POPULATION HEALTH INFORMATICS, A PATH TO HEALTH EQUITY FOR MEDICAID PATIENTS ENROLLED IN THE PATIENT CENTERED MEDICAL HOME (PCMH)

By Carmen Hall

A Dissertation Submitted
In partial fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics
Rutgers, The State University of New Jersey
School of Health Professions

January 2022
Final Dissertation Defense Approval Form

Population Health Informatics, A Path to Health Equity for Medicaid Patients Enrolled in the Patient Centered Medical Home (PCMH)

BY:
Carmen Hall, MA

Dissertation Committee:
Shankar Srinivasan PhD
Frederick Coffman, PhD
Janos Hajagos, PhD

Approved by the Dissertation Committee on 12-13-2022
ABSTRACT

**Background:** The Medicaid population is burdened with social and environmental factors that impede care and result in poor health outcomes.¹ Currently, the collection of social determinants of health (SDH) in electronic databases in the United States is fragmented and inadequate resulting in health gaps leading to substandard care. The integration of SDH with medical information will help achieve health equity for the Medicaid population and determine how SDH affect adherence to physician recommendations.²

**Methods:** Supervised and unsupervised ML algorithms were used to mine data and develop prediction models to determine if social determinants of health were risk factors for appointment compliance for Medicaid patients enrolled in the Patient-Centered Medical Home model from July 1, 2017 – June 30, 2019. This retrospective study analyzed records for 911 patients, ages 18-65 years and 12,118 encounters.

**Results:** k-nearest neighbors had the best performance (AUC=0.743). Support vector machines (AUC = 0.656), logistic regression (0.644) and random forest had the lowest performance (0.599). Tobacco-related disorders, nutritional anemia, age, and gender predicted preventative health appointments.

**Conclusion:** SMOTE oversampling technique can be used to balance a minority class to improve risk predictions significantly compromising performance scores (AUC, precision, recall, F1 score).
ACKNOWLEDGEMENTS

I would like to thank the members of my committee, Dr. Shankar, Dr. Coffman, and Dr. Hajagos, I appreciate their support, recommendations, and advice. My most sincere gratitude to my advisor, Dr. Janos Hajagos for his time and unwavering support. Words cannot express how much I appreciate his mentorship. I am extremely blessed to have had Dr. Hajagos’ guidance.

I want to thank my husband who has been my anchor and source of encouragement, and my children, Sameerah, Sheena and Kyle for believing in me. Finally, I would like to thank all my friends in academia and other walks of life. I am extremely blessed to call them my friends.

I dedicate this work to my parents that departed earth for heaven in 2019.

They are the angels that watch over me.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................ix

LIST OF FIGURES ......................................................................................................x

LIST OF EQUATIONS ..................................................................................................xii

I. INTRODUCTION ......................................................................................................1

1.1 Statement of the problem .......................................................................................1

1.2 Background of the problem ...................................................................................3
  1.2.1 Population Health Informatics .....................................................................5
  1.2.2 Electronic Health Records ...........................................................................6
  1.2.3 International Classification of Disease .......................................................8
  1.2.4 Patient Protection Affordable Care Act ....................................................9
  1.2.5 Comprehensive Health Care .....................................................................10
  1.2.6 Patient Centered Medical Home ...............................................................11
  1.2.7 Delivery System Reform Incentive Payment ............................................11

1.3 Project Purpose ....................................................................................................14

1.4 Hypothesis ..........................................................................................................14

1.5 Need for study .....................................................................................................15

1.6 Definitions ..........................................................................................................16

1.7 Theoretical framework ........................................................................................17

1.8 Challenges & Limitations ....................................................................................18

1.9 Mitigation Plan .....................................................................................................18

II. REVIEW OF RELATED LITERATURE ..................................................................20

2.1 Overview .............................................................................................................20

2.2 Documenting Social Determinants in Electronic Health Records .....................20

2.3 Mining Electronic Health Records ....................................................................23

2.4 Extracting Environmental Data Using Geospatial Technologies ......................25

2.5 Combining Data Collection and Geospatial Technologies ................................28

2.6 Coding Social Determinants in Electronic Health Records .............................30

2.7 Social Determinants of Health and Risk Stratification .....................................32

2.8 Conclusion .........................................................................................................33

III. METHODS ..........................................................................................................39

3.1 Overview .............................................................................................................39

3.2 Subjects ..............................................................................................................39

3.3 Research Design and Rationale .......................................................................39

3.4 Data Sources .....................................................................................................40
IV. RESULTS ............................................................................................................. 55

4.1 Overview ............................................................................................................ 55
4.2 Patient Demographics ..................................................................................... 55
  4.2.1 Age Distribution of Population .................................................................... 56
  4.2.2 Population by Race ..................................................................................... 57
  4.2.3 Population by Region ................................................................................ 58
4.3 Education by Geographic Region ................................................................. 58
  4.3.1 Core Region ............................................................................................. 59
  4.3.2 Central South Region ............................................................................... 60
  4.3.3 Central East Region ................................................................................. 61
  4.3.4 Northfork Region .................................................................................... 62
  4.3.5 Northwest Region ................................................................................... 63
  4.3.6 Southfork Region .................................................................................... 63
  4.3.7 Southwest Region ................................................................................... 64
4.4 Computer and Internet Access ...................................................................... 65
4.5 Patient Comorbidities ..................................................................................... 66
4.6 Preventive Health Appointments .................................................................... 67
  4.6.1 Characteristics of Appointments ................................................................ 68
4.7 Social Determinants of Health ....................................................................... 68
  4.7.1 Characteristics of Social Determinants of Health by Gender ...................... 69
  4.7.2 Characteristics of Social Determinants of Health by Race ......................... 70
4.8 Cluster Analysis .............................................................................................. 71
  4.8.1 Cluster Scatter Plot .................................................................................. 72
  4.8.2 Health Conditions ..................................................................................... 72
  4.8.3 Cluster Health Conditions by Gender ....................................................... 73
  4.8.4 Cluster Health Conditions by Age Group ................................................ 74
  4.8.5 Cluster Health Conditions by Race .......................................................... 75
  4.8.6 Cluster Health Conditions by Region ........................................................ 76
4.9 Classification Models ...................................................................................... 76
4.9.1 Features .....................................................................................................................78
4.9.2 Random Forest (RF) Feature Importance.................................................................79
4.9.3 Classification Model Validation and Accuracy Scores ..............................................80
4.9.4 Random Forest Model Performance ..........................................................................80
4.9.5 Logistic Regression Model Performance .................................................................81
4.9.6 K-Nearest Neighbors Model Performance ...............................................................82
4.9.7 Support Vector Machine Model Performance .........................................................83
4.9.8 SMOTE Precision and Recall Curve .........................................................................84
4.9.9 Summary Metrics ....................................................................................................85
4.9.10 Probability of Appointments and Predictions .........................................................85

V. DISCUSSION .....................................................................................................................87

5.1 Overview .......................................................................................................................87
5.2 Model predictions .........................................................................................................87
5.3 Exploratory Analysis Findings .....................................................................................88
5.4 Patient Social Determinants of Health .......................................................................88
5.5 Racial Demographics of Patient Population and Algorithm Bias .................................89
5.6 Limitations ....................................................................................................................89

VI. CONCLUSIONS AND FUTURE RESEARCH .................................................................91

6.1 Study Summary ...........................................................................................................91
6.2 Future Research ..........................................................................................................91
LIST OF TABLES

Table 1: ICD-10-CM Social Determinants of Health Codes & Descriptions ..................9
Table 2: Summary of Social Determinants of Health Identified in Literature Review.....36
Table 3: CCSR Social Determinants of Health Descriptions and Codes ....................41
Table 4: CCSR Health Condition Descriptions and Codes.....................................42
Table 5: CCSR Preventive Health Appointments Description and Codes......................42
Table 6: Suffolk County Regions and Cities ........................................................43
Table 7: Independent and Dependent Variables ...................................................44
Table 8: Core Region Computer and Internet Access.............................................65
Table 9: Central South Computer and Internet Access.........................................66
Table 10: Social Determinants of Health by Gender.............................................70
Table 11: Social Determinants of Health by Race...............................................71
Table 12: Health Conditions by Cluster.............................................................73
Table 13: SelectKBest Feature Selection Results ...............................................79
LIST OF FIGURES

Figure 1: Conceptual Model of Intervention & Outcomes ..............................................14
Figure 2: Data Wrangling Process Steps ........................................................................46
Figure 3: Feature Vectors .................................................................................................47
Figure 4: Gender Totals ....................................................................................................56
Figure 5: Age Group Totals ...............................................................................................56
Figure 6: Population Race Totals ........................................................................................57
Figure 7: Population Region Totals ...................................................................................58
Figure 8: Region Totals, Bachelor’s degree/higher ............................................................59
Figure 9: Core Region City Totals, Bachelor’s degree/higher .............................................60
Figure 10 Central South Region City Totals, Bachelor’s degree/higher ...............................61
Figure 11: Central East Region City Totals, Bachelor’s degree/higher .................................62
Figure 12: Northfork Region City Totals, Bachelor’s degree/higher .....................................62
Figure 13: Northwest Region City Totals, Bachelor’s degree/higher ....................................63
Figure 14: Southfork Region Totals, Bachelor’s degree/higher ...........................................64
Figure 15: Southfork Region City Totals, Bachelor’s degree/higher .....................................64
Figure 16: Chronic Disease by Gender ..............................................................................66
Figure 17: Chronic Disease by Age Group .........................................................................67
Figure 18: Preventive Health Appointments for Population ................................................67
Figure 19: Social Determinants of Health Totals ................................................................69
Figure 20: Elbow Plot .........................................................................................................71
Figure 21: Clusters ............................................................................................................72
Figure 22: Cluster Analysis, Top 10 Health Conditions .......................................................73
Figure 23: Gender, Clusters 0-3 ........................................................................................74
LIST OF EQUATIONS

Equation 1: K-means Clustering .......................................................................................... 48
Equation 2: Logistic Regression .......................................................................................... 50
Equation 3: Random Forest ............................................................................................... 50
Equation 4: K-Nearest Neighbors ....................................................................................... 51
Equation 5: Support Vector Machine .................................................................................. 51
Equation 6: Accuracy ........................................................................................................ 52
Equation 7: Precision ....................................................................................................... 52
Equation 8: Recall ............................................................................................................ 53
Equation 9: F1-Score ....................................................................................................... 53
Equation 10: Specificity ................................................................................................... 53
Chapter I
INTRODUCTION

1.1 Statement of the Problem

Medicaid insures approximately 18% of the United States population; its beneficiaries are frequently burdened with comorbidities and social determinants that influence health outcomes and contribute to the health care deficit.\textsuperscript{3,4} In 2005, the World Health Organization (WHO) developed an action plan to address social determinants of health (SDH) and called on organizations to improve the conditions of daily life for marginalized populations by addressing the distribution of money, power and resources, raising awareness, and developing a workforce for measuring and evaluating SDH.\textsuperscript{5} The Department of Health and Human Services’, Office of Disease Prevention and Health Promotion Healthy People program is responsible for advancing national priorities for health prevention and promotion. They affirmed that health equity could not be achieved by controlling and treating chronic conditions, but that sustainable health care must include an evaluation of income, education, diet, environment, and social habits.\textsuperscript{6}

WHO defines health as “the state of complete physical, mental and social well-being and not merely the absence of disease or infirmity”\textsuperscript{7}, and social determinants of health as “conditions in which people are born, grow, live, work and age.”\textsuperscript{8} Researchers have estimated that approximately 60% of preventable deaths are associated with social, environmental and behavioral determinants of health.\textsuperscript{9} The harmful consequences that socioeconomic disparities have on a person’s life and health are well documented in the literature. Impoverished living conditions and low income dictate food insecurities and transportation problems affecting the health of individuals. Populations with chronic
diseases, a low level of education or mental conditions may struggle with reading and complying with medication instructions. According to research 36% of individuals covered by Medicaid do not have a high school education. Lastly, risky behaviors such as smoking, and alcohol consumption profoundly affects individuals with or without health conditions. SDH are upstream factors that impede care and health outcomes of Medicaid patients who are poor and have a history of receiving care that is substandard and inequitable.\textsuperscript{10,11} Public Health 3.0 launched in 2016 with recommendations for new positions focused on resolving health disparities related to SDH and the life expectancy of individuals that live in marginalized communities.\textsuperscript{12}

Recent initiatives are encouraging organizations to bring clinical and non-clinical data together. In 2014 the Office of the National Coordinator tasked the National Academy of Medicine (NAM) with identifying and recommending the SDH that should be captured in an EHR.\textsuperscript{11} The NAM formed a Committee on Social Determinants of Health that selected eleven SDH domains, four of which are systematically being collected (e.g., tobacco use, ethnicity/race, alcohol use and address) in EHR systems today. The remaining seven include, education, financial resource strain, stress, depression, physical activity, social isolation, exposure to violence or intimate partner violence, and neighborhood and median-household income. Having a comprehensive collection of SDH in the EHR will enable providers to identify medical interventions that align with patient conditions.\textsuperscript{13}

Despite the insurmountable evidence that link SDH to the Medicaid population, standards for routine data collection, storage, and usage of social determinants in EHRs have not been fully established. Additionally, there is limited research available connecting SDH that are captured electronically as risks factors to health outcomes. This work will explore SDH as risk factors for preventive health appointment compliance for
Medicaid patients that were enrolled in the Suffolk County Collaborative Delivery System Reform Incentive Payment (SCC-DSRIP) at Stony Brook Hospital. The SCC-DSRIP program was a federally funded project led by Stony Brook Hospital in Suffolk County, New York. It was focused on improving health outcomes and equity for Medicaid patients. Social determinants of health were collected, documented, and stored on paper for this project. The EHR was not used to document or access electronically stored SDH information. This study aims to:

- **Aim 1:** Identify SDH that are available in the EHR for Medicaid patients that were enrolled in the SCC-DSRIP.

- **Aim 2:** Identify the SDH that were mapped to the Medicaid population health platform for Medicaid patients that were enrolled in the SCC-DSRIP.

- **Aim 3:** Identify the association of SDH and appointment compliance for Medicaid patients that were enrolled in the SCC-DSRIP.

- **Aim 4:** Identify housing information (i.e., computer, internet access and education) for Medicaid patients that were enrolled in the SCC-DSRIP

### 1.2 Background of the Problem

Social inequalities were identified as precursors to disease for marginalized individuals as early as the 12th century. The Framingham studies in the 1940’s, Britain’s Whitehall civil servants’ studies that began in 1960’s, and the Black Report that documented inequalities in health in the 1980s continued to confirm how health was disproportionately linked to social determinants of health.14-15,16 The leading cause of death in the United States in the 1990’s was tobacco, diet, exercise, alcohol, microbial and toxic agents, firearms, sexual behavior, drugs, and motor vehicle accidents.17 The economic models that controlled the United States health care system prior to the 1990’s failed to support a model of whole-person-care. Countries outside of the United States
have capitalized on addressing social inequalities resulting in better health outcomes and more effective care. In 2020 the United States allocated 20% of its gross domestic product (GDP) on health care compared to other countries where the average spending on health care was approximately 11% of their GDP.18

The National Academy of Medicine (NAM) published several reports that led to the transformation of the United States healthcare system. These evidenced-based reports were embraced as the national agenda and supported by federal and state agencies, healthcare facilities, researchers, policy makers and providers. In 1999, “To Err is Human – Building a Safer Health System” identified that 46,000 – 98,000 people die annually in hospitals from preventable medical errors. Adverse events in hospitals, medical errors and the ordering of unnecessary tests contributed to the rising cost of healthcare in the U.S.19 This study served as evidence of a healthcare system that was broken, with minimal safeguards for providing safe care. Thus, there was a shift in policy changes centered on improving the quality of care for health care consumers.

“Crossing the Quality Chasm – A New Health Care System for the 21st Century”, released by the NAM in 2001, aim for healthcare systems was to embody processes that promoted health equity, safety, and efficiency through timely, effective, and patient-centered services. Health technologies and advanced systems that store, process, and retrieve health information were suggested to improve the U.S. healthcare system. Prominence revolved around high-risk populations, common health care conditions that increased the cost of health care and contributed to morbidity and mortality rates, and systems that rewarded providers for delivering quality services. Multidisciplinary care processes, teams and care coordination programs that placed patients and their families first began to be established.20
In 2004 The World Health Organization (WHO) developed the Commission on Social Determinants of Health to administer policies to reduce health inequities caused by social injustices.\textsuperscript{21} Healthy People 2020 set a goal to make transparent the impact of SDH on wellness and disease.\textsuperscript{22} The Center for Medicaid and Medicare Services (CMS) Center for Medicare and Medicaid Innovation (CMMI) was developed to examine new models of payment, health care delivery and technical innovations to address health inequities to improve the quality of care, and reduce and control health care costs in Medicare and Medicaid reimbursement programs.\textsuperscript{23}

1.2.1 Population Health Informatics

Donald Berwick, the President and CEO of the Institute of Health Improvement introduced the Triple Aim framework which culminated to a standard for improving the U.S. healthcare system. Triple Aim efforts was on improving the patient experience, population health and reducing health care costs. Healthcare organizations started working towards achieving Triple Aim initiatives by creating patient portals that engaged patients with technology, connecting reimbursements models to quality through incentive payments, targeting populations to improve overall health, etc.\textsuperscript{24,25}

Population health brings together interdisciplinary teams to streamline services and processes with technology enabling providers to collaborate to render more effective services. It connects the evaluation of social determinants to clinical practice through public health, medicine, computer science and information processing. Population health informatics is the application of technologies and electronic information such as databases and data analysis software used to store, process, and identify patterns in clinical and non-clinical data to improve the health of a community or population. It combines concepts of clinical, consumer and public health informatics which
interconnects through targeting populations, computer science, clinician, and consumer interactions.\textsuperscript{25,26}

1.2.2 Electronic Health Records

In the 1960’s and 1970’s, Dr. Larry Weed suggested adding financial and social needs into the Problem-Oriented Medical Information System, one of the first electronic medical records to assist with the clinical-decision process.\textsuperscript{27} Significant changes in EHR utilization were not seen until the 2009 Health Information Technology and Economic and Clinical Health (HITECH) legislation. HITECH was part of the American Reinvestment Act (ARRA) which allocated $27 billion in incentive payments through the CMS federal reimbursement programs to accelerate the adoption of interoperable EHRs.\textsuperscript{28} The Office of the National Coordinator (ONC), established under the Bush administration in 2006 to lead health information technology (IT) initiatives, and CMS were charged with developing policies to implement HITECH and Meaningful Use (MU). Additionally, the ONC developed regional extension centers to assist provider practices with system implementations, health IT curriculum to train professionals in informatics, certification requirements for EHR vendors and Beacon communities that focused on health IT research to achieve Triple Aim.\textsuperscript{29} In a 2015 report to congress, 78% of primary practices and 96% of hospitals had certified EHRs because of the HITECH law.\textsuperscript{30}

EHRs are fundamental components of health information technology that launched the healthcare industry into an era of big data, population health, and data analytics. They are used to inform clinicians, make a clinical diagnosis, and identify interventions. In hospitals, ancillary systems such as laboratory, radiology, and pharmacy can be linked to EHRs offering providers the opportunity to login to one system to view clinical notes, orders, and test results. Inpatient and outpatient EHRs are
designed to share one database or healthcare facilities employ integration technology to connect different vendor solutions. Data from EHRs can be extracted and transferred to registries, population health management systems, data warehouses, Regional Health Information Organizations (RHIO), or Health Information Exchanges (HIE) for data aggregation, to look at trends and patterns in patient and population data, and share clinical information with different providers of care and public health agencies.31

MU was organized and implemented in three stages. Stage I, published in 2010 focused on capturing data in the EHR and providing patients with electronic copies of their clinical record. Stage II, published in 2015 encouraging organizations to share data between different systems, communicate electronically among patients and provider organizations, and ensure that EHRs supported the National Quality Strategy (NQS). Attestation for MU occurred between 2011-2013, requiring healthcare facilities to submit reports to the federal government to demonstrate how they were meeting the MU standards established. Organizations that did not meet MU requirements were penalized through CMS reimbursement programs.32,33 Stage III is centered on addressing challenges associated with exchanging health information between systems, organizations, and providers, and supporting the sharing of information across platforms. Thus, the MU program was renamed to the Promoting Interoperability (PI) Program.34

The scope and aggressive timeline for MU stages I and II were met with revisions to dates and phases to adjust to the challenges that organizations stumbled upon when introducing new software, hardware, or processes to clinical practices and healthcare facilities. MU did not result in immediate changes or improvements to clinical care, because legislation failed to account for the cultural changes that would impact the introduction of new functionality and workflows to clinical and non-clinical end
users. Additionally, the infrastructure for information exchange did not exist or support MU goals.  

Clinical decision support (CDS) tools are embedded in EHRs to deliver alerts and reminders (e.g., drug-to-drug and drug allergy interactions) at the point-of-care to reduce medical errors and inform clinicians of important information related to a patient or population. Advanced CDS tools such as artificial intelligence (AI) are currently being explored and tested to improve health care delivery. AI technology has human intelligence capabilities enabling it to act upon data stored or captured real-time by offering solutions or making recommendations for care. Machine learning and data mining techniques combined with AI maximizes the clinical decision process. AI technologies have been used in the retail industry with large companies such as Amazon to increase sales by learning the behaviors of consumers. Unfortunately, ethical barriers and biases have slowed down AI adoption in health care because health decisions are life threatening and the diagnosis process involves many unknown variables that differ from patient-to-patient. Additionally, large volumes and a variety of data, and access to clinical trials presents challenges with research and development. The value of advanced CDS tools will be realized as health care processes evolve using AI technology.

1.2.3 International Classification of Disease

The International Classification of Disease, Clinical Modification (ICD-10-CM) codes have historically been used to report patient diagnosis and inpatient procedures; they are now being used in EHR’s to represent social determinants of health. ICD-10-CM SDH codes are more frequently used in hospitals during patient discharge to document mental conditions and alcohol/substance by physicians and non-physician staff (e.g., social workers, nurses, discharge planners or case managers). Table 1 has a list of the ICD-10-CM codes that are currently being utilized with a brief description. The list is
broad and does not contain all of the SDH that were recommended by the National Academy of Medicine.\textsuperscript{39,40} Additionally, the codes lack granularity leading to the misrepresentation or incomplete documentation of SDH in electronic health records. Expanding ICD-10-CM codes to better represent SDH is necessary for standardizing documentation practices, developing and tracking SDH data routinely and providing more comprehensive services.

<table>
<thead>
<tr>
<th>ICD Code</th>
<th>Social Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z55</td>
<td>Education and literacy</td>
</tr>
<tr>
<td>Z56</td>
<td>Employment and unemployment</td>
</tr>
<tr>
<td>Z57</td>
<td>Occupational exposures</td>
</tr>
<tr>
<td>Z59</td>
<td>Economic issues and housing</td>
</tr>
<tr>
<td>Z60</td>
<td>Social environment</td>
</tr>
<tr>
<td>Z62</td>
<td>Upbringing or how an individual is raised</td>
</tr>
<tr>
<td>Z63</td>
<td>Broad category represents other family circumstances and support groups</td>
</tr>
<tr>
<td>Z64</td>
<td>Psychosocial</td>
</tr>
<tr>
<td>Z65</td>
<td>Broad category represents other psychosocial issues</td>
</tr>
</tbody>
</table>

Source: American Hospital Association\textsuperscript{39}

1.2.4 Patient Protection Affordable Care Act

The Obama administration introduced policy to resolve social issues in the U.S. in 2010 by passing the Patient Protection Affordable Care Act (ACA). The ACA offered states an option to expand state Medicaid programs and made it mandatory for the uninsured to purchase health insurance. Additionally, the expansion of Medicaid redefined preventative health services. In addition to annual preventative health visits, the new model supported screening for breast, cervical and colorectal cancer, cardiovascular risks (e.g., cholesterol), alcohol use, HIV, vaccines, diabetes, etc.\textsuperscript{41,42} The National Strategy for Quality Improvement in Healthcare, established by the Department of Health and Human Services (DHHS) complies with the ACA law. It aims to promote the delivery of quality services and healthy communities through addressing behavioral, social and environmental determinants of health.\textsuperscript{43,44}
The ACA law required tax-exempt hospitals to complete a Community Health Needs Assessments (CHNA) to define the needs and gaps of health care services in local communities. CHNAs are conducted every three years allowing hospitals to build and execute effective strategies for satisfying health and community essentials to improve equity and population health. Details about community related activities are collected during assessments for a careful evaluation of health outcomes. Tax-exempt hospitals are responsible for including a tribal or regional public health department, consumers from marginalized communities, and minorities in the CHNA. The ACA law is grounded in policy that called for the integration of care delivery models, value-based payments connecting the quality of services to reimbursement, teamwork and the cooperation from all healthcare entities to care for vulnerable communities.

1.2.5 Comprehensive Health Care

Hospitals and community health centers started working together to provide a comprehensive approach to health care delivery. This concept initially emerged in Africa in the early 1900’s with two physicians, Sidney and Emily Kark. They conducted a survey on Black Africans which covered 1000 children in different urban centers in 1938. Sidney and Emily had an interest in examining the communities where Blacks lived and the impact that it had on their health. The survey was designed to collect SDH data (e.g., nutrition, environment, and health) which led to the development of the Pholela Health Center in South Africa. The National Health Services Commission recommended that the Pholela Health Center be linked to the national hospital system and local health centers. Thus, in 1945 the Institute of Family and Community Health was set up with community health workers to assist with caring for people in poor communities.
Kark’s establishment of the Community-Oriented Primary Care Center (COPC) was introduced to the United States by Jack Geiger, resulting in the building of Federally Qualified Health Centers (FQHC). These centers care for populations with or without health insurance and provide interventions to address community related issues (e.g., education, housing, food, etc.). Evidence has shown that 22 million of the most vulnerable and high-risk populations that are located in underprivileged communities receive care from FQHC’s where Medicaid is the largest source revenue. Historically, they’ve used a medical home approach to address social and medical issues.

1.2.6 Patient Centered Medical Home

The Patient-Centered Medical Home (PCMH), similar to the medical home, was endorsed by the ACA as a model to change reimbursements by awarding providers that provide quality services. Organizations that use the PCMH model can achieve National Committee for Quality Assurance (NCQA) recognition. The benefits of using the PCMH model were incentive payments and a patient-centered approach to health care. Embedded in the PCMH process were outcomes measurement, ensuring that practices were meeting standards, improving quality, and identifying gaps for improvement. This model encourages providers and care managers to work collaboratively to deliver comprehensive services. The physician coordinates care and support staff are tasked with following up with patients to validate understanding of office and urgent care hours to avoid unnecessary emergency room visits and hospitalizations.

1.2.7 Delivery System Reform Incentive Payment

The Delivery System Reform Incentive Payment (DSRIP) was a federally funded grant program designed for the Medicaid population. It was embedded with Triple Aim, the ACA and HITECH laws. Medicaid is one of the largest insurers in New York State,
with a patient population that suffers from the costliest health conditions. The main goal of DSRIP was utilizing the PCMH model to reduce avoidable hospital and emergency room visits by 25% in five-years for Medicaid recipients. DSRIP established a comprehensive healthcare strategy that weaved together social needs, disease management, community health centers, primary care physicians and specialists, long-term care, medical and behavioral health, and health information technology.

Partnerships were encouraged between hospitals and community healthcare organizations to reduce unnecessary expenditures, improve care coordination and the quality of care delivered. DSRIP entities worked together to achieve hospital utilization goals by shifting to primary care and preventative health services. This project presented the opportunity to narrow the gap of health disparities that plagued disadvantaged populations.\textsuperscript{55,56}

New York State was approved to implement the 5-year (2015-2019) DSRIP program. Years 1-2 focused on achieving implementation milestones and transitioning to pay-for-performance and 3-5 were performance years, each requiring an increase in the percentage of performance from constituents. In year 3, 45% of incentives had to be tied to pay-for-performance, 65% in year 4, and in year 5 it increased to 85%. DSRIP measured performance based on progress in system transformation, clinical improvement and population-wide project implementation using Medicaid and population core measures. Medicaid core measures were associated with preventative care visits, women’s health, chronic conditions, substance and mental abuse, long term and primary care, preventable events, etc. Population measures included clinical performance projects for diabetes, obesity, asthma, lead, tobacco, birth outcomes, etc. Transformation projects encompassed the implementation of integrated delivery systems, an increase of PCMH certifications, care coordination at all levels of health care,
connectivity in primary care settings, and expanded access to community-based care for special populations. Outcomes and process measures were evaluated through quality metrics drawn from Healthcare Effectiveness Data and Information Set (HEDIS), Consumer Assessment of Healthcare Providers and Systems, utilization measures, percentage of primary care providers meeting National Committee on Quality Assurance PCMH standards and the Agency for Healthcare Research and Quality. Performance metrics were focused on population health improvements. DSRIP incentive payments were distributed through the federal government when healthcare organizations met performance metrics and project milestones.57,58,59

Stony Brook Medicine, a 624-bed, tax-exempt, academic facility located in Suffolk County, New York is comprised of 3 hospitals and a portfolio of inpatient and outpatient healthcare services led the Suffolk Care Collaborative (SCC) DSRIP project. The SCC-DSRIP program was made up of three hubs, SBUH, Northwell Health and Catholic Health Services. Each hub was a separate entity with health centers and providers that cared for Medicaid patients.60 Based on the 2014 CHNA, the Suffolk County Medicaid population was estimated at 240,000. There were 119,932 emergency room encounters and 34,944 hospitalizations in 2012. Hospital admissions were associated with cardiovascular disease, substance abuse, psychiatric disorders, diabetes, asthma, cancer, and perinatal care. Significant conditions included, diabetes, mental health, tobacco use, vision health, respiratory disease, violence, injuries, nutrition, weight, and physical activity. Additionally, some patients did not have access to healthcare services. When this study was initiated Suffolk County performed low on unnecessary utilization and there was a variation in the region on preventable readmissions. SCC-DSRIP measured the most prevalent health conditions identified in the CHNA.61
1.3 Project Purpose

The purpose of this project is to perform a qualitative analysis of the SDH documented in Stony Brook Medicine’s Medicaid population health platform and identify SDH risk factors for preventative health appointments. Kane and Radosevich recommend using conceptual models prior to an outcomes study to form the basis of an analysis plan. Patient factors, the patient-centered medical home model, project intervention, anticipated outcomes, and their relationships will be explored (Figure 1).

Figure 1: Conceptual Model of Intervention & Outcomes

<table>
<thead>
<tr>
<th>Social Determinants of Health</th>
<th>Patient Factors</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Education</td>
<td>• Comorbidities</td>
<td>• Preventive health appointment compliance</td>
</tr>
<tr>
<td>• Nutrition</td>
<td>• Age</td>
<td>• Non-compliance with preventive health appointments</td>
</tr>
<tr>
<td>• Alcohol &amp; Substance Abuse</td>
<td>• Race</td>
<td></td>
</tr>
<tr>
<td>• Employment</td>
<td>• Gender</td>
<td></td>
</tr>
<tr>
<td>• Home Environment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Health Goals</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Sexual Information</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Computer &amp; Internet Access</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Intervention                  |               |          |
|                               | • Coordinated Care (PCMH) – SCC-DSRIP Program |

1.4 Hypotheses

The following hypotheses are proposed for this quantitative, correlational study:

- **H1a**: Social determinants of health are predictors of preventive health appointment compliance for Medicaid patients enrolled in the Patient Centered Medical Home (PCMH) model at Stony Brook Hospital.

- **H1b**: Social determinants of health are not predictors of preventive health visit appointment compliance for Medicaid patients enrolled in the Patient Centered Medical Home (PCMH) model at Stony Brook Hospital.
- **H2**: A statistical relationship exists between housing characteristics (i.e., computer and internet access) and Stony Brook Hospital Medicaid patients enrolled in the Patient Centered Medical Home (PCMH) model.

- **H0**: A statistical relationship does not exist between housing characteristics (i.e., computer and internet access) and Stony Brook Hospital Medicaid patients enrolled in the Patient Centered Medical Home (PCMH) model.

### 1.5 Need for Study

Chen et. al. findings from a systematic review have shown that studies used SDH documented in EHR’s as risk predictions for hospital 30-day readmissions, HIV risks and suicide attempt. Most SDH research has focused on screening for SDH and sifting through EHRs to identify unstructured social determinant of health data. Interestingly, there were no studies identified in the literature that specifically used the social history tab to gather SDH as predictors for preventative health appointment compliance for the Medicaid population. Additionally, community-level technology assessments (i.e., computer and internet connectivity) have been limited. Structured SDH data fields have been developed on the social history tab in Stony Brook Hospital's EHR, that are mapped to the Medicaid population health database. At the time of this study, the EHR was not used to capture or identify SDH, these data were collected on paper. A quality analysis of the Medicaid population health database and public health databases that store environmental information (e.g., education, computer, and internet access) will inform the Stony Brook Medicine community and contribute to the literature.
1.6 Definitions

- **Agency for Healthcare Research & Quality (AHRQ):** Agency responsible for building a safer health care system with a focus on evidence, equity, quality and health care that is affordable.\(^64\)

- **Consumer Assessment of Healthcare Providers & Systems (CAHPS):** AHRQ program developed for advancing the patient health care experience through survey distribution.\(^65\)

- **Community Health Record:** Is also referred to a community information system that contains population-level indicators associated with a geographic community, such as social, physical and lifestyle determinants that describe health and equity.\(^66\)

- **Electronic Health Record:** A longitudinal health record containing demographic and clinical information that is shared with healthcare providers in and across organizations.\(^31\)

- **Health Effectiveness Data and Information Set (HEDIS):** A tool used by health insurance plans in the United States to measure the performance of service and care.\(^67\)

- **Meaningful Use:** Developed by the federal government HITECH to identify EHRs that capture data elements defined by CMS for capture into the EHR.\(^32\)

- **National Center for Quality Assurance (NCQA):** An organization responsible for improving the quality of health care in the United States.\(^68\)

- **Population Health:** A concept used to describe a group of individuals and the distribution of health outcomes associated with the group.\(^69\)
- **Population Health Management**: Interventions that are developed to address a population or group and their associated risks. The goal is to slow the progression of health risks to improve health outcomes.\(^{69}\)

- **Preferred Provider Organization (PPS)**: Integrated delivery networks composted of providers and community-based organizations that is led by a safety-net provider or hospital. PPSs work together under a lead hospital that plans, implements and manages the DSRIP programs.\(^{57}\)

### 1.7 Theoretical Framework

Ecological and Socio-Technical theories are proposed in this project since the premise of this investigation seeks to identify SDH documented in electronic databases and how they influence behaviors of the Medicaid population. The Ecological theory emphasis is on understanding the relationships between people and social and environmental factors. While the socio-technical theory primarily encompasses the social and technical aspects of a system to determine usage. Both theories have been applied in health-related research and have evolved to include multiple interdependent levels/dimensions.\(^{70,71}\)

Ecological models have been used to explore links between SDH through the evaluation of community conditions using a singular approach. A multi-level model was developed by Sweat and Dennison enabling a more thorough investigation of patient-centered medical homes. The five levels of Sweat and Dennison’s theory align with the SDH of this study, individual behaviors and traits, relationships and social support, environmental conditions, structural (policies, laws, and politics); super structural (social justice and disadvantages related to socioeconomic, status, sexism). Scott and Wilson used this framework to identify social determinants in African Americans.\(^{72}\)
Sittig and Singh’s socio-technical theory was derived from Henrickson et al., Vincent et al., Carayon et al., and Harrison et al. and specifically designed to address the complexities of health information technology systems. This comprehensive multi-level theory is comprised of the following levels: processes, technology, culture, infrastructure, people, and effective communication. These dimensions are interdependent and interrelated and apply to the many challenges associated with the success of HIT projects. Sittig and Singh’s framework will guide this research through the identification of gaps related to socio-technical inefficiencies in reference to SDH electronic documentation and usability.73

1.8 Challenges & Limitations

The data elements for this study presents several challenges for analysis and addressing the hypothesis. Firstly, since it is not mandatory to document SDH, the electronic health record may not have this information stored consistently for each data element for the patient population. This study uses data stored in social fields and not in clinical notes. Secondly, it is unknown if the SDH are mapped accurately from the EHR to the population health platform. Thirdly, the limitation to using zip codes to evaluate environmental surroundings is that nomadic and homeless populations primarily use zip codes to receive their mail and may not reside at the addresses documented in the system. Additionally, how these data are collected will impact how they are reported in this study. Fourthly, the American Community Survey collects and stores data using surveys which are completed on a voluntary basis. Finally, this study will not specifically examine which methods were employed to reach health visit utilization.

1.9 Mitigation Plan

SDH data stored in the population health platform will be cross referenced with the EHR to ensure that data elements are accurately mapped. When possible, when the
same SDH data elements are pulled from different databases, they will be cross referenced to assess both consistencies and inconsistencies. Other limitations associated with inaccuracies and limited data availability will be reported in the study findings to inform future research. This is the first investigation that involves an analysis of SDH data on the social history tab in the EHR, therefore it serves as a groundwork for future studies to improve the documentation process and transition to whole-person-care.
2.1 Overview

Social Determinants of Health (SDH) have been documented on paper and in electronic health records (EHR) on forms and clinical notes, in a structured and unstructured format. The electronic storage of SDH is explored in this chapter to identify patterns in documentation practices that could help establish best practices for population health informatics. It investigates tools used to extract SDH from electronic health records and public databases, advantages and disadvantages, and studies that link SDH to health outcomes. Finally, studies that utilize SDH as risk factors for appointment compliance using machine learning tools are described. CINAHL, Google Scholar, PubMed, Scopus and Science Direct were searched to identify articles for this literature review. This review excludes studies conducted for individuals below the age of 18 or over 65 years.

2.2 Documenting Social Determinants in Electronic Health Records

The challenges associated with collecting SDH from underrepresented individuals include, creating questions that will capture SDH data needed for providing care, providing patients with a valid explanation for collecting SDH information, and developing strategies to increase patient comfort with sharing these data. Thus, studies in this section used validated survey tools to gather data from participants. The community characteristics were also similar consisting of a diverse groups of people with socioeconomic challenges, except for one study. However, SDH data collected varied based on the study and population. These studies used numerous databases to extract
data which has informed the research community with tested methodologies and sources of data.

Pinto et al. (2016) study in Toronto with the Centre for Addiction and Mental Health, Mount Sinai Hospital and Toronto Public Health had a focus on testing the feasibility of collecting SDH data from patients using a survey. The authors collected the data using tablet technology in patient waiting rooms which was fed directly into the EHR to share with providers of care. This was the only study in which surveys were password protected and linked to the patients’ electronic medical number and that posted SDH within 48 hours to the physicians’ inbox. Physicians were required to acknowledge in the system that they viewed SDH data. The surveys were administered by college students and participants were given an option not to respond to questions that they deemed personal or preferred not to answer. A total of 407 patients completed the Toronto survey. There were 5 participants of the 407 that stated that they felt uncomfortable responding to questions they felt were too personal. In addition, one person suggested adding gender to address queer and transgender individuals. The study neglected to track the number of people that declined to complete the survey and to save the language preferences for the participants, resulting in the need for future research on non-response bias.75

In 2018 Steiner and colleagues investigated older adults and food insecurities at Kaiser Permanente of Colorado (KPCO). This article was selected because of its association with SDH, the patient population does not represent the age range of this study. A total of 51,446 patients over 65 years of age completed at least one survey. Like Lofters et al, this study collected household information and housing status. Since the focus was on food insecurities in senior community, information was also gathered related to mental health conditions, social isolation, activities of daily living, alcohol and
tobacco use, physical activity, nutrition, and safety. The investigation used ICD-9 common conditions (e.g., diabetes, hypertension, and depression) and claims data to identify food insecurities and patient encounters. In addition, the BMI was calculated for all the participants. The survey was linked to administrative and clinical data at KPCO Institute for Health Research. Most patients that completed the survey were closer to 65 years of age, White and did not have multiple comorbidities. A total of 50,097 completed the food survey questions of which 2,859 or 5.8% reported that they had food insecurities. Patients that were 85 years and older that identified as Hispanic or Black were obese and insured by both the Medicare and Medicaid. Food insecurity was found to be more prevalent in woman than in men and in patients who presented at the emergency room, hospitals and admitted into a skilled nursing facility. Study limitations included missing data for income. Surveys were completed at well visits and food insecurity was assessed using a single question in the survey.76

Gold et al. research occurred at three community health centers at Oregon Community Health Information Network (OCHIN), the nation’s largest CHC network that utilizes Epic. This research builds upon recommendations from previous work with a focus on the integration of SDH data in the workflows of clinicians. There was an interest in acting upon the SDH captured through referrals in Epic, a popular EHR platform for hospitals in the United States. The socioeconomic status of the patient population was assessed prior to the study validating the need to collect SDH data. Comparable to Pinto et al., stakeholders in the community centers participated in this project. This was the first study to include referral specialists and community health workers. Survey tools included questions aligned with the NAM recommendations for SDH inclusion in EHRs and were also pulled from KP questionnaires. Surveys were distributed through patient portals making it convenient for patients to document SDHs.
Participants were not provided with the option to “refuse to answer.” Finally, a summary tab was created in the EHR that populated SDH data. Referrals were generated if there was a corresponding ICD-10 code. In a similar study conducted at the Boston Medical Center, adult patients at a primary clinic were screened for SDH and referrals were generated based on their social needs. The most common SDH identified were employment, food insecurity and problems paying for medications.

The authors that tested the surveys suggested that additional research be performed to validate the findings that were generalized based on the patient population. They concluded that future studies should give patients the option to “refuse” to answer questions, to decline services if a referral is going to be generated and develop processes to avoid duplicate entry. Additionally, it was suggested that a SDH box within the problem list be created to minimize issues related to tracking referrals. Lastly, an investigation of how to act upon SDH data collected, an evaluation of how SDH affect health, and further guidance on how to collect SDH data in clinic settings.

2.3 Mining Electronic Health Records

Electronic health records have been mined in studies to identify SDH that were available. The following tools have been used to extract these data: natural language process (NLP) or data mining, artificial intelligence, and algorithms. Though these methods have been applied to gather structured SDH, they have been used more frequently for unstructured data which requires the development of algorithms to search for words.

Oreskovic et al. study at Massachusetts General Hospital applied Queriable Patient Inference Dossier (QPID), a word recognition program, to extract psychosocial risk-factors from the EHR. Unlike the previous Massachusetts General Hospital, this study tested mining tools. Oreskovic et al. identified 22 search terms that aligned with
psychosocial risks for Medicaid patients. The mean number of terms found in the EHR for the Medicaid population was 14.1, compared to 6 for non-Medicaid (e.g., control group) patients. Extracting data using word recognition software was a much faster or robust process compared to the previous study that used a manual process to identify terms and then developed software to search for psychosocial risks.\textsuperscript{79}

Bejan et al. evaluated the use of word2vec and lexical associations to mine 100 million notes for words related to childhood experience (ACD) and homelessness at Vanderbilt University Medical Center. The mining tools used were effective, 70\% of ACE patients’ experienced sexual abuse, 50.6\% suffered from physical abuse; for homeless individuals 62.8\% did not have housing and 61.5\% used tobacco. Bejan et al. concluded that it was easier to extract homeless information. A total of 70\% of the homeless were men and 67.8\% of the patients with childhood experiences were women.\textsuperscript{80}

Feller et al. developed an algorithm to identify housing instability and drug use at New York Presbyterian-Columbia University Medical Center (NYPH-CUMC). They used three predictive and machine learning algorithms to search clinical notes to identify risk-factors for HIV patients. NYPH-CUMC three models identified baseline data using ICD codes associated with drug and non-drug dependence, emergency room visits, and drug induced mental disorders, inpatient visits, other joint disorders, diabetes, and benign neoplasm. STD infections, gender, fractures, and dislocations related to men having sex with men, cervical disorders, lymph and episodic disorders, past visits and unprotected sexual behaviors was used as baseline data. The following SDH information was identified from the databases as HIV risks: sexual orientation, sexual activities, previous tests or diagnosis, terms related to drug use, psychological comorbidities, and health utilization history.\textsuperscript{81}
2.4 Extracting Environmental Data Using Geospatial Technologies

Geographical Information Systems (GIS) are embedded with geospatial technologies that utilize zip codes from EHRs to map communities providing environmental data. The U.S. Census, Agency for Toxic Substances, and other local and state neighborhood databases that gather social and environmental information are all examples of Community Information Systems (CIS). These systems are frequently built and maintained by organizations with the ability to pull community related data from local and state public systems. The application of GIS technology to capture neighborhood characteristics offers context-informed information that is meaningful for providing effective services or whole-person-care.82

In 2011 Comer et al. created use cases for the Indiana Center of Excellence in Public Health Informatics for real-time integration of EHR and CIS systems. They had a specific interest in combining the Indiana Network for Patient Care (INPC), an EHR, with the SAVI Community Information System at Indiana University-Purdue in Indianapolis. SAVI has over 10,000 community characteristics such as, education, human services, housing demographics, public safety, and advanced functionality for using spatial technologies for providing community information in the state of Indiana. It is larger than any other CIS in the U.S. with data from more than 30 state, federal and local providers, capable of presenting data in an integrated format. INPC is Regenstrief Institute’s HIE that provides data at the point-of-care for over 14,000 physicians. The use cases provide a seamless, automated process for linking systems with large volumes of data, replacing the manual process which is complex and prone to error. According to the authors, the use cases are currently being tested.83

Dixon et al. built on Comer et al. study by integrating INPC & SAVI in 2015 to assess the prevalence of diabetes and neighborhood indicators associated with this
population. The study found that 8.9% of diabetes patients lived in neighborhoods in a metropolitan area; however, the authors do not describe how the integration process was performed (e.g., manual, or automated). Findings identified the following biases with using EHR zip code data: EHR data represents people that need health care; when linking EHR and CIS data the outcomes will be influenced by how the data was matched; and HIE data matching can create duplicate records and may contain larger quantities of data for low-income providers. Future research is needed to address these biases.\textsuperscript{84}

Bazemore et al. (2015) research was a collaboration with the Robert Graham Center (RGC), Oregon Community Health Information Network (OCHIN) and HealthLandscape to integrate CIS and EHRs. Unlike Dixon et al., the SDH was gathered from two systems, RGC and HealthLandscape. These systems store social, economic, and behavioral data from various state, local and national resources. They partnered with Federally Qualified Health Centers and linked U.S census data to patient zip codes. The EHR used in this project was the “Accelerating Data Value Access a National Community Health Center Network (ADVANCE).” A Community VS Geocoding API was developed that could accept address requests and assign geographic identifiers using HealthLandscape geospatial technology. Using this process, the authors were able to link the geographic information directly to the patients’ EHR. Data was transported to the EHR using a secure platform and data encryption to comply with HIPAA standards.\textsuperscript{82}

The API used in the ADVANCE project is currently undergoing additional testing. The authors suggested that additional research be conducted in SDH workflow integration and the use of public databases for community characteristics. They recommended that researcher’s partner with health care providers to identify best practices for storing SDH data in EHRs. Finally, they suggested studying the
development and application of clinical decision support tools to identify patients that live in a demographic region.\textsuperscript{82}

Biro et al. and Roth et al. research in 2016 builds on previous studies using geospatial technology. Their studies examined environmental factors surrounding communities where obese patients live. The CIS systems used in the studies were the Canadian census data and Nielsen PrimeLocation database at Ohio State University Wexner Medical Center (OSUWMC). The purpose of the Ohio study was to evaluate the neighborhoods of obese patients to determine the availability of fast-food restaurants, places that offered physical activity and to test algorithms for extracting data to link systems using the geocoding process. Canada’s research in 2016 was focused on determining the association of social determinants and obesity in the Kingston Ontario Practice Based Research Network using the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Both studies used a manual process to link the zip code data extracted from the EHR to match to CIS systems.\textsuperscript{85,86}

Roth et al. linked the percentage of unemployment, median household income, poverty level percentage, and education to 62,701 patient zip codes, ages 18-67. Data was extracted from the Nielsen PrimeLocation database that is populated with socioeconomic and population health information from the internet, mail, smartphone devices, barcode scanners and telephone surveys. Demographic data (e.g., height, gender, race, DOB, zip code, and weight) was pulled from the OSOWMC information warehouse.\textsuperscript{86} In contrast, Biro et al. study population consisted of 7153 patients’ ages 20 years and older. Demographic information was extracted from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). CPCSSN is used for monitoring the prevalence of chronic conditions in Canada. Unlike Roth et al, the data gathered in
CPCSSN came from various EHRs (e.g. ten primary care practices) related to different diseases.85

Biro et al. study found that 35% of obese patients lived in substandard areas when compared to adults in areas that were least deprived. Obesity and community deprivation varied for all age groups in the study; therefore, additional research will be required to understand the factors that influence health outcomes. Alternatively, Roth et al. study identified community characteristics such as fast-food restaurants, neighborhood composition, social deprivation, and socioeconomic composition of obese patients. This data could be used to initiate community changes that could potentially encourage exercise, nutritional foods, etc.85,86

Consistent with Dixon et al. findings, the previous studies were limited to the patients that were visiting their practices which could compromise the study for people in the neighborhoods that do not use local facilities. When patient populations are confined, the results from studies may vary depending on where the patient population was selected. In addition, EHR systems have missing data or non-standardized approaches for documentation that affect data quality and/or may limit data from answering the research question. There were also challenges with linking data from multiple sources resulting in unmatched data. While the cleaning of data is a standard process, it may not avoid potential errors that can influence the outcome of a study. Adding a control group may provide comparisons and outcomes that can further inform the community.84,85,86

2.5 Combining Data Collection & Geospatial Technologies

The Ontario, Canada study conducted by Lofters et al. included six clinics that use the medical home model for caring for its population. The purpose was to determine the association of SDH and colorectal, cervical and breast cancer screening. This
investigation differs because it combines self-reported surveys to collect SDH data and geospatial technologies to ascertain neighborhood characteristics of patients. Geospatial technologies compromise privacy and security enabling locations to be tracked and or sensitive information being shared. Survey results were captured on paper and/or electronically and then transferred to an EHR. Zip codes were linked to the 2006 Census in Canada. A total of 5766 patients completed the Ontario survey. Study participants were selected based on age, family histories, need and recommendations for colorectal, cervical or breast cancer screening. The following SDH data were collected: preferred language, sexual orientation, household income and numbers, immigrant status, ethnicity and housing status.87

Lofters et al. found that some survey questions were not answered by participants. Older patients that completed the survey were found to be more up to date with screenings, while younger patients were found to be more up to date with cervical cancer screenings. Patients were least likely to be updated on colorectal cancer (72.9%), as opposed to cancer screenings (78.7%) which could be associated with the colorectal procedure. Findings associated poverty or disadvantaged patients with limited or no cancer and breast cancer screenings. There was a significant association with income, housing, immigrant status and cancer screenings. People with a low-income level and housing instability were not as likely to screen for cancer. Study limitations include indigenous individuals not feeling comfortable with responding to questions. Additionally, similar to other studies, participants did not represent the entire community.87
2.6 Coding Social Determinants in Electronic Health Records

Blosnich et al. conducted the first study that utilized coded data to identify transgender individuals and SDH data in the Veterans Administrator (VA), an integrated delivery healthcare. The project evaluated medical disorders among transgender veterans to seek an understanding of the prevalence of SDH among these individuals. Medical and socio-demographic variables selected for the population were mood disorders, HIV, post-traumatic stress, hepatitis C, tobacco use, alcohol use, suicide disorder, housing instability, violence, and financial contention. Demographic variables for the study included race, marital status, and gender. Transgender individuals were identified with gender disorder in the VA system. Gender disorders and SDH were both documented using ICD-9 codes. SDH data was gathered from transgender individuals during clinical screenings using a questionnaire.88

Blosnich et al. found a direct association with housing instability and financial contention with all the medical disorders except for HIV. Violence was linked to all medical conditions except tobacco use and HIV, 48.1% of the transgender population had one or more SDH and 6.5% had three social determinants. The study further identified that 12.9% of the participants had financial challenges (30.8%) and housing instabilities (28.3%), compared to 19.5% of individuals involved in violence. Additionally, 22.5% used illicit drugs and 26.2% were addicted to alcohol, 43.7% used tobacco, and approximately 17.8% had codes related to suicide. A larger percentage of Blacks had housing instability and financial challenges compared to white transgender participants. Transgender women presented lower odds of housing challenges and financial limitations than males. The socio-demographic data distinguished individuals based on social determinants needs.88
Findings validate the importance of capturing social and environmental conditions in EHRs, because of the strong association with medical disorders, SDH and transgender populations. The limitations of the study include having a comparison group of non-transgender veterans to examine the difference between both groups. Most of the SDH collected came from self-reported surveys and gender was not known since the system did not offer sections to identify how these individuals switched their identities. There was a strong potential for codes to mistakenly not be associated with the initial documentation. For example, there could also be an underestimate of the ICD-9 codes used for documenting SDH. Finally, the VA does not have a classification for unmarried individuals that may be living together.88

Chambers et al. collaborated with the Bronx Community Network, Montefiore Medical Center and the City Department of Health and Mental Hygiene to conduct a geographic study on behavioral determinants of patients in Bronx neighborhoods. Survey questions were embedded in Montefiore Medical Centers (MMC) clinical records. MMC hospitals are in community settings that have FQHCs. The questions appeared as a popup when the vital signs were assessed or during the intake process for appointments or screenings in three health centers. Clinical staff could not bypass the screen forcing them to collect SDH data. The survey assessed dietary intake and physical activity for adult patients (18 and older). Neighborhood characteristics were evaluated using the US Census and zip code data.89

Findings from the Bronx study indicate that 35 – 43% of the individuals reported that they did not have physical activity in 30 days across all health centers. In the neighborhoods located near the health centers, 18-29% reported no activity in the past 30 days, approximately half of the patients documented they walked or biked for transportation in the past 30 days. The study revealed that there was a higher number of
patients in the clinics that reported unhealthy behaviors compared to the neighborhoods. Chambers et al. associated survey responses to neighborhood characteristics but assumed that the geographic locations could have had an impact on survey responses. Additionally, not all patients seen were asked questions which could reflect selection bias and/or clinician discretion during patient visits.89

2.7 Social Determinants of Health & Risk Stratification

Risk stratification is a process that utilizes claims data or data from EHRs to identify high and low-risk populations so that providers can seek out solutions for providing care that is both cost effective and efficient. Additionally, payment models that provide incentives for improving the quality of care can ensure that fairness is granted for all patient types and conditions for measurement and reimbursement. Traditionally, risk-factors were calculated based on medical care. With current policy changes there is a need to integrate SDH risk-factors into the quality equation.90 These data will assist providers with offering social services that could potentially improve health outcomes and reduce the rising cost of healthcare. There were several studies that focused on investigating SDH risk-factors.

Jamei et al. investigated the association of SDH for patients readmitted in Northern California Sutter Health system. Alcohol and tobacco use were found to be a risk for 30-day readmissions. The authors also gathered patient addresses from the EHR to match geographical locations using Google Geocoding Application Programming Interface (API). Unfortunately, the investigators failed to associate the neighborhood information found to the patient population.91 In two similar studies, the authors identified SDH as predictors for hospital readmissions. Low et al. findings associated economic status and treatment with antidepressants as predictors for hospital readmissions (e.g., three or more admissions within a year). Patients on anti-
depressants were found to be at a higher risk for readmissions. Watson et al. evaluated systems at Massachusetts General Hospital by searching for key terms using a computer program to extract psychosocial risk factors for patients with heart failure. Of the 729 patients included in the study, 93 were readmitted within 30-days. Patients with missed appointments, dementia or depression, adherence and refusal of treatment were more likely to be readmitted. Findings from these studies validate the association of behaviors with health outcomes and the use of machine learning algorithms for prediction analysis.

Alemi et. al. VA study used Naïve Bays and logistic regression, machine learning algorithms to assess if SDH were predictors of suicide and intentional self-injury. The SDH measured included homelessness, housing events, social isolation, arguments with spouse, changes in business/work, poverty, court involvement, family disruption, unemployed, school events, substance abuse, counseling, stress, and psychological trauma. In another VA study, Davis et. al. tested emergency room utilization using the following SDH in the electronic health record: housing instability, violence, unemployment, financial and legal problems, social or family issues, lack of access to transportation and psychosocial in the EHR. Both studies found that SDH were predictors for suicide and emergency room utilization.

2.8 Conclusion

Vulnerable populations represent a different group of people with different social and economic challenges. Within the population of the underrepresented are subpopulations that have unique SDH characteristics. Although the studies in this literature review were retrospective, they validate what has been known about the effects of social determinants for marginalized groups. Across most studies, participants that lived in substandard neighborhoods were poor, or had conditions that appeared to be
influenced by the neighborhoods they live in. Further research is necessary to identify the most prevalent risk-factors for different marginalized groups and associated health conditions. The association of risk-factors with health conditions can potentially lead to the development of interventions tailored to individuals. Finally, further research in the generation of referrals in EHRs, demonstrated by the Kaiser Permanente study, will help pinpoint the best interventions for SDH.

Researchers gathered data using technology and paper, both remotely and local to the facility through surveys and entry into the EHRs. Unfortunately, findings fail to conclude the best practices for collecting these data from patients or demonstrate the collection at-the-point of care by clinicians. Most surveys were distributed in patient waiting rooms and included questions related to participants comfort level with answering sensitive SDH questions. Overall, the survey responses were positive, but most failed to track survey non-response rates. As a result, the studies did not uncover the biases for using surveys. Collecting data real-time, directly into EHRs, will minimize errors and enable organizations to apply CDS tools to build reminders to update SDH information more efficiently.

A few studies demonstrated the advantages of coding SDH; however, most indicated the need for coding SDH in the EHR to simplify data extraction, referral generation and/or bill for SDH interventions. The Systematized Nomenclature of Medicine Current Terminology (SNOMED) includes codes for ethnicity, occupation, environment, and geographic location. Alternatively, the International Classification of Disease (ICD-10) includes codes for food insecurities, transportation, economic circumstances, housing, and employment. These codes do not represent all the SDH data elements recommended for capture in the EHR by NAM. Kaiser Permanente’s
research and the VA studies demonstrated the advantages of coding social determinants of health.96

Each study collected similar, but different social determinants of health information from participants. Interestingly, the most common SDH was related to the environment or housing using GIS technology. Table 2 has a list of all social determinants of health included in this literature review with associated health conditions for comparison analysis. It provides a visual of the gaps in research for future work.
Table 2: Summary of Social Determinants of Health Identified in Literature Review

<table>
<thead>
<tr>
<th>Author</th>
<th>Social Determinants of Health</th>
<th>Medical Conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alemi et al.</td>
<td>Homelessness, housing events, social isolation, arguments with spouse, changes in business/work, poverty, court involvement, family disruption, unemployed, school events, substance abuse, counseling, stress, psychological trauma[^94]</td>
<td>N/A</td>
</tr>
<tr>
<td>Bazemore et al.</td>
<td>Economics, workforce, population estimates, vital statistics, demographics, poverty, education, migration, healthcare quality indicators, physical and environment, mental health and substance abuse[^82]</td>
<td>N/A</td>
</tr>
<tr>
<td>Bejan et al.</td>
<td>Childhood experience, homelessness, and tobacco use[^86]</td>
<td>N/A</td>
</tr>
<tr>
<td>Biro et al.</td>
<td>Fast food restaurants, neighborhood composition, social deprivation and socioeconomic composition[^85]</td>
<td>Obesity</td>
</tr>
<tr>
<td>Chambers et al.</td>
<td>Dietary intake and physical activity for adult patients[^89]</td>
<td>N/A</td>
</tr>
<tr>
<td>Davies et al.</td>
<td>Housing stability, violence, unemployment, financial and legal problems, social and family issues, lack of transportation psychosocial[^95]</td>
<td>N/A</td>
</tr>
<tr>
<td>Dixon et al.</td>
<td>Neighborhood[^84]</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Feller et al.</td>
<td>Sexual orientation, sexual activities, previous tests or diagnosis, terms related to drug use, psychological comorbidities, and health utilization history.[^81]</td>
<td>HIV</td>
</tr>
<tr>
<td>Gold et al.</td>
<td>All of the social determinants domains recommended by NAM[^77]</td>
<td>N/A</td>
</tr>
<tr>
<td>Jamei et al.</td>
<td>Tobacco, alcohol, and drug use</td>
<td>Multiple comorbidities</td>
</tr>
<tr>
<td>Lofters et al.</td>
<td>Sexual orientation, household income and numbers, immigrant status, ethnicity and housing status[^87]</td>
<td>Colorectal, and breast cancer,</td>
</tr>
<tr>
<td>Low et al.</td>
<td>Mental health and low income[^92]</td>
<td>Multiple comorbidities</td>
</tr>
<tr>
<td>Oreskovic et al.</td>
<td>Psychosocial risk factors[^79]</td>
<td>N/A</td>
</tr>
<tr>
<td>Pinto et al.</td>
<td>Sexual orientation, income, housing[^75]</td>
<td>N/A</td>
</tr>
<tr>
<td>Author</td>
<td>Social Determinants of Health</td>
<td>Medical Conditions</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------------------------------------------------------------------</td>
<td>-----------------------------------------</td>
</tr>
<tr>
<td>Roth et al.</td>
<td>Income, unemployment, education, poverty level and population size[^6]</td>
<td>Obesity</td>
</tr>
<tr>
<td>Steiner et al.</td>
<td>Household, housing status, food insecurity, social isolation, activities of daily living, alcohol and tobacco use, physical activity, nutrition, and safety[^6]</td>
<td>Mental health</td>
</tr>
<tr>
<td>Vest, RV, Ben-Assuli O</td>
<td>Socioeconomic status, material conditions, tobacco, substance abuse, environment, health conditions (asthma, congestive heart failure, diabetes, depression), social circumstances (education, racial dissimilarity)^[97]</td>
<td>ED readmission rates (EHR &amp; HIE)</td>
</tr>
</tbody>
</table>

The authors used various databases to gather demographics and social determinants of health for GIS and survey studies. Data was extracted from disease registries, HIEs, community information systems, public and state databases. Thus, the EHR served as the primary resource for feeding data to most of these systems, except the public and state databases. The process of mapping zip codes to the neighborhoods with GIS is tedious and involves various steps. Interoperability appears to be the challenge with linking EHRs to CIS. The authors demonstrated the information garnered from geocoding provides a more in depth understanding of the social issues that may interfere with healthcare. The examination of different tools has the potential to standardize practices for gathering environmental SDH. Further research is needed to determine how geospatial technologies could feed relevant information to clinical systems seamlessly. Hence, secondary use of public databases or community information systems provides information that healthcare organizations may struggle to collect from patients using surveys.

Studies that gathered unstructured data demonstrated how complicated and time consuming it is to pull SDH from clinical notes. There was only one study that captured all the SDH recommended by NAM. Designing systems and processes to collect structured data simplifies report generation. The integration of systems presents an
opportunity for streamlining data collection and availability through interoperable systems. Connectivity will not limit how these data are collected but allow real-time updates to an integrated delivery systems database. Integrated delivery systems enable the sharing of both EHR data in both inpatient and outpatient venues. In addition to collecting SDH during patient encounters, technologies such as smartphones, iPads and patient portals can be used to capture “Real-time” information. Questions should be designed to track user concerns which will help inform organizations and providers for making changes to improve the data collection process remotely and at the point-of-care.
3.1 Overview

This chapter details the methods and procedures utilized to test the hypotheses for this study for a thorough understanding of the interpretation of results. It includes information related to research design, methodology, data sources, variables, statistical and machine learning procedures, and the pre-analysis process. The machine learning predictive models’ results were reported based on Lou et. al. recommendations.98

3.2 Subjects

The study sample included 911 adult patients and 12,118 outpatient encounters from a safety-net hospital system between July 1, 2017 – June 30, 2019. Patients were enrolled in the Suffolk Care Collaborative Delivery System Reform Incentive Payment (SCC-DSRIP) program in the Suffolk County region of New York. The inclusion criteria for patient demographics are males and females between 18-65 years of age. Subjects were required to have one inpatient or outpatient encounter, a zip code, and a minimum of one social determinant of health (SDH) variable documented in the electronic health record (EHR).

3.3 Research Design and Rationale

A correlational retrospective study was selected to examine the association of SDH and preventive health visit appointment compliance. According to Mertler & Reinhart, this design evaluates relationships among independent and dependent variables.99 This non-experimental investigation utilized secondary data and did not
recruit participants. Through statistical analysis, predictor variables were evaluated to assess changes on the dependent variable. If there is a difference identified on the dependent variable after analysis it can be assumed that the independent and dependent variables are associated.

Logistic regression observes the outcome of dependent variables which can be binary, nominal, ordinal or continuous. Binary logistic regression has two categorical dependent variables, multinomial has more than two categorical dependent variables; ordinal has two dependent variables that can be placed in logical order, and continuous variables are numeric and non-categorical. Binary regression was selected for this study because the dependent variables have two possible outcomes. Designs, such as multiple regression, path analysis, bivariate correlation, ANOVA, ANCOVA and MANCOVA were not selected because these methods are for dependent variables that are quantitative and discriminant analysis utilizes more than two categorical dependent variables.99,100

### 3.4 Data Sources

Secondary data was extracted from the population health database or data warehouse at Stony Brook Medicine. Additional resources were used to aggregate data to simplify the analysis of information gathered from large data sets.

#### 3.4.1 Population Health Database

Cerner’s HealtheIntent population health database at Stony Brook Medicine was used to store data for Medicaid patients that were enrolled in the SCC-DSRIP program. It receives medical record, encounter, and SDH information from the EHR which is linked through an interface.
3.4.2 U.S. Census Bureau

The EHR is unable to seamlessly capture environmental characteristics; therefore, the U.S. Census Bureau and the American Community Survey was used to identify neighborhood determinants using zip codes extracted from the HealtheIntent population health database.

3.4.3 Clinical Classification Software Refined (CCSR)

To simplify the grouping of SDH, health conditions and preventive health appointments, the Clinical Classifications Software Refined (CCSR) data set that aggregates the International Classification of Disease – Clinical Modification (ICD-10-CM) diagnosis codes was downloaded from the Agency for Healthcare Research and Quality website. The CCSR codes and descriptions were carefully examined to ensure they would align with the goals of this project. Table 3 lists social determinants of health, Table 4 health conditions, and Table 5 preventive health appointments with CCSR codes and their corresponding descriptions that were used to describe the results of this study.

<table>
<thead>
<tr>
<th>Conditions</th>
<th>CCSR Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol-related disorders</td>
<td>MBD017</td>
</tr>
<tr>
<td>Cannabis-related disorders</td>
<td>MBD019</td>
</tr>
<tr>
<td>Lifestyle/life management factors</td>
<td>FAC020</td>
</tr>
<tr>
<td>Malnutrition</td>
<td>END008</td>
</tr>
<tr>
<td>Nutritional anemia</td>
<td>BLD001</td>
</tr>
<tr>
<td>Nutritional deficiencies</td>
<td>END007</td>
</tr>
<tr>
<td>Other specified and unspecified Nutritional and metabolic disorders</td>
<td>END016</td>
</tr>
<tr>
<td>Opioid-related disorders</td>
<td>MBD018</td>
</tr>
<tr>
<td>Socioeconomic/psychosocial factors</td>
<td>FAC019</td>
</tr>
<tr>
<td>Tobacco-related disorders</td>
<td>MBD024</td>
</tr>
</tbody>
</table>

Source: CCSR/Clinical Classification Software Refined Description: social determinants of health
### Table 4: CCSR Health Condition Descriptions and Codes

<table>
<thead>
<tr>
<th>Conditions</th>
<th>CCSR Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>RSP009</td>
</tr>
<tr>
<td>Cardiac arrest and ventricular fibrillation</td>
<td>CIR018</td>
</tr>
<tr>
<td>Coronary atherosclerosis and other heart disease</td>
<td>CIR011</td>
</tr>
<tr>
<td>Diabetes mellitus with complication</td>
<td>END003</td>
</tr>
<tr>
<td>Diabetes mellitus without complication</td>
<td>END002</td>
</tr>
<tr>
<td>Essential hypertension</td>
<td>CIR007</td>
</tr>
<tr>
<td>Heart failure</td>
<td>CIR019</td>
</tr>
<tr>
<td>Hypertension with complications and secondary hypertension</td>
<td>CIR008</td>
</tr>
<tr>
<td>Pulmonary heart disease</td>
<td>CIR014</td>
</tr>
</tbody>
</table>

Source: CCSR/Clinical Classification Software Refined Description: social determinants of health

### Table 5: CCSR Preventive Health Appointments

<table>
<thead>
<tr>
<th>Conditions</th>
<th>CCSR Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encounter for observation and examination ruled out (excludes infectious disease, neoplasm, mental health disorders)</td>
<td>FAC003</td>
</tr>
<tr>
<td>Encounter for prophylactic measures (excludes immunizations)</td>
<td>FAC005</td>
</tr>
<tr>
<td>Neoplasm-related encounters</td>
<td>FAC008</td>
</tr>
<tr>
<td>Other specified encounters and counseling</td>
<td>FAC012</td>
</tr>
<tr>
<td>Contraceptive and procreative management</td>
<td>FAC013</td>
</tr>
<tr>
<td>Medical examination/evaluation</td>
<td>FAC14</td>
</tr>
<tr>
<td>Exposure, encounters, screening or contact with infectious disease</td>
<td>FAC016</td>
</tr>
</tbody>
</table>

Source: CCSR/Clinical Classification Software Refined Description: social determinants of health
3.4.4 Suffolk County Regions

Cities were grouped into regions with a master spreadsheet from Stony Brook Medicine consisting of the Suffolk County regions and cities (Table 6).61

<table>
<thead>
<tr>
<th>Region</th>
<th>Cities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core</td>
<td>Stony Brook, Mount Sinai, Miller Place, Coram, Rocky Point, Middle Island, Lake Grove, Port Jefferson Station, Shoreham, Nesconset, Saint James, Centereach, Sound Beach, Smithtown, Ridge, Kings Park, Hauppauge, Port Jefferson, Ronkonkoma, Selden, Setauket</td>
</tr>
<tr>
<td>Southwest</td>
<td>Great River, Islandia, Brentwood, Copiague, West Sayville, Bay Shore, Bohemia, Amityville, Central Islip, Brightwaters, Wyandanch. North Babylon, West Babylon, Islip, Deer Park, Lindenhurst, Oakdale, Babylon, Islip Terrace, West Islip, East Islip, Ocean Beach</td>
</tr>
<tr>
<td>Central South</td>
<td>Brookhaven, Yaphank, Medford, Bellport, Holtsville. Blue Point, Patchogue, Shirley, Farmingville, Bayport, Sayville, Holbrook</td>
</tr>
<tr>
<td>Northwest</td>
<td>Melville, Huntington Station, Cold Spring Harbor, Huntington, Centerport, Northport, Commack, Greenlawn, East Northport</td>
</tr>
<tr>
<td>Central East</td>
<td>Aquebogue, Upton, Speonk, Jamesport, South Jamesport, Remsenburg, Manorville, Moriches, East Moriches, Wading River, Riverhead, Calverton, East Quogue, Westhampton, Mastic, Center Moriches, Mastic Beach, Quogue, Eastport, Westhampton Beach</td>
</tr>
<tr>
<td>Northfork</td>
<td>Shelter island, Peconic, Laurel, East Marion, Greenport, Orient, New Suffolk, Southold, Mattituck, Cutchogue, Shelter Island Heights</td>
</tr>
<tr>
<td>Southfork</td>
<td>Sagaponack, East Hampton, Amagansett, Hampton Bays, Wainscott, Sag Harbor, Bridgehampton, Southampton, Montauk, Watermill</td>
</tr>
</tbody>
</table>

3.5 Data Variables

The dependent variables (DV) for this study results in an outcome of yes/no, or participants with (1) or without (0) a preventive health visit. The independent variables (IV) are categorical (e.g., nominal or ordinal) and continuous. Covariates examine the individual characteristics of patients (e.g., age, gender, race, and comorbidities) and related social determinants of health. The covariates are dichotomous, continuous and/or nominal (Table 7).
### Table 7: Independent and Dependent Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>IV/DV</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preventive health visits</td>
<td>DV</td>
<td>Binary/Dichotomous</td>
</tr>
<tr>
<td>Gender</td>
<td>Covariate/ IV</td>
<td>Binary/Dichotomous</td>
</tr>
<tr>
<td>Medical Comorbidities</td>
<td>Covariate/ IV</td>
<td>Nominal</td>
</tr>
<tr>
<td>Race</td>
<td>Covariate/ IV</td>
<td>Nominal</td>
</tr>
<tr>
<td>Age</td>
<td>Covariate/ IV</td>
<td>Continuous</td>
</tr>
<tr>
<td>Education</td>
<td>IV</td>
<td>Ordinal</td>
</tr>
<tr>
<td>Nutritional health</td>
<td>IV</td>
<td>Ordinal</td>
</tr>
<tr>
<td>Alcohol abuse</td>
<td>IV</td>
<td>Categorical</td>
</tr>
<tr>
<td>Employment</td>
<td>IV</td>
<td>Ordinal</td>
</tr>
<tr>
<td>Home environment (ACS-Fact Finder)</td>
<td>IV</td>
<td>Nominal</td>
</tr>
<tr>
<td>Patient health goals</td>
<td>IV</td>
<td>Nominal</td>
</tr>
<tr>
<td>Sexual information</td>
<td>IV</td>
<td>Nominal</td>
</tr>
<tr>
<td>Substance abuse</td>
<td>IV</td>
<td>Categorical</td>
</tr>
</tbody>
</table>

### 3.6 Procedures

MS Excel and Python (version 3.7.6) were used to preprocess, organize, aggregate, analyze, and graph data. A cluster analysis was performed to discover how health conditions were distributed amongst the population. Classification algorithms, from the Scikit-learn library (version 0.24.0) were selected to build models. Finally, machine learning models were evaluated using various metrics.

### 3.6.1 Pre-Analysis & Data Organization

Data screening is a critical step for quantitative studies. According to Mertler and Reinhart, multivariate research requires data to be examined carefully for errors, missing data, and outliers to ensure that assumptions of the study can be addressed to prevent analyzing incorrect values. The following six datasets were extracted from Stony Brook Medicine’s HealtheIntent data warehouse for years 3 and 4 of the SCC-DSRIP program:

1. encounter (268,284 records, 121 columns)
2. condition (961,570 records, 78 columns)
3. person demographics (65,704 records, 63 columns)
4. person race (43,890 records, 16 columns)
5. person (45,512 records, 103 columns)
6. cohort (268,284 records, 1 column)

Each data set was inspected for duplicate records and encounters, missing values, erroneous data, similarities, and differences to determine how they would be manipulated individually and collectively. The cohort data set was eliminated from the study because the data elements were consistent with the encounter data set. Columns from the remaining data sets were used to identify patient demographics, social determinants of health, comorbidities, ICD-10-CM codes, and encounter information.

To identify the population, the demographic data sets were merged (person, demographics, person race and person) on the MPI number, duplicate records, and records with missing values for race, age, gender, and zip code were eliminated. The patients’ birthday was converted to age and records of patients between the ages of 18 – 65 were extracted. Regions were mapped using the zip code column on the demographic data set to simplify data aggregation and analysis by grouping cities into Core, Southwest, Central South, Northwest, Central East, Northfork, and Southfork regions (Table 4).

Services that were not associated with the study (e.g., inpatient, perioperative, emergency services, and pre-post op services) were removed from the encounter data set and service years 3 and 4 of the DSRIP program were identified. To simplify grouping of patient conditions and SDH, the Clinical Classifications Software Refined (CCSR) International Classification of Disease ICD-10-CM codes were merged with the condition data set. Lactose intolerance unspecified code (E73.9) and anemia due to blood loss (D50.0) were removed because they were classified as social determinants of health.
Finally, billing codes were eliminated since they were duplicate encounters created to bill patients.

The patient characteristics (e.g., gender, race), region, preventive health visit information and social determinant of health domains or predictor variables (alcohol abuse, education, nutritional health, employment, patient health goals, sexual information, and social abuse) and the corresponding CCSR descriptions were all coded with numerical variables for data analysis. Figure 2 provides a high-level overview of the process steps, for data aggregation and elimination.

**Figure 2: Data Wrangling Process Steps**

- Merged race, person, and person demographics datasets = 66,687 demographic encounters
- Removed duplicate records using MPI number = 42,761 demographic encounters
- Removed records with missing values = 41,031 demographic encounters
- Merged race, person, and person demographics datasets = 66,687 demographic encounters
- Removed non-outpatient visits from encounter dataset = 13,480 encounters
- Merged regions to dataset to zip code = 38,643 demographic encounters
- Removed duplicates from condition dataset = 218,047 condition encounters
- Identified number of patients with one SDH in condition dataset = 911 final medical records
- Merged unique records and demographics datasets = 922 demographic medical records
- Removed billing codes and missing dates = 12,118 final encounters
- Removed (E73.9 and D50.0) from condition dataset = 15,748 encounters
- Merged unique records and condition datasets = 15,790 encounters
3.6.2 Feature Engineering and Selection

To prepare data for supervised and unsupervised machine learning algorithms, CCSR diagnosis codes and descriptions were written to JSON (JavaScript Object Notation), then the DictVectorizer was used to transformed categorical data to feature vectors creating extra columns to indicate the presence (1) or absence (0) of a CCSR category. The newly coded data was converted to a data frame to build models. As shown in Figure 3 patients had multiple CCSR categories with multiple 0/1 values. To address multicollinearity or collinear relationships between independent variables, features were evaluated using Pearson correlation. Features highly correlated were removed from the final analysis. Finally, Scikit-learn SelectKBest, a technique that uses chi-square to compute non-negative features and target classes was applied to identify the highest performing features.

![Figure 3: Feature Vectors](image)

3.6.3 K-means Clustering Algorithm

CCSR patient diagnosis codes were mined using the k-means clustering algorithm to partition unlabeled data with similar characteristics. Principal component analysis (PCA) was applied prior to building clusters to reduce dimensions and improve performance. The number of principal components was examined to identify a predetermined threshold of total variance. Principal components were then used to construct the elbow method to select the optimal number of clusters (k centers) for k-
Finally, features were joined to demographic information (age, race, gender, and region) to examine structures and describe the communities within larger groups. Using a set of observations \((x_1, x_2, x_3, \ldots, x_n)\), the k-means algorithm separates data points into clusters \((C)\) described by the mean \((\mu_j)\) of the observations. The algorithm selects centroids that minimize inertia (sum-of-squares) within each cluster. New centroids are created using the mean value of observations assigned to each centroid and the differences between the first and second centroid values are computed. The algorithm repeats these steps until the centroids are no longer moving significantly.

\[
\sum_{i=0}^{n} \min_{\mu_j \in C} \left( \| x_i - \mu_j \| ^2 \right)
\]

3.6.4 Classification Algorithms

In a systematic review logistic regression, classification trees, random forests, artificial neural networks, and support vector machines were found to be used more frequently for binary classification studies. Srinivas and Salah study identified random forests, stochastic gradient boosted trees, and deep neural networks as popular techniques for binary variables that are used in health care for length of stay, mortality, and disease risks. Naives bayes and k-Nearest neighbors were also identified as effective binary algorithms. Therefore, in addition to logistic regression, random forests, k-Nearest Neighbors, and support vector machine classification algorithms were selected for comparison analysis.

Binary classification is subject to imbalanced classes potentially leading to bias of the majority class and reduced model performance. Synthetic Minority Over-Sampling (SMOTE) is a popular technique for balancing imbalanced classes using interpolation to
generate synthetic balanced samples in the majority class. It has been used in various studies and there are different variations of SMOTE that have been developed to improve the performance of imbalanced classes.\textsuperscript{109,110,111} This technique will be applied to address class imbalance and a comparison analysis will be conducted on both balanced and imbalanced datasets.

Independent variables (X) and response vectors (y) were defined and randomly split into testing (.30) and training sets (.70) to learn the relationship between X and y. The train/test size was selected based on the sample size of the dataset. Singh et al. found that model performance improved for small samples when the testing size was increased above standard sizes (25/75, 20/80).\textsuperscript{112} The target variables had a predetermined threshold of $p = 0.05$, values $p > 0.05$ are positive predictions (1), while values $p < 0.05$ are non-predictive values (0). Features were scaled using Scikit-learn standard scaler to standardize data and improve model performance.

3.6.5 Logistic Regression (LR)

Logistic regression is linear parametric algorithm that is widely used for binary classification models. It utilizes probabilities and odds ratios to determine if an event will occur. The $\ell_2$ regularization penalty for logistic regression mitigates overfitting and improves model performance. When sample sizes are biased or small using the maximum-likelihood approach to logistic regression results in overfitting or inaccurate class predictions.\textsuperscript{112} $\ell_2$ penalized logistic regression for binary classes utilizes a cost function and ordinary least of squares. Where X and Y are dependent and independent variables and (i) are data points. Lambda ($\lambda$) is multiplied by the two norms squared of theta ($\theta$) to penalize or shrink coefficients.\textsuperscript{113}
Random Forest (RF)

Random Forest (RF) models have become increasingly popular because of its ability to handle noise enabling it to process imbalanced datasets. It is a non-parametric method that generates decision trees to classify labeled objects. The n-estimator parameter enables the selection of trees resulting in improved accuracy minimizing overfitting. RF uses the Gini index to identify the Gini of each branch on a node; it determines which branch is more likely to occur. Predict_proba is represented by P_{mk} and C is the number of unique classes.\textsuperscript{104,114}

\[
Gini = \sum_{i=1}^{C} P_{mk}(1 - P_{mk})
\]  

K-Nearest Neighbors (KNN)

The K-Nearest Neighbors classifier (KNN) is a non-parametric algorithm that identifies k data points that are nearest to other unknown data points using distance functions. It counts the number of neighbors that belong to a class (0/1) and then assigns classes based on uniform weights. The k parameter plays a significant role in the performance of the KNN algorithm because it indicates the number of neighbors used for estimation. The Euclidean distance metric calculates the distances between two points and then assigns the point to a class or nearest neighbor. A (X_1, X_2, X_3, \ldots, X_m) and B (Y_1, Y_2, \ldots, Y_m) are feature vectors and m is the dimensionality of the feature space. To calculate the normalized Euclidean metric the following equation below.\textsuperscript{104,105,115}
\[ d(A, B) = \sqrt{\frac{m}{\sum_{i=1}^{m} (x_i - y_i)^2}} \]

### 3.6.8 Support Vector Machine (SVM)

Support Vector Machines (SVM) have become increasingly popular. Similar to RF, it can handle noise and process imbalanced datasets. It is a non-parametric model that separates data into a binary class using the maximum distance between the two classes and their data points. It classifies data through hyper-planes which is a straight line that separates 2-dimensional space into two halves.\textsuperscript{104,108} It maximizes the distance between different classes also known as margin maximization. The overall goal with SVM is to find \((w \in \mathbb{R}^p)\) and \((b \in \mathbb{R})\) so that the prediction given by \(\text{sign}(w^T \phi(x) + b)\) is correct for most samples given training vectors \((X_i \in \mathbb{R}^p, i = 1, ..., n),\) in two classes, and vector \((y \in \{1, -1\}^n)\).\textsuperscript{116}

\[
\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i
\]

subject to \(y_i (w^T \phi(x_i) + b) \geq 1 - \xi_i\)

\(\xi_i \geq 0, i = 1, ..., n\)

### 3.6.9 Parameter Selection

Scikit-learn K-Fold cross validation is a technique widely used to identify the best performing parameters prior to building models. K-Fold cross validation estimates parameters on new data and delivers an accuracy score for model evaluation by repetitively splitting dating into folds. It trains all folds except for one that is used for validation. Classification models will be estimated using 10 folds for optimal performance.\textsuperscript{117,118} Random forest will be estimated with 50 ensemble trees and k-
nearest neighbors 30 nearest neighbors. Logistic regression will be estimated with a penalty of \( \ell_1 \) and \( \ell_2 \), liblinear solver which is good for small datasets, and C parameters (1, 5, 10) to improve performance. Support vector machines will be estimated using a radial bias function (RBF) kernel, with gamma set to “auto,” and C values (1, 5, 10, 20). The (C) is a tuning parameter that softens SVM margins for a better fit.\(^{102}\)

### 3.6.10 Evaluation of Classification Models

Model evaluation was executed using the confusion matrix, receiver operating character (ROC), area under the curve (AUC), accuracy score, precision, recall, specificity and F1 score. A confusion matrix categorizes binary classes into true positives (TP), true negatives (TN), false positives (FP), or false negatives (FN). TP, TN, FP, and FN are reported as percentages, 100% being the optimal score. The ROC curve is a graphical plot of sensitivity (TP) verses specificity (FP) at different thresholds. The AUC targets values above and below the probability threshold, the ideal AUC score is equal to 1.0 (100%).\(^{104,108}\) Predictive algorithms have a higher AUC. ROC-AUC is the most reported performance metric for prediction models.\(^{119}\)

Accuracy evaluates model performance by calculating the number of correct predictions.\(^{104,108}\)

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]

Precision scores are used to measure the correctly classified predictions and the total number of true positives and negatives. It helps with understanding the positive predicted values.\(^{104,108}\)

\[
\text{Precision} = \frac{TP}{TP + FP}
\]
Sensitivity is also known as recall. It is the ability of a test to maximize true positives, or correctly identify positive predictions.\textsuperscript{104,108}

\[
\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}
\]  \hspace{1cm} (9)

F1 score brings together the total of positively and correctly positive predicted values; it represents the harmonic mean of both precision and recall.\textsuperscript{104,108}

\[
\text{F1 Score} = \frac{2 + \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]  \hspace{1cm} (10)

Specificity is a measure that can classify true negatives, with ROC it can show the relationship to minimize false positives.\textsuperscript{104,108}

\[
\text{Specificity} = \frac{TN}{FP + TN}
\]  \hspace{1cm} (11)

3.6.11 Pearson’s Chi-Square Test

The chi-square test for independence tests the relationship between categorical variables. Comparisons were made with age groups, gender, and race for clusters 0-3 that were developed with the k-means clustering algorithm. A contingency table calculated the expected (E) and observed frequencies (o) to determine if the demographic variables and clusters were independent. The degrees of freedom (\(df = r - 1/c - 1\)) was calculated which approximates the number of categories with a .05
significance level.\textsuperscript{104} Metrics were evaluated based on 95% confidence level or p-value < .05.

\subsection*{3.6.12 Silhouette Coefficient Score}

The silhouette coefficient evaluated the compactness or distance between k-means clusters. The optimal value for silhouette score is 1 and least optimal is -1. Values that are closer to 0 indicate overlapping in clusters. Negative values can be interpreted as values being assigned to the incorrect cluster.\textsuperscript{108,120}

\subsection*{3.6.13 Environmental Analysis}

The zip codes in the final data set were entered into the American Community Survey (ACS) to identify the educational level of the communities. Additionally, computer and internet access were retrieved from the Unites States Census Bureau and reported based on the zip codes of the population.
Chapter IV
RESULTS

4.1 Overview

This chapter describes the study results for participants enrolled in the SCC-DSRIP program between 7/1/2017 – 6/30/2019. The results show the characteristics of the population across age groups, gender, race, social determinants of health, comorbidities, service, and geographic location. It provides a list of the social determinants of health captured in the electronic health record for patients and how they are distributed based on age, race, region, and gender. Furthermore, the patient conditions and the predictive models for the correlational analysis of appointment compliance and social determinants of health are presented.

4.2. Patient Demographics

There were 911 subjects enrolled in the SCC-DSRIP program, 64.7% females and 35.3% males (Figure 4). The median age for the population was 46 years and the average was 45 years. The average age for males was 45.6, the median was 49 years. The average age for females was 44.7, the median was 46 years. Patient care for subjects in this study was rendered in the following outpatient practices in the Suffolk County region of New York: Medicaid General Medicine (22%), General Medicine Primary Care (13.6%), Family Medicine Primary Care (35.3%), Outpatient Offsite Location (26.5), and Pediatrics (2.8%) offices.
4.2.1 Age Distribution of Population

*Figure 5* has a breakdown of the population by age groups. The 18-29 age group had the least number of subjects, 62% (77) females and 38% (48) males. The 30-41 group was composed of 68% (166) females and 32% (79) males. Ages 42-53 had 69% (167) females and 31% (74) males. Finally, ages 54-65 had the most subjects, 60% females (179) and 40% (121) males.
4.2.2 Population by Race

There were seven different racial groups, White and Black/African American represented the largest groups. A total of 632 identified as White and 106 as Black/African Americans. There was a small population of Asians (27), American Ind/Alas Nat (7), Hispanic (2) and Nat Haw/Pac Islanders (1). Females represented 44.2% (403) of White subjects, 7.6% (70) Black/African Americans, 1.7% (16) Asians, .54% (5) Amer/Alask, .1% (1) of Hispanic and Nat/Haw .1%. Males represented 25% (229) of Whites, 3.8% (35) Black/African Americans, 1.2% (11) Asians, .21% (2) American/Alaskans, and .1 % (1) of Hispanics. Subjects are not racially diverse, given the small percentage of minorities in study. Additionally, a significant number of people did not identify their race (Figure 6).
4.2.3 Population by Region

Figure 7 describes the concentration of subjects in the Suffolk County regions. A complete list of the cities within each region is in Table 4. Overall, most people occupied the Core (432) and Central South (286) regions. A small percentage occupied the Central East (99), Southwest (79), North fork (2), Northwest (11) and Southfork (3) regions.

![Figure 7: Population Region Totals](image)

4.3 Education by Geographic Region

The education level (bachelor's degree/higher) is summarized by region and city to further understand the SDH of the population for adults 25 years and older. Emphasis is placed on the largest racial groups in this study (White and Black/African Americans), who were concentrated in the Central South and Core regions. A total of 32.5% of the Central South and 42.8% of the Core regions population reported having a degree. In the Northwest and Southfork region, unlike other regions, over 50% of the people reported having a bachelor’s degree/higher. The Central South and Southwest regions had the least amount of people with a college education (Figure 8). According to the United States Census Bureau, Whites represent 84.2%, Black/African Americans 8.8%, American Indian and Alaskan Native 0.6%, Asian 0.1%, Native Hawaiian and
Other Pacific Islanders 2.0% and Hispanics 20.2%, Two/More Races 2.0% of the population in Suffolk County, New York.\textsuperscript{121,122}

![Figure 8: Region Totals, Bachelor’s degree/higher](image)

**4.3.1 Core Region**

The Core region was composed of 47.3% of the population, 42.8% of the population that live in the region reported their education. As shown in Figure 9, Stony Brook is comprised of more people with a college degree than any other city, Whites represent 86.8% and Blacks 2.3% of its population. In Middle Island, 27.3% of the people had a bachelor’s degree/higher where Whites represent 83.4% and Blacks 12.4% of the population. Additionally, Selden where 26.2% of the population have a bachelor’s degree/higher, Whites represent 82.2% and Blacks 4.7% of the overall population. Ridge had the lowest number of people with a college degree, 7% of its population is Black and 87.1% is White.\textsuperscript{121,122}
4.3.2 Central South Region

A total of 31.4% of the people in this study lived in the Central South region, 32.5% of the people that live in the region reported their education. As shown in Figure 10, there were more people with a college education in Sayville than any other city. Sayville is comprised of 94.7% Whites and 0.9% Black/African American compared to Bellport where Whites represent 67.5% and Blacks 23.1% of the population. In Shirley, Whites represent 85.7% and Blacks 7.1% of the population. Bellport and Shirley have the lowest number of people with a bachelor’s degree/higher. Interestingly, Bellport has the largest population of Blacks and 23.1% of people have a bachelor’s degree/higher. In Medford Whites represent 85.4% and Blacks 9.3% of the population and 25.9% reported having a bachelor degree/higher.\textsuperscript{121,122}
4.3.3 Central East Region

A total of 10.9% of the people in this study lived in the Central East region, 39.4% of the people that live in this region reported their education. Upton had the most educated population; however, only 7 people reported having a degree because it is a very small town. Quogue had the most educated population; East Quogue is composed of 97% White and .4% is Black. In Mastic where only 14.9% of the people have a bachelor’s degree/higher, Whites represent 84%, and Blacks 9.8% of the population. In Riverhead 71.4% of the population is White and 18.8% is Black/African American and 20.8% of the people reported having a bachelor’s degree/higher (Figure 11).\textsuperscript{121,122}
4.3.4 Northfork Region

In this study, 2% of the people lived in the Northfork region; however, 47.2% of the population in this region reported their education. This region is composed of 95.3% Whites and 2.1% Black/African American, it is in eastern Long Island. There is a large population of educated people in this region compared to other regions (Figure 12).121,122

Source: American Community Survey120

Figure 11: Central East Region City Totals, Bachelor's degree/higher

Figure 12: Northfork Region City Totals, Bachelor's degree/higher
4.3.5 Northwest Region

A small population (1.2%) of people in this study lived in Northwest region, 56.1% of the people that live in the region reported their education. This region had a significant number of people with college degrees. Cold Spring Harbor (95% Black, 0.6% Black) and Centerport (96% White, 0.5% Black) were the most educated cities and East Northport and Huntington Station were the least (Figure 13).121,122

![Figure 13: Northwest Region City Totals, Bachelor’s degree/higher](image)

Source: American Community Survey122

4.3.6 Southfork Region

Comparable to Northfork and Northwest regions, only .3% of the people in this study lived in the Southfork region. Most of the towns in the Southfork regions are in Southampton (Figure 14). This area is composed of 85.3% Whites and 5.7% Blacks.121,122
4.3.7 Southwest Region

A total of 8.7% of the people in this study lived in the Southwest region, 32.9% of the population that live in this region reported their education. Great River (97% White, 0.5% Black) and Brightwaters (94% White, 1.5% Black) had the most educated populations and Brentwood (63.3% White, 8.6% Black) and Central Islip (52.6% White, 25.1% Black) had the least.\textsuperscript{121,122}
4.4 Computer and Internet Access

Table 8 and Table 9 provide a snapshot of the computer and internet access for the Core and Central South regions. Cities in the Core and Central regions had strong internet connectivity and computer access (over 90%). There were only three cities below 90% for computer and internet access, Ridge and Selden (Core region) and North Bellport (Central South region). These cities also have a lower percentage of people with a bachelor’s degree/higher.  

<p>| Table 8: Core Region Computer and Internet Access |</p>
<table>
<thead>
<tr>
<th>City</th>
<th>Computer</th>
<th>Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Centereach</td>
<td>94.6%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Coram</td>
<td>92.9%</td>
<td>89.6%</td>
</tr>
<tr>
<td>Hauppauge</td>
<td>96%</td>
<td>92.6%</td>
</tr>
<tr>
<td>Kings Park</td>
<td>92.3%</td>
<td>89.3%</td>
</tr>
<tr>
<td>Lake Grove</td>
<td>90.3%</td>
<td>87.8%</td>
</tr>
<tr>
<td>Middle Island</td>
<td>90.5%</td>
<td>86.4%</td>
</tr>
<tr>
<td>Miller Place</td>
<td>98.1%</td>
<td>94.6%</td>
</tr>
<tr>
<td>Mount Sinai</td>
<td>92.9%</td>
<td>90.6%</td>
</tr>
<tr>
<td>Nesconset</td>
<td>92.7%</td>
<td>90.3%</td>
</tr>
<tr>
<td>Port Jefferson</td>
<td>92%</td>
<td>90.2%</td>
</tr>
<tr>
<td>Port Jefferson Station</td>
<td>92%</td>
<td>90.2%</td>
</tr>
<tr>
<td>Ridge</td>
<td>84.9%</td>
<td>81.1%</td>
</tr>
<tr>
<td>Rocky Point</td>
<td>93.3%</td>
<td>90.1%</td>
</tr>
<tr>
<td>Ronkonkoma</td>
<td>96.6%</td>
<td>93.8%</td>
</tr>
<tr>
<td>St. James</td>
<td>93.5%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Selden</td>
<td>89.3%</td>
<td>85.5%</td>
</tr>
<tr>
<td>Smithtown</td>
<td>93.6%</td>
<td>90.7%</td>
</tr>
<tr>
<td>Sound Beach</td>
<td>91.1%</td>
<td>88.2%</td>
</tr>
<tr>
<td>Stony Brook</td>
<td>94.9%</td>
<td>93.5%</td>
</tr>
<tr>
<td>Wading River</td>
<td>92.8%</td>
<td>86.4%</td>
</tr>
</tbody>
</table>

Source: U.S. Census Bureau
### Table 9: Central South Region Computer and Internet Access

<table>
<thead>
<tr>
<th>City</th>
<th>Computer</th>
<th>Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayport</td>
<td>94.7%</td>
<td>90.9%</td>
</tr>
<tr>
<td>Blue Point</td>
<td>96.5%</td>
<td>95.8%</td>
</tr>
<tr>
<td>Brookhaven Town</td>
<td>92.8%</td>
<td>89%</td>
</tr>
<tr>
<td>Farmingville</td>
<td>95.4%</td>
<td>88.5%</td>
</tr>
<tr>
<td>Holbrook</td>
<td>95.8%</td>
<td>93.2%</td>
</tr>
<tr>
<td>Holtsville</td>
<td>93.7%</td>
<td>91.1%</td>
</tr>
<tr>
<td>Medford</td>
<td>94.9%</td>
<td>91.2%</td>
</tr>
<tr>
<td>North Bellport</td>
<td>89.5%</td>
<td>82.5%</td>
</tr>
<tr>
<td>Patchogue Village</td>
<td>93.2%</td>
<td>85.1%</td>
</tr>
<tr>
<td>Sayville</td>
<td>92.5%</td>
<td>87.1%</td>
</tr>
<tr>
<td>Shirley</td>
<td>94.1%</td>
<td>90.2%</td>
</tr>
<tr>
<td>Yaphank</td>
<td>92.4%</td>
<td>88.7%</td>
</tr>
</tbody>
</table>

Source: U.S. Census Bureau

#### 4.5 Patient Comorbidities

*Figure 16* summarizes some of the major comorbidities for males and females. Females suffered more from essential hypertension (10%) and diabetes mellitus without complication (18.4%) than males. Males had a larger percentage of diabetes mellitus with complications (18.7%) and diabetes mellitus without complication (15%). There was a low percentage of males and females with hypertension with complications, secondary hypertension, heart failure, pulmonary heart disease, and cardiac arrest and ventricular fibrillation. The 54–65-year age group had the most chronic conditions compared to other age groups (*Figure 17*).
4.6 Preventive Health Appointments

Preventive health appointments included screenings, physical exams, preoperative exams, physical exams, medical clearances, immunizations, counseling, and birth control. There were 394 preventive health appointments identified. A large majority of appointments were for medical examination/evaluation (43.4%) and exposure, screenings or contact with infectious disease (34.5%). The appointment types, the percentage of appointments for each appointment type, and the corresponding CCSR codes and descriptions are shown in Figure 18.

![Figure 17: Chronic Diseases by Age Group](image1)

![Figure 18: Preventive Health Appointment for Population](image2)
4.6.1 Characteristics of Appointments

Females had 332 appointments, compared to males who had 62. There were more people with appointments in the 30-41 age group (139), than the 18-29 (58), 42-53 (82) and 54-65 (115) age groups. The White (263) population had more appointments than Black/African Americans (51). The Hispanics and Native/Hawaiian did not have appointments and the Asian (11) and American Indian/Alaskan Natives (3) had the least number of appointments. Most people that had appointments were from the Central South (126) and Core (173) regions. The Southwest (51), Central East (41) and Northwest (3) had the least amount of people with appointments and there were no appointments for people in the Northfork and Southfork regions. The SDH that were prominent for subjects with appointments were nutritional anemia (73), nutritional deficiencies (123), other nutritional and metabolic disorders (53), and tobacco related disorders (87).

4.7 Social Determinates of Health

A total of 2,883 SDH were documented in the EHR on the social history tab. Nutritional deficiencies, other specified and unspecified nutritional and metabolic disorders, tobacco-related disorders, and nutritional anemia were the most common; while lifestyle/life management, malnutrition, and socioeconomic/ psychosocial factors were the least common (*Figure 19*). Most of the CCSR categories were consistent with the SDH descriptions in the HealtheIntent database. There were a few exceptions, lifestyle/life management issues were related to unprotected sex and tobacco abuse; socioeconomic/psychosocial factors were associated with stressful life events affecting family.
The z-codes identified in the system included, z63.79 (other stressful life events affecting family); z72.51 (unprotected sex), z68.38 (obesity), and z72.0 (tobacco abuse disorder) in the system. Notably, unprotected sex and tobacco were documented twice, and stressful life events and obesity were documented once.

4.7.1 Characteristics of Social determinates by Gender

Females outnumbered males in social determinants by age group and race out of 2,883 SDH documented in the system. Females smoked more tobacco (61%), suffered with more nutritional deficiencies (73.9%), opioid-related disorders (66.7%), alcohol-related disorders (54.1%), nutritional anemia (80.8%), other nutritional and metabolic disorders (61.4%), while males consumed more cannabis-related substances (62.8%). There were a limited number of lifestyle/life management factors documented in the system for males and females, and females had 1 socioeconomic/psychosocial factor and males had zero (Table 10).
Table 10: Social determinants of health by gender

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age Groups</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-29</td>
<td>143(39.8%)</td>
<td>216(60.1%)</td>
</tr>
<tr>
<td>30-41</td>
<td>231(29.1%)</td>
<td>561(70.8%)</td>
</tr>
<tr>
<td>42-53</td>
<td>159(23.3%)</td>
<td>524(76.7%)</td>
</tr>
<tr>
<td>54-65</td>
<td>385(36.7%)</td>
<td>664(63.3%)</td>
</tr>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>736(34.8%)</td>
<td>1377(65.2%)</td>
</tr>
<tr>
<td>Black</td>
<td>68(21.5%)</td>
<td>249(78.5%)</td>
</tr>
<tr>
<td>Asian</td>
<td>20(22.5%)</td>
<td>69(77.5%)</td>
</tr>
<tr>
<td>Amer/Alas</td>
<td>5(45.5%)</td>
<td>6(54.5%)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>1(50%)</td>
<td>1(50%)</td>
</tr>
<tr>
<td>Nat/Haw</td>
<td></td>
<td>2(100%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>88(25.2%)</td>
<td>261(74.8%)</td>
</tr>
<tr>
<td><strong>SDH</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alcohol-related disorders</td>
<td>83(45.9%)</td>
<td>98(54.1%)</td>
</tr>
<tr>
<td>Cannabis-related disorders</td>
<td>27(62.8%)</td>
<td>16(37.2%)</td>
</tr>
<tr>
<td>Lifestyle/life management factors</td>
<td>1(25%)</td>
<td>3(75%)</td>
</tr>
<tr>
<td>Malnutrition</td>
<td>3(27.3%)</td>
<td>8(72.7%)</td>
</tr>
<tr>
<td>Nutritional anemia</td>
<td>40(19.1%)</td>
<td>169(80.8%)</td>
</tr>
<tr>
<td>Nutritional deficiencies</td>
<td>123(26.1%)</td>
<td>348(73.9%)</td>
</tr>
<tr>
<td>Other nutritional and metabolic disorders</td>
<td>102(38.6%)</td>
<td>162(61.4%)</td>
</tr>
<tr>
<td>Opioid-related disorders</td>
<td>38(33.3%)</td>
<td>76(66.7%)</td>
</tr>
<tr>
<td>Socioeconomic/psychosocial factors</td>
<td></td>
<td>1(100)</td>
</tr>
<tr>
<td>Tobacco-related disorders</td>
<td>90(39%)</td>
<td>141(61%)</td>
</tr>
</tbody>
</table>

Source: CCSR/Clinical Classification Software Refined Description: social determinants of health

4.7.2 Characteristics of Social determinates by Race

White and Black/African American subjects had more social determinants of health documented electronically than any other racial group (Table 11). Asians, American Indian/Alaskan Natives, Hispanics, Native Hawaiian/Pac Islander had the least amount of SDH documented. There was a significant number of participants that were not willing to share SDH that were classified in the “Unknown” category.
4.8 Cluster Analysis

The dataset for the cluster analysis contained 911 patients and 301 columns.

Principal component analysis reduced the features from 301 to 2 principal components.

The number of components identified at 95% variance for the principal component analysis was 5.74 which equated to 148 (5.75) eigenvectors. The 148 eigenvectors were used to plot the Inertia plot which identified 4 clusters for the k-means clustering analysis (**Figure 20**).

![Inertia Plot](image.png)

**Figure 20: Inertia Plot**
4.8.1 Cluster Scatter Plot

The scatterplot shown in *Figure 21* illustrates how health conditions were clustered, indicating some overlapping and data points belonging to more than one group. There was a significantly larger spread in clusters 0 and 3 making these clusters less distinct than clusters 1 and 2. This was consistent with the silhouette score of .025 which indicates features crossing boundaries and being a part of more than one cluster. The distribution of health conditions in clusters were as follows: cluster 0: 12.4% (113 patients), cluster 1: 27.6% (252 features), cluster 2: 46.7% (426 features), and cluster 3: 13.1% (120 features).

![Figure 21: Clusters](image)

4.8.2 Health Conditions

The top 10 health conditions identified in the cluster analysis by MPI number (911) and not encounters (12,118) are shown in *Figure 22*. These conditions include some of the social determinants of health (e.g., nutritional deficiencies, nutritional and metabolic disorders, tobacco-related disorders, and nutritional anemia) that are discussed in section 4.5. Nutritional deficiencies, obesity, musculoskeletal pain, and other nutritional and metabolic disorders were most prominent.
The sections that follow will explain how the most prominent health conditions in Figure 22 were distributed amongst age groups, race, gender, and region in each cluster (Table 12).

<table>
<thead>
<tr>
<th>Health Condition</th>
<th>Cluster 0</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety and fear-related disorders</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aplastic Anemia</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipid metabolism</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Musculoskeletal pain, not in lower back</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Nutritional and metabolic disorders</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nutritional deficiencies</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Obesity</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spondylopathies/spondyloarthropathy</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Tobacco-related disorders</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Source: CCSR/Clinical Classification Software Refined Description: social determinants of health

4.8.3 Cluster Health Conditions by Gender

Figure 23 illustrates how health conditions were distributed by gender in clusters 0-3. Females outnumbered males in all health conditions. In cluster 0 both males and females suffered most from depression and anxiety and fear-related disorders. In cluster 1 they had more conditions related to nutritional deficiencies, obesity, and disorders of
lipid metabolism. In cluster 2, the most prominent health conditions for males and females were nutritional and metabolic disorders, obesity, tobacco-related disorders, and aplastic anemia. Finally in cluster 3, health issues were related to spondylopathies/spondyloarthropathy and musculoskeletal pain, not in lower back. A chi-square test of independence for gender and clusters yielded, $p = 0.675$ indicating that there was no significant relationship between the variables.

![Figure 23: Gender (Clusters 0-3)](image)

### 4.8.4 Cluster Health Conditions by Age Group

*Figure 24* illustrates how health conditions were distributed among age groups in clusters 0-3 (*Table 12*). The 54–65 age group had the most health conditions and the 18-29 group had the least. In cluster 0 health conditions were prominent for all age groups except 18-29, the 54-65 and 42-53 groups suffered more from health conditions. Cluster 1 health conditions were prominent for age groups 30-41, 42-53 and 54-65. In cluster 2, the 18-29 group had the least amount of health conditions, while all other age groups had the most. Finally, in cluster 3 the 18-29 group had the least amount of health conditions compared to the other groups.
A chi-square test of independence for age-groups and clusters yielded, \( p = 0.987 \) indicating no significant relationship between variables.

**4.8.5 Cluster Health Conditions by Race**

*Figure 25* illustrates how health conditions were distributed among racial groups in clusters 0-3. As previously mentioned, White and Black/African Americans represented most of the population and had the most health conditions compared to other racial groups. A chi-square test of independence for race and clusters yielded \( p = 0.998 \) indicating no significant relationship between the variables.
4.8.6 Cluster Health Conditions by Region

*Figure 26* illustrates how health conditions were distributed among geographical regions in clusters 0-3. The Core and Central South regions had the most health conditions compared to other regions. A chi-square test of independence for region and clusters yielded a $p = 0.291$ indicating no significant relationship between the variables.

![Clusters by Region](image)

*Figure 26: Region (Clusters 0-3)*

4.9 Classification Models

To build classification models to make future predictions, diagnostic and service dates were streamlined to one appointment date for each encounter. Diagnostic and service dates that were the same (1898) were removed from the final encounter numbers (12,118). Service dates that were prior to the diagnosis date (8726) and duplicate encounters (1,103) were removed. There was a significant variance between the two classes on the remaining 392 features, 8% (31) appointments and 92% (368) non-appointments. The imbalanced classes are shown in *Figure 27*. A statistical mean of the features variables shows that some feature variables were extremely weak predictors (*Figure 28*). Most of the SDH appear to be strong predictors of appointment compliance except lifestyle/life management, malnutrition, and socioeconomic/psychosocial disorders (*Figure 29*). Since the binary class was significantly imbalanced SMOTE was used to balance the minority class.
Figure 27: Model Sample Data Appointments

Figure 28: Statistical Mean of Features
4.9.1 Features

A Pearson Correlation test found that there was no correlational relationships with the feature variables selected for inclusion in the classification models, all values were closer to -1. The top 10 features identified by SelectKBest are shown in Table 13, tobacco-related disorders, lifestyle/life management, opioid related disorders, and socioeconomic/psychosocial factors were excluded. After reviewing the mean scores and testing different
features, the final features selected for inclusion were age groups, race, nutritional deficiencies, malnutrition, gender, other nutritional and metabolic disorders, tobacco-related disorders, nutritional anemia, lifestyle/life management disorders, alcohol-related disorders, opioid-related disorders, and cannabis-related disorders.

<table>
<thead>
<tr>
<th>Features</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region</td>
<td>254</td>
</tr>
<tr>
<td>Alcohol-related disorders</td>
<td>30</td>
</tr>
<tr>
<td>Other nutritional and metabolic disorders</td>
<td>8.54</td>
</tr>
<tr>
<td>Race</td>
<td>7</td>
</tr>
<tr>
<td>Nutritional anemia</td>
<td>6.6</td>
</tr>
<tr>
<td>Nutritional deficiencies</td>
<td>4.8</td>
</tr>
<tr>
<td>Malnutrition</td>
<td>4.0</td>
</tr>
<tr>
<td>Age groups</td>
<td>3.3</td>
</tr>
<tr>
<td>Gender</td>
<td>1.5</td>
</tr>
<tr>
<td>Cannabis-related disorders</td>
<td>1.3</td>
</tr>
</tbody>
</table>

**4.9.2 Random Forest (RF) Feature Importance**

RF feature importance is calculated using the mean decrease in impurity (MDI) which ranks features by level of importance. It’s pivotal to understanding how features are distinguished and how they contribute to the performance of a model. Surprisingly, age, gender, tobacco-related disorders, nutritional anemia, and nutritional deficiencies were the most important features for appointment compliance (*Figure 30*).
4.9.3 Classification Model Validation and Accuracy Scores

Optimal parameters were selected using the mean performance of cross validation. The random forest parameters were set to 40 ensemble trees achieving a validation score of .89. Logistic regression parameters were set to l2 with a c-value constraint of 1.0 using the liblinear solver which received a validation of .92. Support vector machines achieved a .92 validation score using the RBF kernel and (C) misclassification penalty coefficient or tuning parameter of 1.0. Finally, k-nearest neighbors scored a .92 with k = 10 neighbors. These scores were consistent with the accuracy scores of the Imbalanced dataset. SMOTE’s accuracy scores were significantly lower. Random forests had the highest score (89%), k-nearest neighbors had the second, (87%), and logistic regression and support vector machines were the lowest (84%).

4.9.4 Random Forest (RF) Model Performance

Figure 31 show the ROC curves for the Imbalanced and SMOTE models. The RF Imbalanced AUC performance was 581, compared to SMOTE’s 599. Both models performed poorly with no significant differences. The Imbalanced model had 0 predictions and 9 false negatives with a precision, recall and f1-score of 0. While SMOTE had 1 prediction and 8 false positives with .20 precision, .11 recall, f1-score .14. Both models had over 85% true negative predictions and 2-9 false positives and false negatives (Figure 32).
4.9.5 Logistic Regression (LR) Model Performance

Figure 33 show the ROC curves for the Imbalanced and SMOTE models. The LR Imbalanced AUC performance was 627 compared to SMOTE’s 644 which performed better. The Imbalanced model had 0 predictions and 9 false negatives with a precision, recall and f1-score of 0. While SMOTE had 3 appointment predictions and 6 false positives with .20 precision, .33 recall, f1-score .25. Both models had over 80% true negative predictions, however, SMOTE had .17% false positives and the Imbalanced model had 0% and both models had false negatives (Figure 34).
4.9.6 K-Nearest Neighbors (KNN) Model Performance

Figure 35 show the ROC curves for the Imbalanced and SMOTE models. The KNN Imbalanced AUC performance was 732 compared to SMOTE’s 743, both models performed well. Performance improved with the balanced model. The Imbalanced model had 0 predictions and 9 false negatives with a precision, recall and f1-score of 0. While SMOTE had 1 appointment prediction and 8 false positives with .12 precision, .11 recall, f1-score .12. Both models had over 80% true negative predictions, however, SMOTE had 5.9% false positives and the Imbalanced model had 0%. Both models had 8-9 false positives (Figure 36).
4.9.7 Support Vector Machines (SVM) Model Performance

Figure 37 shows the ROC curves for the Imbalanced and SMOTE models. The SVM Imbalanced AUC performance was .671 compared to SMOTE’s 656. Model performance decreased with SMOTE. The Imbalanced model had 0 predictions and 9 false negatives with a precision, recall and f1-score of 0. While SMOTE had 3 appointment predictions and 6 false positives with .20 precision, .33 recall, f1-score .25. Both models had over 80% true negative predictions, however, SMOTE had 10.1% false positives compared to the Imbalanced model of 0%. Both models had 6-9 false positives (Figure 38).
4.9.8 SMOTE Precision and Recall Curves

The precision and recall curves AUC scores were extremely low reflecting the precision and recall scores for SMOTE models. Logistic regression performance was slightly higher than the KNN, RF and SVM.
4.9.9 Summary of Metrics

*Figure 40* show improvement in the ROC-AUC performance for most models using the *SMOTE* technique. KNN achieved the highest performance for both the *Imbalanced* (732) and *SMOTE* balanced (743) models and RF (599) had the lowest. SVM performed better with the Imbalanced dataset with a higher ROC-AUC (.671). While accuracy scores were above 90% for all models except RF for the Imbalanced model, there were no appointment predictions. Unlike the *Imbalanced* models, *SMOTE* was able to make predictions for all models. However, *SMOTE*’s accuracy scores were significantly lower (below 90%). LR and SVM had the most predictions (3), but they were met with many FP and FN (type I and type II errors) and extremely low metric scores (precision, recall, f1).

4.9.10 Probability of Appointments and Predictions

*Table 41* summarizes the probability of appointments for each model. All the models projected the probability of appointments (threshold of *p* > .05). The logit summary for the LR models shown in Figure 42 show nutritional anemia (*p* = 0.006) and tobacco-
related disorders ($p = 0.02$) as predictors of appointment compliance; as well as gender ($p = 0.03$) and age groups ($p = 0.04$).

![Histogram of Probabilities for Classification Models](image1)

**Figure 41:** Histogram of Probabilities for Classification Models

![Logit Summary](image2)

**Figure 42:** Logit Summary
Chapter V

DISCUSSION

5.1 Overview

The primary goal of this study was to investigate through model analysis if social determinants of health could predict appointment compliance, determine if there was a statistical relationship with computers, the internet, and population, and to conduct a qualitative analysis of SDH. The purpose for pursuing this research was the shift to population health informatics, advanced technologies and federal policies encouraging whole person care and a passion for improving health equity for marginalized groups. This chapter focuses on interpreting the results of the study. It provides a summary of what was demonstrated through using machine learning algorithms and the SMOTE oversampling technique. Furthermore, it includes an analysis of the strength of the study compared to previous work.

5.2 Model Predictions

The four supervised algorithms explored, compared, and analyzed in this study demonstrated that machine learning tools could be used to predict a minimum of 1 and a maximum of 3 appointments on a dataset that was balanced using SMOTE oversampling technique. Tobacco-related disorders and nutritional anemia were found to have a relationship with preventative health appointments. Non-clinical determinants associated with appointments included age and gender. However, the predictions were met with low precision, recall, f1 scores, false positives, false negatives, and true negatives. Tobacco is a common SDH that has been identified in the literature, Jamei et al. investigation found that tobacco was a risk factor 30-day readmissions.91
5.3 Exploratory Analysis Findings

The cluster analysis revealed that there were a significant number of people with nutritional deficiencies, nutritional anemia, other nutritional and metabolic disorders which were reported in the social determinants of health section of this thesis. Interestingly, it identified additional SDH that were not documented on the social history tab (e.g., obesity, depression, and anxiety and fear-related disorders). Because this analysis was done on unique records, some of the more common comorbidities did not surface in the cluster analysis. An analysis of health conditions based on encounters identified that males and females had other health conditions such as diabetes, hypertension, coronary atherosclerosis, asthma, etc. Findings from a systematic review by Walker et. al. describes the impact that SDH has on glycemic control and blood pressure.123

5.4 Patient Social Determinants of Health

The strength of this study compared to prior studies is that it identified structured SDH on the social history tab in the population health database. Most of the previous research focused on sifting through large datasets to find SDH information using natural language processing techniques in EHRs. Out of 2,883 social determinants of health found in the system only 11 had z-codes documented which is consistent with previous research. The expansion of z-codes with granularity is necessary to simplify the research process and build decision support tools that could assist providers in making a comprehensive clinical diagnosis based on the whole-person care model.

Consistent with previous studies, alcohol, tobacco, malnutrition, obesity, and drug use (cannabis and opioid), unprotected sex, depressive disorders (mental health specifically), and other stressful life events affecting the family was identified in the system. Although the social history tab has fields to capture education, employment and patient health goals, there was nothing documented. While there was some information
related to lifestyle and life management, and socioeconomic and psychosocial factors it was limited. Based on NAM recommendations and this analysis some of the SDH that were not found in the EHR/population health database were as follows: education, financial resource strain, exposure to violence or intimate partner violence, neighborhood, median-household income, and stress validating the challenges to collect these data from patients.

5.5 Racial Demographics of Patient Population and Algorithm Bias

One of the most significant observations identified was the difference in the racial profile of the population and in the Suffolk County region. The White population was significantly larger than any other group in this study. They had more social determinants and conditions documented than any other racial group. Blacks/African Americans are clustered in cities that have a lower education level (e.g., Riverhead, Bellport, Ridge, Islip, etc.). Education is a social determinant of health that has had a negative impact on the health of individuals and populations. Additionally, in some of these locations there was a weaker signal for internet connectivity and computer access. Since algorithms learn from the data that it is fed, populations or groups that are outnumbered can potentially result in significantly biased results.

5.6 Limitations

There were several study limitations. Firstly, the exclusion of ethnic groups from my IRB which may have reduced the diversity of the subjects significantly. Secondly, several encounters had to be eliminated to reduce the model sample to one medical appointment per record for model analysis, which resulted in an imbalanced dataset. Since this was a retrospective study, data elements could not be changed to improve model performance. Thirdly, this study only evaluated four models, different models and parameters could have potentially produced better results. Fourthly, A statistical analysis of environmental data collected from the American Community
Survey or U.S. Census Bureau could not be performed because the data collected was aggregated based on zip code. Finally, the validation of data mapping from the EHR to the population health database for social determinants of health could not be performed because reports in this study were only generated from the population health database.
Chapter VI

CONCLUSIONS AND FUTURE RESEARCH

6.1 Study Summary

This study demonstrates the importance of conducting a routine evaluation or assessment of the social determinants of health that are being captured and stored in electronic health databases and/or population health platforms. Similar efforts can be used to improve the overall electronic documentation of social determinants of health. The study validates the benefits of utilizing machine learning tools to determine if social determinant of health are risk factors for appointment compliance, etc. and exploring data to identify patterns with patient conditions and social determinants of health. Finally, the study confirmed how algorithms can create bias when the predominant race is White and there is a low percentage of minorities represented in the study.

6.2 Future Research

Additional studies using SDH as predictors will validate the findings in this study. More efforts are needed to target vulnerable populations that are segregated in regions and cities that may not have access to the resources that are available in predominantly White neighborhoods in the Suffolk County region. Identifying how algorithms can be used to reduce bias is necessary in studies with a significantly smaller number of minority patients in predominantly White regions or neighborhoods. Further research on social determinants of health using machine learning tools will help establish the evidence that is required to improve best practices in population health informatics.
References


5. Closing the gap in a generation Health equity through action on the social determinants of health.


8. A Framework for Educating Health Professionals to Address the Social Determinants of Health.; 2016. doi:10.17226/21923


92


29. Blumenthal D. spe ci a l r e p or t Implementation of the Federal Health Information Technology Initiative. Published online 2011.


40. Gottlieb BL, Tobey R, Cantor J, Hessler D, Adler NE. Integrating Social And Medical Data To Improve Population Health: Opportunities And Barriers. Published online 2017.


94


55. York State Department of Health N. A Plan to Transform the Empire State’s Medicaid Program.


60. Plan PC, Plan PC, Name PPS. DSRIP PPS Primary Care Plan. Published online 2016.

61. 2014 Community Needs Assessment Suffolk County, New York Delivery System Reform Incentive Program (DSRIP). Published online 2014.


78. de la Vega PB, Losi S, Martinez LS, et al. Developing an Emr-Based Screening and Referral System To Address Social Determinants of Health (Sdoh) in Primary Care: a Feasibility Study. Med Care. 2018;33(2).

79. Oreskovic NM, Maniates J, Weilburg J, Choy G. Optimizing the Use of Electronic Health Records to Identify High-Risk Psychosocial Determinants of Health. JMIR Med Informatics. Published online 2017. doi:10.2196/medinform.8240


97


100. Tabachnick BG, Fidell LS, Ullman JB. *Using Multivariate Statistics.*


111. SMWireless Sensor Networks Intrusion Detection Based on SMOTE and the Random Forest Algorithm _Enhanced Reader.


