhar Card/Social Security number and using this, the Government may identify hotspots where certain risks are increased and in turn try to better them. Along with this, there may also be the availability of a central recommendation system which will recommend ways of reducing the risk profiles of an individual, thus improving the quality of life.

7 The Indian Context

India faces unique healthcare challenges where social determinants of health play a crucial role. We can address these factors through innovative data sources, leveraging them to create a proactive care approach that can have a lasting impact on the nation's health outcomes.

Here are some potential data sources and approaches that could be explored:

- Leveraging existing government data sources: India already has several nationwide surveys and data collection mechanisms, such as
  - the National Family Health Survey (NFHS),
  - District Level Household and Facility Survey (DLHS),
  - Census data of India
  - Statistical Election Reports from Election Commission of India,
  - National Ambient Air Quality Status & trends in India; Central Pollution Control Board Delhi.
  - Annual Report on Public Health, GOI, New Delhi.
  - Targeted Public Distribution System by Department of Food and Public Distribution, GOI.
  - Integrated Child Development Services, by Ministry of Labour and Ministry of Women and Child Development, GOI

These sources offer demographic, socioeconomic, and health data at different levels (national, state, district, village). Analyzing these integrated datasets reveals patterns and correlations between social determinants and health outcomes.

- Partnering with primary healthcare providers like state government health providers, community health workers, and NGOs, such as the Mitanin Programme in Chhattisgarh, India, can enhance data collection by leveraging their deep understanding of local contexts and community dynamics, supplementing government agency data. (Nandi, S., & Schneider, H. (2014). Health Policy and Planning, 29(suppl_2), ii71-ii81.): https://academic.oup.com/heapol/article/29/suppl_2/ii71/587209

- GIS and satellite imagery provide insights into environmental factors like air quality and land use. Overlapping with demographic data pinpoints hotspots, aiding effective intervention targeting.

Utilizing Social Determinants of Health (SDOH) data effectively, especially through the Multidimensional Poverty Index (MPI) derived from sources like the National Family Health Survey, enables a proactive healthcare approach. By employing predictive and risk analysis, patterns can be extracted among different determinants such as income, education, living conditions, and health outcomes. The MPI, established in 2011, tracks progress from 1990 to 2011 in various indicators including indoor biomass fuel use, sanitation, child malnutrition, and gender gaps in education and governance participation. Regression models using MPI indicators can identify high-risk individuals and communities, guiding targeted interventions like health education campaigns, screening programs, and initiatives to improve water access, sanitation, and healthy behaviors. This approach supports community-based responses and preventive care programs, addressing specific SDOH indicators comprehensively and enhancing healthcare outcomes at the grassroots level. (Cowli R., & Dandona, L. (2014). Social determinants of health in India: progress and inequities across states. International journal for equity in health, 13, 1-12.)

Integrated Care Pathways, involving collaboration among medical professionals, community organizations, social workers, and government agencies, can address both medical and social factors of individuals. Data-driven insights optimize resource allocation, prioritize investments in needy areas, and help India transition from reactive to proactive healthcare, improving outcomes and reducing costs for more equitable communities.

This approach streamlines identifying high-risk communities and vulnerable populations, tailoring health promotion strategies, informing urban planning, empowering stakeholders, and fostering collaboration. Adopting a data-driven approach in India can address social determinants of health, improving outcomes and equity.

Analyzing social determinants guides resource allocation and interventions for better health outcomes, prioritizing underprivileged areas. Education, income, and health literacy data inform poverty alleviation and urban planning, fostering healthier living. This approach ensures informed decisions for building equitable communities in India.

8 References

[1] Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., ... & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. Journal of the American Medical Informatics Association, 25(3), 230-238.

[2] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[4] Yenduri, G., Srivastava, G., Maddikunta, P. K. R., Jhaveri, R. H., Wang, W., Vasilakos, A. V., & Gadekallu, T. R. (2023). Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. arXiv preprint arXiv:2305.10435.

[5] Shao, Z., Gong, Y., Shen, Y., Huang, M., Duan, N., & Chen, W. (2023, July). Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In International Conference on Machine Learning (pp. 30706-30775). PMLR.

[6] To, O. (1993). Actual Causes of Death in the United States. JAMA, 270(18).

[7] Makridis, C. A., Zhao, D. Y., Bejan, C. A., & Alterovitz, G. (2021). Leveraging machine learning to characterize the role of socio-economic determinants on physical health and well-being among veterans. Computers in Biology and Medicine, 133, 104354.

[8] McKay, L. (2001). Changing Approaches to Health: The History of a Federal/Provincial/Territorial Advisory Committee. CPRN= RCRPP.