2022

A Transdisciplinary Emergent Approach for Systems and Interventions (EASI)

Chaithanya Renduchintala
University of Central Florida

A TRANSDICIPLINARY EMERGENT APPROACH FOR SYSTEMS AND INTERVENTIONS (EASI™)

by

CHAITHANYA RENDUCHINTALA

M.S., University of Central Florida

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the School of Modeling, Simulation and Training
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Fall Term
2022

Committee Chair: Varadraj P. Gurupur
Thomas T. H. Wan

ABSTRACT

In modeling human behavior and social structures several factors can emerge over time this can be attributed to the availability of new data, increased complexity, changes to the organizational structure, interventions, introduction of innovative technology or services and due to improved knowledge and treatments. We hypothesize a new class of emergent decision support systems that continually evolve to account for this "Causal Drift". In this work, we illustrate the application of the Emergent Approach to Systems and Intervention (EASI™) methodology with the example of Community Intervention Activity Model (CIAM) to reduce the rate of diabetic hospitalization at the local/ county level. A key contribution of this work is the design of an efficient theoretically informed emergent data collection system. A second key contribution of this work is that it offers practitioners a methodology to systematically determine data that needs to be collected and then model the collected data. Thus EASI™ methodology supports the efficient capture of data that has utility in decision making. To ensure applicability of this work publicly available Behavioral Risk Factor Surveillance System (BRFSS) and Social Vulnerability Index (SVI) data sets have been utilized. The EASI™ method has four significant advantages: a) the prediction is based on theoretically informed causal structure; this allows it to be used as a basis for evaluation of interventions as opposed to deep learning and other machine-based structure learning methods which are susceptible to spurious associations, b) existing data is utilized to evaluate clinical relevance of predictions, c) leveraging high dimensional synthetic observational health data to model health objectives, and d) provides guidance on transformation of system from the emergent basis to the new emergent system as new knowledge is gained. The dissertation proposes, implements, and evaluates the EASI™ methodology as applied to a CAIM for the reduction in diabetic hospitalizations.

ACKNOWLEDGEMENTS

I wish to thank and recognize the contribution and guidance of the rest of my committee without whose support it would not have been possible for me to complete this dissertation. Roger Azevedo PhD., Christian King, Ph.D., Alexandra Xan C.H. Nowakowski, Ph.D.and Michael A. Xynidis, Ph.D.

I will be remiss if I do not acknowledge the invaluable and timely support of Dr. Charles Hughes in ensuring I complete my dissertation in a timely manner. Dr. Barbara Fritzsche for her support and guidance. I want to also thank Kirsten Seitz for all her help over the last couple of years.

I would not have started on this journey without the motivation and encouragement from Dr. Peter J Kincaid. I express my deep gratitude to my former supervisor Gordon Hogan.

There are so many other people who have helped me restore my health, find companionship and friendships though this journey. To each of you I owe a debt of gratitude for your presence in my life.

I want to thank my mother, Nirmala Prabhu and my father, Prabhu Kumar for their encouragement through this journey.

Last but not least my doggie KRISH who is a constant reminder that it is possible to remain happy and excited regardless of circumstances as long as there is a little bit of play every single day!

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES xi

CHAPTER ONE: INTRODUCTION 1

1.1 Emergence in Healthcare 1

1.2 Overview of the Problem 3

1.1.1 Significance of the Problem 5

1.3 Research Questions 5

1.3.1 General Hypothesis 6

1.4 Overview of the Dissertation Research 7

1.5 Summary 7

CHAPTER TWO: LITERATURE REVIEW 10

2.1 Community Health Improvement Plan (CHIP) Implementation 10

2.2 Causal Modeling in Health Care 15

2.3 Inference of Causality and Causal Structure 17

2.3.2 Structure Learning in Bayesian Networks 18

2.3.3 Advances in Causal Bayesian Modeling 18

2.4 Synthetic Data Generation 19

2.4.1 Generating Synthetic Datasets for Healthcare Modeling using GANs 19

2.5 Public Datasets in Community Health Planning 20
2.5.1 BRFSS Data Set for Community Health 20
2.5.2 SVI data set for community heath 21
2.5.3 Mimic iii Hospitalization Data 22
2.6 Modeling Approaches Applicable to Diabetic Hospitalization 22
2.6.1 Linear Regression 23
2.6.2 Least Absolute Shrinkage and Selection Operator (LASSO) Regression 24
2.6.3 Ridge Regression 25
2.6.4 K-Nearest Neighbors (KNN) 25
CHAPTER THREE: METHODOLOGY 28
3.1 Emergent Approach to Systems and Interventions (EASI™) 31
3.1.1 EASI™ Methodology Modules 32
3.1.2 EASI™ Methodology Algorithm Explained 32
3.2 EASI™ Process Applied in Community Health Improvement Implementation 36
3.3 EASI – Diabetic Hospitalization Health Objective Improvement model 37
3.5 EASI Synthetic Data Generation For Diabetic Hospitalization Simulation 41
3.6 EASI™ Data Collection Model for Diabetic Hospitalization (Outcome 1) 45
3.6.1 Dataset Synthesis for selected Community Health Improvement Plans 45
3.6.2 System Design Criteria 45
3.7 EASI Data Collection Model and Design Methodology 47

3.8 EASI Methodology Summary 47

CHAPTER FOUR: DATA ANALYSIS AND RESULTS 49

4.1 Diabetic Hospitalization – Length of Stay MIMIC-iii Data Set.....49

4.1.1 Data Set Description 50

4.1.2 Application of Linear Regression 50

4.1.3 Application of LASSO Regression 51

4.1.4 Application of Ridge Regression 53

4.1.5 Application of K-Nearest Neighbors (KNN).....54

4.1.6 Application of Multinomial Logistic Regression.....55

4.2 Diabetic Hospitalization with BRFSS Data Analysis 57

4.2.1 Description of the Key predictors and response variable in BRFSS 59

4.2.2 BRFSS Exploratory Data analysis 63

4.2.3.1 Unsupervised Machine Learning Models 72

4.3.2.2 Supervised Machine Learning Model 75

4.3. Data Collection System Design 82

4.4. Data Collection System Implementation as a Relational Model 83

CHAPTER FIVE DISCUSSSION, CONCLUSIONS AND FUTURE WORK 85

5.1 Key Contributions.....85

5.1.1 Contribution to Modeling 86

5.1.2 Contribution to Simulation 88

5.1.2.1 Contribution 1: Utilization of Synthetic Data for Intervention Performance Modeling 88

5.1.2.2 Contribution 2: Utilization of Public Data Sets such as BRFSS and SVI in Intervention Planning.....88

5.1.3 EASI Methodology contribution to localized Diabetic Hospitalization Intervention Models 89

5.2 Limitations 90

5.2.2 Limitation 2: Synthetic Data Generation.....91

5.2.3 Limitation 3: Data Modeling 92

5.2.4 Limitation 4: Interpretability and Intervention Design.....92

5.3 Comparison with most closely related work.....92

5.3.1 Comparison with Health as a System Model.....92

5.3.2 Comparison with GAIN: Missing Data Imputation using Generative Adversarial Networks 93

5.4 Direction of Future Work 94

REFERENCES 96

LIST OF FIGURES

Figure 1: overview of EASI(TM) methodology - data preparation, synthesis, collection and modeling process 7

Figure 2: CHIP development process 11

Figure 3: EASI(Tm) Methodology Key Outcomes 29

Figure 4: Overview of EASI(TM) Methodology 32

Figure 5: Algorithm EASI Model Development and Validation procedure 35

Figure 6: EASI(TM) Methodology Application in Community Health Improvement 36

Figure 7: Data hierarchy representation in Care Planning for chronic diabetics. 38

Figure 8:Figure 8: EASI(TM) Evidence Incorporation Model – Methodology 39

Figure 9: Representational Framework for Health Objective Specific Casual Model Generation using EASI(TM) Methodology 40

Figure 10Synthetic Data Generation with BRFSS and SVI data for selected census tracts 41

Figure 11 Brute Force - Constraint and Counters Based Generation Model 42

Figure 12: Marginal Distribution Based on Copula GAN Generation Model 43

Figure 13: Community Health Activity and Engagement Record system (CHEARS) components 47

Figure 14: λ vs MSE for varying degrees of λ in Lasso modeling 52

Figure 15: LASSO zeroed and non-zeroed coefficients 53

Figure 16: λ vs MSE for varying degrees of λ for Ridge modeling 54

Figure 17: KNN confusion matrix 55

Figure 18: Multinomial logistic regression results. 57

Figure 19: BRFSS data set feature selection 60

Figure 20: SVI dataset features 60
Figure 21: Raw data non-diabetic, pre-diabetic and diabetic 61
Figure 22 Merging Pre-diabetic and Non-Diabetic 62
Figure 23: The premilitary data exploration - An optimization problem in resource allocation... 62
Figure 24: BMI distribution bar chart 63
Figure 25: Age vs Mental Health days 64
Figure 26: Income and Diabetic rate as a proportion 64
Figure 27: Income and Mental health unbalanced data 65
Figure 28:(a) Age and diagnosis of diabetic diagnosis (b) Age and diagnosis of High BP in the Florida BRFSS data set 66
Figure 29: BMI and Physical Health showing increase in diabetes when BMI is high and physical health is poor 67
Figure 30: Correlation plot of observed data (predictors and response variable) 68
Figure 31: Checking for Multicollinearity 69
Figure 32: Pair plot for the numeric features 69
Figure 33: PCA for the numeric data 70
Figure 34: Elbow graph of the PCA for the numeric features 70
Figure 35: PCA Elbow graph representation using all features 71
Figure 36: GMM vs K-Means for categorical and numeric features 72
Figure 37: T-SNE scatter plot 73
Figure 38: K-prototypes after 10 runs with tqdm range (2 and 6) 74
Figure 39: KNN model metrics 75
Figure 40: The KNN performs well with more than 2 clusters 75

Figure 41: Comparison of Supervised Machine Learning Models 77

Figure 42: (a) Full Logistic Model with All Features (b) 95% confidence intervals of the coefficients 77

Figure 43: A step wise feature selection 78

Figure 44: Prediction of diabetes with only BMI 79

Figure 45: Increase in precision, recall and F-1 score of KNN with SMOTE ENN 80

Figure 46: The Community Health Activity and Engagement Record Systems (CHEARS) Architecture 82

Figure 47: The Community Health Activity and Engagement Record Systems (CHEARS) Architecture 83

LIST OF TABLES

Table 1 : Feature Comparison and Cross Walk between BRFSS, SVI and MIMIC-iii Data Sets 44

Table 3: MIMIC-iii data linear regression model analysis.....50

Table 4: Linear regression coefficients and R2.....51

Table 5: Comparison of Regression Modeling performance.....54

Table 6: Coefficients of multinomial logistic regression.....55

Table 7: Bin probabilities for 6 random observations 56

CHAPTER ONE: INTRODUCTION

1.1 Emergence in Healthcare

In common parlance, emergence is a phenomenon of change that results in new patterns. In terms of evolutionary theory it is defined as, "the rise of a new system that cannot be predicted or explained by antecedent conditions" (Britannica, 2017). However, in healthcare causal structures and their associated ability to explain a new system is a prerequisite for informed intervention design and evaluation of performance. In healthcare planning temporal datasets inform the investigators or stakeholders on the future intervention strategy. Studies using Behavioral Risk Factor Surveillance System (BRFSS) data analysis have demonstrated the seasonal nature of health amongst US adults (Jia & Lubetkin, 2009). Therefore, there exists an emergent dataset that is temporally separated from an emergent basis dataset. The empirical relationship between the emergent and its emergence basis is known as emergence (Sartenaer, 2015). There exists several public data sets such as the Social Vulnerability Index (SVI) and the BRFSS that readily provide a basis of evidence for the phenomenon of emergence at the regional, county and census tract level (CDC, 2015).

In the context of health improvement planning emergence can thus be defined as, "an empirical relation wherein the emergent while ontologically determined in part by the emergence basis is qualitatively novel." This definition captures the observation that as health improvement activities along with systemic changes such as infrastructure improvements, improved access to care and environmental changes the causal association between observable predictors and health objective outcomes may change (Zhang et al., 2012). Therefore, the emergent and the emergence basis in some cases have topological nonequivalence because of temporal separation. Further emergence

is typically two-faceted in that it comprises of both synchronic and diachronic aspects while it is not always possible to trace the determinative path from the emergent basis to the emergent. Subsequently, a revised or new causal structure maybe required to sufficiently explain the emergent dataset (Rueger, 2000). While the emergent dataset maybe causally and constitutively determined by its emergence basis it is often not possible to trace the causal or constitutive chain from the basis to the emergent (Mossio, Bich, & Moreno, 2013). There might be several unknown or unobservable factors (constituents) that can mediate or moderate outcomes (Braithwaite, 2018).

The community activity intervention model (CAIM) consists of a set of activities that intend to improve a specific health outcome at the community level (Layde et al., 2012). In most cases this means at the census tract or the county level. This is because population within a census tract typically has similar level of access to care, built environment, schools and other infrastructure resources. The racial, economic, and demographic profiles tend to be similar within a census tract. Furthermore, the SVI measures reported by the CDC are at this level. SVI data has been used in the planning of community health care activities (Flanagan, Hallisey, Adams, & Lavery, 2018). At its core CAIM aim to allocate community partner resources to mitigate serve disparities in target populations thereby improving health outcomes for those individuals and in effect improving the overall health outcomes for the community. The recent COVID-19 pandemic has motivated several evidence-based resource allocation models to mitigate the adverse impacts at the local level. These models utilized the SVI dataset and other publicly available localized data (Arling et al., 2021). Similar models can be applied in the care and resource allocation for chronic diseases. Previous studies have utilized SVI data in heart failure readmission modeling (Regmi et al., 2021).

1.2 Overview of the Problem

There is an abundance of scientific literature on various CAIM models that attempt to explain the causal relationships between observational health predictors and indicators of community heath such as diabetic hospitalization rates (Thomas T. H. Wan, Terry, McKee, & Kattan, 2017). These models are developed based on datasets that are accessible to the investigators. Large studies that are well regarded in the scientific community are usually associated with large and complex datasets. However, it is important to note that many individual clinics especially those that operate in rural areas in United States do not have access to large populations. Thereby, limiting their access to large datasets needed for prediction and decision support. It is important to note that many clinics may serve only 10 to 20 patients a day per physician. Rural care is further complicated by the fact that resources are highly constrained and therefore collaborative care models that involve several partners are essential to serve the needs of low-income residents (Powers et al., 2020). Patients may be diagnosed with different conditions. In rural clinics it is quite possible that for each disease condition a typical clinic may only see a few hundred patients that year (Mehrotra, Wang, Lave, Adams, & McGlynn, 2008). This number may not be sufficient enough to support data driven structure learning methods such as deep learning or score-based and constraint-based learning Bayesian Models (Raghu, Poon, & Benos, 2018). Furthermore, data driven structure learning can result in spurious associations (Scutari, 2017). At this juncture it is important to discus the specific problems targeted in this dissertation.

Problem 1: Health improvement intervention designs at specific clinics and communities need to be empirically validated using localized data sets that might have a small number of observations. This is extremely important given the fact that the local characteristics such as social vulnerability and built environment have impact on health outcomes.

It is also important to note that currently most health care interventions in clinical and hospital settings are based on casual structures as described in the available literature (Wan, 2002). Tacit knowledge or know-how and knowledge gained from such literature that is most often used to design behavioral intervention design. However, these interventions do not consider implicit knowledge available in Electronic Health Record (EHR) systems (Thomas T. H. Wan, 2002), Implicit knowledge generated by the clinic from actual care provided to its patients is not fully utilized in clinical decision making and intervention design. The situation is more adverse in CHIP development. CHIP development is informed by local SMEs who suggest activities to improve outcome based on personal knowledge and bias. In these settings the associated activities are determined without the benefit of theoretically informed causal models and there are no systems to capture data to generate implicit knowledge. There are initiatives to bring academic partners to health coalitions to facilitate clinical services for the community (Wells et al., 2006). There are however, several tool kits readily available to facilitate the community health assessment and subsequent plan generation (Hewitt & Dykstra). Observational Health data collection systems are typically not available at the point of planning to all participants in the planning process. As a result, processes, systems, and interventions are designed based on subject matter expert opinion and not informed by empirically validated theoretical model based on literature. Recent literature recognizes the need to incorporate the use of publicly available local population health data in the community health assessment and planning phase (Stoto, Davis, & Atkins, 2019). The same idea can be extended towards implementation and progress monitoring. Publicly available health outcome summaries from data sources such as Florida Health Charts are used to ascertain targets for improvement. (FDOH, 2021)

Problem 2: Localized data collection system for empirical validation of theoretically suggested causal associations between proposed activities and outcomes have not been implemented to monitor the progress of community level health improvement interventions.

A fundamental assumption during development of interventions in the community health improvement planning process is that causal structure is temporally static. However, it is well understood from the available literature that the causal structure may change to better explain unexplained variance (Zelta) in endogenous/ Outcome variables over time (Hu & Bentler, 1999).

1.1.1 Significance of the Problem

The significance of this work is that it introduces a formal methodology that enables the selection of appropriate quantitative methods that can utilize localized small sample sizes for health outcome prediction based on causal inference while factoring in the phenomenon of emergence. Further, it supports design of theoretically informed data collection system to support the health improvement objectives.

1.3 Research Questions

Based on the aforementioned ideas and facts this dissertation attempts to answer the following research questions from a systems design perspective:

RQ. Design a theoretically informed data collection system for assessing the effectiveness of community health intervention using the EASI™ Methodology?

A. Can we model a health objective such as diabetic hospitalization and generate a theoretically informed Data Collection architecture?

B. Can the data collection architecture be evaluated with Simulated Data Health Condition Specific Model in CHIP and implementation based on Simulated Data?

1.3.1 General Hypothesis

The central hypothesis related to intervention support, causal structure modelling and prediction based on the research questions mentioned that will be tested in this dissertation research are as follows:

H1. Selected features from the BRFSS data set can be utilized to model a data collection system for diabetic hospitalizations.

H2. EASI™ methodology can guide the generation of synthetic data set to validate the data collection system.

1.4 Overview of the Dissertation Research

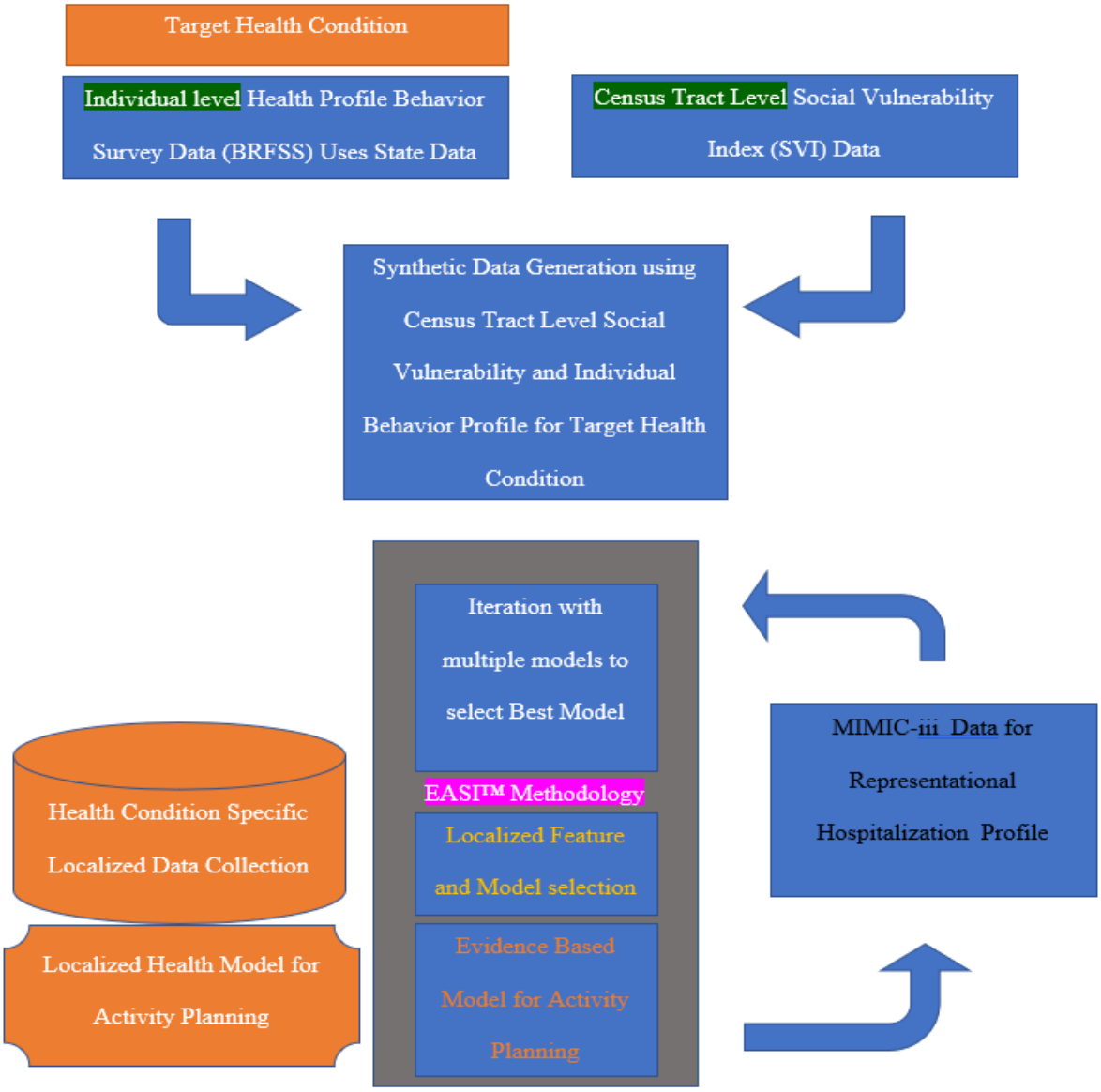


Figure 1: overview of EASI(TM) methodology - data preparation, synthesis, collection and modeling process

1.5 Summary

In this dissertation EASI™ methodology is the core topic and the critical contribution presented to the community of engineers and scientists. The purpose of this project is to develop a localized diabetic prediction model that considers census tract level social vulnerability and behavior data. A fundamental idea that drives this dissertation research is the fact that local social vulnerability

and behavior impact the risk of developing diabetes and length of stay owing to complications and comorbidities associated with diabetes leading to hospital admissions. MIMIC-iii data set is utilized as representational hospitalization data to determine the length of stay. Recently, GANs have been utilized to generate synthetic health data for several applications (Chen, Lu, Chen, Williamson, & Mahmood, 2021). In this research we outline an algorithm to generate synthetic localized population health data sets using GANs.

Typically, social vulnerability and behavioral features are not a part of EHR data collected in the hospital upon admission. However, SVI data is commonly used in community health improvement planning, generating the community health needs assessment and in other programs that require baseline data to inform vulnerability and need-based resource allocation (Flanagan et al., 2018; Lara-Garcia et al., 2020). Here the intention is to create an integrated synthetic dataset for the purposes of community health intervention planning that leverages BRFSS and SVI data sets along with MIMIC-iii data set that is utilized as a representational hospital admission data. There are several statistical approaches in literature that provide guidance on how data from different sources might be combined. Health information system researchers have made the case for data combination from different sources to facilitate open data exchange platforms (Hayashi et al., 2021). The census tract level social vulnerability index data is used as counter to draw samples from the BRFSS dataset. The synthetic data provides representation of the population for a given census tract. This research provides a foundational framework for utilization of evidence-based strategies in community health interventions. It is important to note that diabetic hospitalization is a multifactorial problem; a combination of genetic, comorbidities, demographics, behavioral and environmental factors are involved in the disease progression that ultimately results in hospitalization and readmission. This disease is responsive to behavioral practice changes such as

improvement in diet, exercise, and regular testing. Individual overall perception of health, social perception of health and exercise can influence outcomes at a community / county level.

Therefore, the core contribution of this dissertation is that it introduces a methodology that supports evidence-based decision making and theoretically informed efficient data collection in the emergent health data environments such as in community health improvement planning and implementation. Such work can result in an open yet secure data exchange framework to facilitate community health improvement coalitions in addition to its contribution to science.

CHAPTER TWO: LITERATURE REVIEW

The fundamental contribution of this dissertation is to describe an ensemble methodology known as the Emergent Approach to Systems and Interventions (EASI™). This methodology is applicable to any dataset that shows temporal emergence(Sartenaer, 2015). Observational health data, such as community health data are typically emergent in nature because casual relationships between components of the intervention i.e activities to improve health and outcomes may change due to the interaction between systemic, ecological and infrastructure improvement other random factors that cannot be predicted at the time of CHIP.

It is noteworthy to state that the EASI™ methodology is transdisciplinary in nature. When applied in the context of community health improvement it draws from several established in scientific literature. This chapter will systematically review relevant literature and describe how specific techniques will be applied to develop the EASI™ methodology.

2.1 Community Health Improvement Plan (CHIP) Implementation

Existing observations indicate that patient behavior, behavioral change, self-adherence, knowledge, education, engagement along with attitude, family support, social capital, and motivation have huge impact on daily habits, practices, beliefs and ultimately on clinical outcomes. The BRFSS and SVI datasets contain features that impact diabetic hospitalizations. Similarly, in the community activity intervention model specific localized census tract level SVI data features that impact the health condition (diabetic hospitalization) are identified. A criterion for selection of an SVI feature is that it can be improved by allocating resources. For instance, if the SVI feature education indicates that the target population has low educational attainment then a possible the community activity intervention model (CAIM) will include education on that health condition. Based on the improvement desired in the outcome activities and resource allocation is planned.

There have been similar approaches that use health administration data and machine learning models to predict adverse outcomes. (Ravaut et al., 2021)

The objective of using community activity intervention model (CAIM) is to accomplish the following:

1. Describe the process and a method to develop an intervention that is based on knowledge learned from prior intervention
2. Demonstrate the refinement of the intervention based on prior knowledge.
3. Each iteration of the model should improve prediction practice measures for an individual.
4. Demonstrate that refinements and using parameters identified in the causal model can be used to improve predictions with small data sets.
5. Facilitate high risk groups for poor practice and resultant health outcomes
6. Monitor and confirm sustained practice with minimal number of survey items.
7. Study the impact of interventions on practice

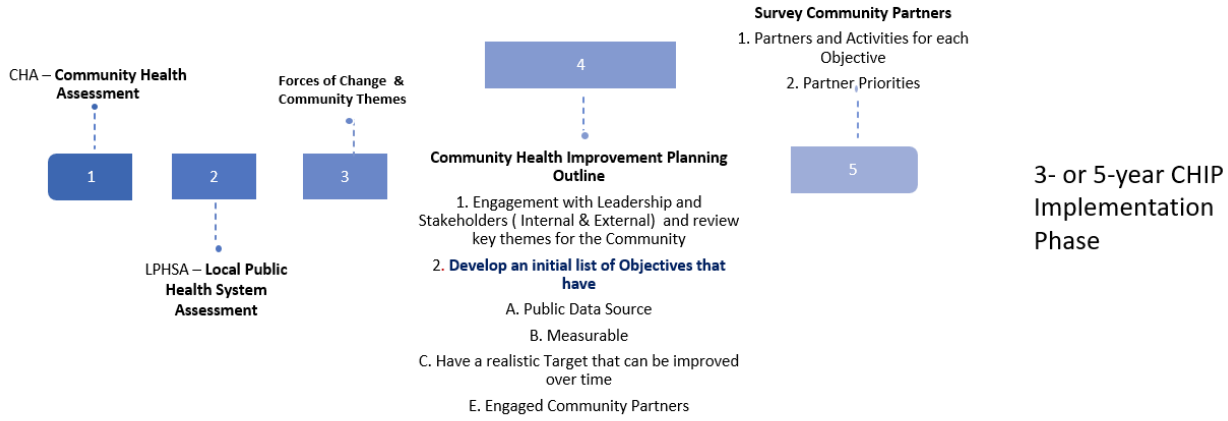


Figure 2: CHIP development process

There is a large body of literature on the design of community-based interventions. The meaning of community-based intervention can differ based on the context. (McLeroy, Norton, Kegler, Burdine, & Sumaya, 2003)

2.1.1 The Reduction of Diabetic Hospitalizations Health Objective

In this section we will use a typical community health improvement planning objective and review literature from research and practice. Hospitals, CHDs in partnership with community organizations can improve the continuum of care, by working as a collaborative. (Philis-Tsimikas & Gallo, 2014) As an example of a target Orange County, Florida Dept of Health plans to reduce the rate of diabetic hospitalization by 5% from a baseline of 3069 in 2019 by 1% a year. During the Community Health Improvement Planning (CHIP) phase community partners discuss possible activities. This discussion is guided by subject matter experts, who are typically representatives of the various partners of the CHIP coalition. Activities to reduce diabetic hospitalizations include **screening** and identification/ screening of susceptible individuals, distribution of food vouchers to selected individuals, provide lifestyle, cooking and nutrition **training,** access to **education** on healthy lifestyle and food choices, biometrics screenings to **evaluate** potential risk of hospitalization. These activities are done by a combination of partners. It is therefore a challenge to develop a uniform set of measures for many of the activities. There are various techniques to collect data both direct and indirect that have been discussed in literature. Evidence-based approaches to intervention planning, that incorporates input from the members of the health coalition, leverages empirical evidence and residents have proven to be effective. (Fernandez, Ruiter, Markham, & Kok, 2019) Further, unlike hospital data which is a single entity that manages its data Community Health Departments (CHDs) work with several partners in the CHIP and must utilize data from diverse sources. (Weinick, Caglia, Friedman, & Flaherty, 2007) Probabilistic

approaches allow the combination of data from diverse sources into a joint probability distribution. Methods of estimating data points when administrative EHR are not available have been described in literature. (Weinick et al., 2007)

The Chronic Care Model (CCM) is a well-accepted framework for managing type 2 diabetes. This model is based on data collected at the community, clinic and at individual patient EHR records. This model is well aligned with the operational modality of CHIP implementation and is promotes efficient use of resources, provides a basis of allocation of new resources, while aligning activities between health teams and patient interactions. Literature suggests that incorporation of the CCM model for the management of diabetic patients results in improvement in health outcomes. (Baptista et al., 2016). Research in the field has identified the following as **important risk factors** for diabetes Research in the field has identified the following as important risk factors for diabetes blood pressure (high), cholesterol (high), smoking, diabetes, obesity, age, sex, race, diet, exercise, alcohol consumption, BMI, Household Income, Marital Status, Sleep, Time since last checkup, Education, Health care coverage and mental health.

The meal restrictions may include education on limitations on the type of foods, food consumption recommendations thought the day as Knowledge and Attitude are known to have impact on adherence and daily practice. As a result models have been proposed to explain the causal association between knowledge, attitude and daily practice (Rahaman, Majdzadeh, Holakouie Naieni, & Raza, 2017) This model has been extended to include motivation as a factor. Inclusion of motivation improves the explanation of the variance encountered in the previous the model. (Thomas T. H. Wan et al., 2017)

A practical implementation of such theory might be to send notifications related adherence to caregivers and the social network of the patient including medical service providers and activity

partners. Network analysis of patients evaluate the impact of the structure of human connections and social capital of a patient. Studies demonstrate that the care provider network has a measurable impact on outcomes for the patient. (Davis, Lim, Taira, & Chen, 2019) Others measures to assess daily motivation maybe collected and covariance determined, motivation for self-care is an important factor that is almost entirely depended on the individual circumstance of the patient. (Shigaki et al., 2010) Therefore, a data collection system based on such theory might be an efficient way to predict adherence in small patient populations. (Vaona et al., 2017) Given the importance of data collection there have been several efforts to optimize data collection methods and models in hospital and community settings.(Holden, McDougald Scott, Hoonakker, Hundt, & Carayon, 2015) As new insights are gained, specific to certain populations, data measures collected are often updated. (O'Connor et al., 2011)

As more data is collected over time additional causal factors may emerge. The model can conceivably be expanded to include co morbidities, outcomes variables can be expanded to include obesity, cardiovascular issues, hospitalizations, and readmission. The EASI™ methodology can allow for such flexible expansion and accommodation of new factors based on health objectives and community goals for improvement.

Mediation and Moderation effects are important phenomenon in physiological studies (Baron & Kenny, 1986). In the context of diabetic patients, the size of the effect positive and negative affect on daily practice is important. Motivation to use the system and attitudes towards patient and self-care may reveal which patients will benefit the most from the use of the system. The factor loadings form the causal model can be used in the development of causal Bayesian prediction algorithms (Gupta & Kim, 2008). The EASI™ methodology can help to answer questions of clinical relevance at an individual level including predicting the possibility and

severity of adverse events. Furthermore, the EASI™ allows the use of interaction effects as an exogenous variable for prediction of important outcomes.

The EASI™ methodology allows community health practitioners to adopt a systems approach to develop predictive models that target specific diabetic related health outcome measures such as hospitalization, length of stay, and mortality rate. This methodology provides an analytic framework to guide community health decision making for optimal outcomes and to optimize resource allocation. Similar approaches towards developing resource allocation optimization criteria can be found in recent literature (T. T. H. Wan et al., 2022). In this paper the concept of G (goal attainment) is expressed as a regression model that considers individual and interaction effects of efficiency and effectiveness along with a constant to optimize local factors that are not evaluated. This work extends the "Health as a System" logic model for diabetes care performance and outcomes. It extends the logic model by leveraging existing public data sets and survey tools. The diabetic logic model incorporates individual, ecological, interaction effects, discrete event models, latent constructs, and growth model's overtime. This results in the Community Health Engagement and Activity Record System (CHEARS™) data architecture that can both implement this model and leverage the optimization algorithm outlined in this work.

2.2 Causal Modeling in Health Care

To improve interventions and predictions in observational health data utilization of CASUAL STRUCTURE in prediction and clinical decision support by means of linking it with causal BN (Gupta & Kim, 2008). The initial causal structures maybe based on previous literature and is used as a basis to inform the emergent base prior to the start of theory-based data collection (Sartenaer, 2015). Here we combine system design with casual structure validation and utilization of the causal structure in prediction (with respect to forward inference) for the purpose of synthesizing data

models and simulation systems based on these data models (applying backward inference)(Blodgett & Anderson, 2000). This approach together with its implementation in the form of a simulation system allows researchers to leverage their validated causal structure models and develop data driven causal reasoning-based decision support systems. The data acquisition modules of the system allow researchers to acquire data that can then be used by both causal structure and Causal Bayesian Modeling. These two techniques can be then linked together to overcome the limitations of the other resulting in a decision support system capable of prediction, diagnosis, and evaluation of interventions (Gupta & Kim, 2008; X.-f. Xu, Sun, Nie, Yuan, & Tao, 2016). Causal Discovery from observational data fall under two methods: constraint-based and score-based. Based on available literature it is common knowledge that hybrid models that combine the two methods can also be used to learn causal structure of BN. This task is commonly known as structure learning. The main objective here is to determine the Directed Acyclic Graphs (DAG) that best captures the conditional independencies present in the data (Scutari, 2017).

Bayesian modelling may be interpreted as causal models if certain conditions are met. These conditions are causal sufficiency, effect variables are only dependent on their direct causes and independent of all other causes (Markov Condition), probabilistic independence due to specified causal structure (Faithfulness condition) and when specified and unspecified causes are independent (Gupta & Kim, 2008).

As an example, CASUAL STRUCTURE has been used to explain the casual associations in the diabetes caregiving process (Holmes et al., 2005) . A priori model is used this model is them refined by dropping the weak associations to develop a parsimonious model. In this experiment the investigators have taken into consideration the following categories of variables: a) cognitive variables, b) psychosocial variables, c) diabetes care variables, and d) disease and demographic

variables. A key feature of this research is the emphasis on parsimonious biopsychosocial models for the purpose of improving caregiving for adults with type 2 diabetes. Developing an efficient and parsimonious model for diabetes caregiving happens to be a key step towards synthesizing effective interventions with a higher ease of implementation (Holmes et al., 2005).

2.3 Inference of Causality and Causal Structure

2.3.1 Structural Equation Modeling and Bayesian Networks

Structural Equation Modelling (SEM) is a method to empirically validate theoretically informed causal structures. SEM models are parsimonious. They attempt to provide the simplest theory to explain the observed data. Since a SEM is developed based on causal reasoning it can be used in both design and evaluation of interventions (Thomas T. H. Wan, 2002). SEM while excellent in empirical validation of theoretically informed causal structure is not as useful in forward (prediction) and backward (diagnostic) inference. To overcome this limitation, it is useful to link the SEM to a Casual Bayesian Network (CN) or develop a Dynamic Bayesian Network representation (Roversi, Tavazzi, Vettoretti, & Camillo, 2021). While an SEM is a parsimonious representation of the observed variables and conceptualized latent variables BN provide predictions that can be described in terms of probabilities and percentages. (Anderson & Vastag, 2004). Bayesian Networks are limited in terms of their ability to explain causality when used independently. The problem of structure learning in BN has been addressed in many ways. Score-based, constraint-based, and hybrid methods have been used to continually learn structures in BN (Beretta, Castelli, Gonçalves, Henriques, & Ramazzotti, 2018). Using a SEM to test casual relationships and then linking an empirically validated SEM to a BN may be of particular interest in observational community health data sets that show temporal emergence. In this chapter we

will discuss these approaches and their limitations in application to clinical decision support or intervention design.

Additionally, emergence as a social phenomenon has been widely researched (Lichtenstein, 2015). The approach delineated in this dissertation is based on the emergence of improved understanding causal associations over time. Here it is important to review relevant research from the area of emergence and its application in our work. To achieve the objectives of this dissertation we synthesize a deidentified population dataset from the SVI data and the BRFSS datasets. We use this synthesized data to model the impact of community health interventions on the targeted health condition outcome. The outcome taken into consideration is hospitalizations due to diabetes.

2.3.2 Structure Learning in Bayesian Networks

It is important to note that BN can use score-based, constraint-based, and hybrid methods that combine the two methods to learn the causal structure from the data (Chan, Wong, Hon, & Choi, 2018). These learning networks have strong mathematical basis and application in decision support systems (Heckerman, Geiger, & Chickering, 1995). However, they require large data sets to learn the causal structure. Additionally, the structure thus derived is not based on causal reasoning which in turn limits its use in intervention design and implementation (Chan et al., 2018).

2.3.3 Advances in Causal Bayesian Modeling

BN was used to infer that blood pressure had limited or no causality to the incidence of type 2 diabetes. In this experimentation the investigators normalized the data, extracted the required features using suitable algorithms and then applied Bayesian analysis to the data thus extracted to analyze the causal connections (Hays, Revicki, & Coyne, 2005). It is to be noted that Bayesian analysis is a wide and complex area of heuristics. DAG an aspect of Bayesian analysis was

successfully to identify the causal relationship between educational level and incidence of diabetes (Sacerdote et al., 2012).

2.4 Synthetic Data Generation

Existing literature indicate that access to data sets is critical to the development and testing of novel machine learning algorithms. (Tucker, Wang, Rotalinti, & Myles, 2020) Synthetic data sets can be helpful in generation of machine learning models for sparsely available real world data sets. Furthermore, synthetic health data sets can help to validate data collection system design and architecture in a cost-effective manner. Such data sets can facilitate the design of novel data collection systems such as the CHEARS™ architecture presented in this dissertation (Pollack, Simon, Snyder, & Pratt, 2019).

2.4.1 Generating Synthetic Datasets for Healthcare Modeling using GANs

GANs were conceptualized by Goodfellow et al., in 2014 for image generation (Goodfellow et al., 2014). There have been several efforts since to use GANs to generate observational health data and EHR data (Choi et al., 2017). This work resulted in the medGAN algorithm to generate high fidelity synthetic patient records. The team demonstrated results of data distributions and predictive modeling were similar in both synthetic and real datasets. Further advances in GANs have resulted in methodologies to generate multi-model synthetic medical time series data. (Esteban, Hyland, & Rätsch, 2017).

Copula GAN and CTGANS are helpful in generating synthetic tabular data. We have utilized both GANs to generate the synthetic data set that combines BRFSS and SVI data using a few selected common features. Similar applications of GANs in combing datasets have been reported in literature (Yoon, Jordon, & Schaar, 2018). Such a data set can help to leverage freely available localized datasets to validate health condition specific data collection and measurement models.

These models can help to transition community health practice from a subjective art based on local subject matter expert intuition towards a more evidence-based planning, implementation and measurement cycle with data collection that is informed by theory.

2.5 Public Datasets in Community Health Planning

There are several data sets that are freely available which might find use in community health planning at a local level. Data driven, machine learning approaches have shown utility in the prediction of diabetic hospitalizations and outcomes for target populations (Dinh, Miertschin, Young, & Mohanty, 2019). In this dissertation the MIMIC-iii, BRFSS and SVI datasets have been selected intentionally. The SVI dataset is helpful in mapping population disease vulnerability (Linder, Marko, Tian, & Wisniewski, 2018). It is well established in literature that social determinants of health have a measurable impact on disease prognosis, specifically diabetic outcomes such has hospitalization and mortality (Hill-Briggs et al., 2020). The BRFSS and the SVI dataset are CDC conducted survey and standardized across the nation. Therefore, any methodology built using the features captured in these two data sets has national significance. Further the MIMIC-iii data set is selected as a representational hospitalization data set. It essentially mimics data collected in a New York hospital that has been utilized by researchers to make meaningful diabetic predictive models (Anand et al., 2018). The content of this data is based on real data in a hospital EHR system. Therefore, similar techniques can be applied to any hospital data system. By selecting these three data sets we maximize the applicability of the methodology described in this dissertation.

2.5.1 BRFSS Data Set for Community Health

The CDC conducts an annual survey in all 50 states known as the BRFSS (https://www.cdc.gov/brfss/index.html) the state of Florida is a participant in this

program. The data set was extracted an prepared for analysis. A total of 279 factors are present in the data set. There are several supplemental questions on prediabetes and diabetes. The survey records behavioral and demographic Reponses. The publicly available diabetes health indicators dataset (https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset) is used in this project. This contains 253,680 survey responses from cleaned BRFSS 2015 - binary classification**.**

The BRFSS data set will be used to develop probability of hospitalization and readmission for various groups such as people over 65, obese and people in other high-risk groups as listed in the Floridahealthcharts (https://www.flhealthcharts.gov/ChartsReports/rdPage.aspx?rdReport=NonVitalIndNoGrp.Dataviewer&cid=8574). Classification models to identify high risk pools for diabetic hospitalization and readmission will be developed. Using the BRFSS data diabetic hospitalization rate for various groups and counties will be predicted and compared with the true values in flhealthcharts. The model will be revised to minimize loss using outcome data from 40 counties and the remaining counties will be used as test outcomes to assess the performance of the model. The ability to predict hospitalization rates based on behavioral data will help in the planning and implementation on health care interventions to reduce the hospitalization and readmission rates.

2.5.2 SVI data set for community heath

Health disparities are a result of several historical, geographical, demographic, community and individual health profiles. Restoring health equity as a community heath objective has been a focus for many decades. Elimination of disparities is a multidimensional problem further complicated by several spatial and temporal factors that are specific to health conditions. For instance, the temporal characteristics of diabetic progression is different from covid-19, however similar social determinants of health features might be shared. A simulation study shows that mortality from

COVID-19 shares similar social determinants with diabetic progression. (Seligman, Ferranna, & Bloom, 2021) This indicates health disparities result in skewed outcomes for vulnerable population in both cases. However, the profile of vulnerable population might differ in other aspects and the population might respond better to community level health interventions if specific groups can be identified and targeted for specific vulnerabilities. Therefore, disease specific, spatial -temporal, identification of vulnerability may support implementation of evidence- based preventive interventions by community health departments. (Neelon, Mutiso, Mueller, Pearce, & Benjamin-Neelon, 2021) The CDC conducts census tract level surveys of social vulnerability https://www.atsdr.cdc.gov/placeandhealth/svi/index.html and makes this data available to the public.

2.5.3 Mimic iii Hospitalization Data

The MIMIC-III dataset in our analysis. This publicly available dataset includes information on over 40,000 patients who attended the Beth Israel Deaconess Medical Center's ICU and Hospital from between the years 2001 and 2012. The information dataset was created with data from what was described as routine hospital care (i.e., patient demographics, blood tests, urine tests, etc.). The data was sourced from critical care information systems, electronic health records, and death records. The dataset itself is available in a PostgreSQL database format and was sourced by following the instructions at https://physionet.org/content/mimiciii/1.4/.

2.6 Modeling Approaches Applicable to Diabetic Hospitalization

In this section we discuss the regression and classification models that we utilized in our analysis of the mimic data. We discuss how each statistical model works in the general and introduce how we applied the statistical techniques to our specific instance of the data. There is

extensive literature on the appropriate use of modeling techniques and consideration while developing predictive models for specific health outcomes. (Grant, Collins, & Nashef, 2018)

2.6.1 Linear Regression

Linear Regression models have been extensively used in health care costs and utilization predictions and application of this technique is well established in both research literature and academic texts. (Gregori et al., 2011) One of the challenges for linear regression is developing a model that is not overfitted. High model accuracy using training data (low training MSE), but low accuracy while using test data (high test MSE) is a key sign that a model may be overfitted. There is extensive literature on proper application of linear regression models and derivative models such as penalized regression models in prediction of costs in specific settings. (Thongpeth, Lim, Wongpairin, Thongpeth, & Chaimontree, 2021) Overfitted models are not generalized enough to account for unseen data and therefore perform poorly during the testing phase. Using a very large number of predictors may lead to overfitting because it makes the model very flexible to the training data. To prevent model overfitting, a subset of the predictors must be selected, especially the predictors that are best related to the response. Best subset selection, forward stepwise selection, and backward stepwise selection are three subset selection methods that can be used to determine the best combination of predictors for a linear regression model. Forward and backward stepwise selection are typically used for models with a very large number of predictors for computational reasons. These selection methods work well but may not select the optimal subset due to their selection algorithm. The best subset selection method is computationally expensive but guarantees the optimal subset of predictors because the algorithm iterates over every possible combination of predictors and creates models that minimize training RSS. Since minimizing RSS leads to model overfitting, the final step in best subset selection is to estimate test MSE for each

of the models created in the best subset algorithm by using several statistics such as Mallow's Cp, Bayesian information criterion (BIC), and adjusted R^2. Mallow's Cp is an unbiased estimate of test MSE and will take on a small value for models with low test MSE, therefore, the model with the lowest Cp should be selected. BIC is similar to Cp in which you one should select a model with the lowest BIC value because it corresponds to the lowest test MSE estimate. However, unlike Cp, BIC places a heavy penalty on models with too many variables therefore it tends to prefer smaller models. Unlike Cp and BIC, one should select the model that maximizes adjusted R^2 because it indicates it will have the lowest test MSE. Any of these metrics can be selected to determine the final predictors of the model. The optimal predictors are then fitted into a linear regression model along with the response to estimate the beta coefficients that minimize RSS using a least squares approach. The final model is finally used to predict the response of unseen test data.

2.6.2 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

The LASSO regression performs feature selection by decreasing the coefficient values of the models to zero. In order to allow the algorithm to determine which features were best, we executed the LASSO regression using our full feature set. We used the *glmnet* library in R to execute the LASSO regression by setting *alpha=1* in the *glmnet* function call[1]. The LASSO regression utilizes a lambda (λ) and ℓ_1 normalization penalty term to force some of the coefficients to have a value of zero thus performing feature selection. As different values of λ generate different regression models, we verified the selection of the best λ by using 10-fold cross validation and selecting the λ with the minimum Mean Squared Error (MSE). LASSO regression is helpful in

[1] https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

development of accurate risk models in small populations typically encountered in rural settings. (Pavlou et al., 2015)

2.6.3 Ridge Regression

Unlike the LASSO regression, the Ridge regression does not perform feature selection. Instead, it decreases the coefficient estimates to have values close to, but never equal to zero. This means that all features that are used to train the model remain in the resulting regression model. We used ridge regression on the full feature set. The *glmnet* library in R was also used to execute the ridge regression by setting *alpha=0* in the *glmnet* function call. The ridge regression utilizes a λ and ℓ_2 normalization penalty term to force some of the coefficients to have smaller values as λ increases. As with LASSO regression different values of λ also generate different regression models, so also we verified the selection of the best λ by using 10-fold cross validation and selecting the λ with the minimum Mean Squared Error (MSE).

2.6.4 K-Nearest Neighbors (KNN)

KNN makes a prediction for an observation by identifying the k nearest neighbors. The observation is then classified into the class that holds the majority within the k nearest neighbors. KNN is accurate when the number of observations is much greater than the number of predictors. KNN is non-parametric and therefore reduces bias while increasing variance. KNN does not highlight important predictors because it does not output coefficients, however, a prediction accuracy for classification can be obtained from a confusion matrix. The number of nearest neighbors, k, is a tuning parameter for the KNN model. A small k creates a flexible model that can lead to overfitting while a large k makes the model less flexible. Both very high and very low k values can lead to high classification errors therefore k must be carefully selected.

2.6.5 Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is extensively used in healthcare related models where in several predictors are binary, categorical, or numerical. The outcome or the dependent variable is typically a binary or a categorical variable. For example, outcome variable could have 0 representing not admitted and 1 representing admission. An example of a categorial outcome could be mild, moderate, and severe for a disease condition at a particular point in time. MLR predicts the probability of the outcome variable to belong to a particular category or class based on a set of multiple independent predictor variables which can be categorical, binary or continuous. Multinomial logistic regression is a derivation of the binary logistic regression that permits the classification of more than one level in the response or outcome variable. As in the binary logistic regression, the multinomial logistic regression uses the maximum likelihood function to estimate the class that the response probability will fit in.

As in all data analysis processes careful consideration should be taken when implementing the multinomial logistic regression. Prior to performing logistic regression, exploratory analysis of the data should be conducted to justify the use of multinomial logistic regression. It is also crucially important to analyze your independent variables for collinearity as to not undermine the statistical significance of your predictor variables. In addition, univariate analysis should be performed to assess for outliers, high leverage points, and skewed data.

Multinomial logistic regression is a particularly useful tool because it performs on few assumptions. It does not assume linearity, normality, or homoscedasticity. Yet we must be aware of the few assumptions that it does presume. In respect to the independent variables, multinomial logistic regression assumes the independence of response variable classes. In other words, the

choice class of a response variable does not depend on the class on another response variable class. Furthermore, it is imperative to understand that the response variable classes should not be equally separated. This causes issues with the estimation of the coefficients of the model.

CHAPTER THREE: METHODOLOGY

This dissertation terms a methodology to work with temporal emergence in the context of observational health data. This chapter will provide an overview of the problem addressed, outline three illustrative examples that will be used throughout this work to demonstrate application and list the research questions. Additionally, the significance of the EASI™ methodology in the context of observational health data is discussed.

The EASI™ methodology has been developed to work with emergent data; a phenomenon that result in unpredictable change to the causal associations and components of the causal association. By emergent we mean the data shape, components, and causal associations between components of data varies overtime from a given emergence basis. This is typical in traditional healthcare community health settings. Community health improvement planning focused on improving certain health objectives is generally driven by local subject matter experts. Subject matter experts draw upon their experience and suggest activities that can likely perceive the outcomes. In this process the notion of causality is based on expert understanding and awareness of the local community health environment. However, as part of the implementation of the improvement plan the divisions are often made upon review of performance versus expected performance. Such revisions usually entail the collection of additional data the performance of additional activities change in the number of activities performed based on expert opinion on the impact of these activities have on the outcome variable for instance in this thesis we will discuss the reduction in the rate of diabetic hospitalizations as an illustrative example of the EASI™ methodology.

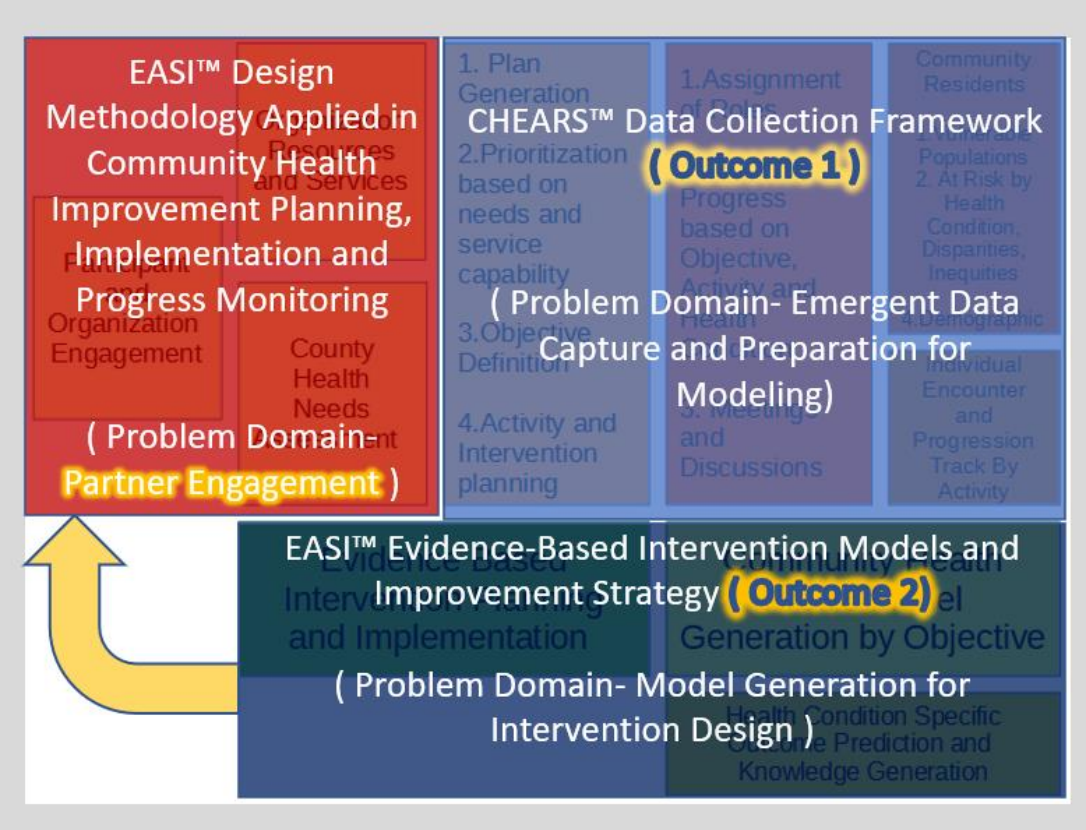


Figure 3: EASI(Tm) Methodology Key Outcomes

The Community Health Improvement Plan (CHIP) has an initial set of activities that can be broadly grouped as: a) training, b) outreach and identification of possible beneficiaries, c) education, and d) services that might include the provision of food and finally evaluation and testing type activities to see how a patient performs through a period of time. It is conceivable that additional activities might be performed by the community partners to achieve the desired improvement based on review of improvements resulting from the initial set of activities. The EASI™ methodology has two major outcomes as shown in figure 3, the first outcome is a theory-based data collection system that informs evidence-based intervention models for targeted health outcomes.

A review of literature and SME experience indicates that comorbidities may play a significant role in outcomes for any given patient. This necessitates retrieval of EHR type data about the population that is being served via the community health improvement plan implementation. Data collection is expensive and imposes an additional burden. To enable adoption the collection system needs to be data efficient and only pull a subset of the EHR data to make meaningful decisions to plan activity that help to achieve community health objectives. This means additional background data about the patient is needed to perceive a concrete pattern. Consequently, the data collection platform must evolve overtime to gather new information and accommodate emergent data shapes.

The EASI™ methodology estimates outcomes (predicted by BN), based on synthetic datasets these estimates are then compared with actual outcomes in existing data. As mentioned before, we will be using simulated data associated with diabetes care the purpose of validation and developing a proof of concept. This choice has been made based on the critical need for developing decision support systems for diabetic care. Diabetic populations are not uniform across all clinics. Diabetic care clinics, specifically rural care clinics, serve small populations with distinct characteristics. Factors such as income, education, ethnicity, social cohesion, and access to care can vary greatly depending on the clinic location. The approach described in this proposal leverages implicit knowledge in the form of EHR records specific to the population served by the clinic. The knowledge acquisition component of the system is allowing the creation survey items based on theory developed using tacit knowledge. Consequently, implicit knowledge generated is fully utilized in behavioral intervention design and clinical outcome prediction. In evidence-based decision-making organized incorporation of implicit, tacit and explicit knowledge can improve performance and outcome (Thomas T. H. Wan, 2002).

EASI™ outlines a novel system design strategy comprising of a method of system design to facilitate data acquisition for both causal model structure specification, validation, and utilization of the causal structure for prediction and diagnosis. The proposed methodology describes an approach, defines a process, and provides mathematical boundaries for its application. A critical step in any modelling process happens to be data acquisition and initial analysis. This is many times accomplished by using a software system. Based on the objectives it is important to note that EASI™ will also be critical in data acquisition that facilitates the required causal modelling of acquired data. In other words, this methodology and its associated software systems will help structure the data acquisition process that will facilitate causal modelling; thereby, obviating the need for reformatting or restructuring the acquired data for the purpose of causal modelling.

3.1 Emergent Approach to Systems and Interventions (EASI™)

Figure 1 below describes the EASI™ that will be validated using a proof-of-concept study on diabetic care. The delineated approach combines elements of data science specifically, causality and prediction with system design for data acquisition and information architecture that supports the generation, modelling, and application of new knowledge. As mentioned before, we will use the KMAP-O model as an example to illustrate this method. The resulting data networks will then be validated by its ability to simulate and predict Outcome and Practice measures. The dataset used for validation of this approach and method will be used from previous literature (Thomas T. H. Wan, 2002).

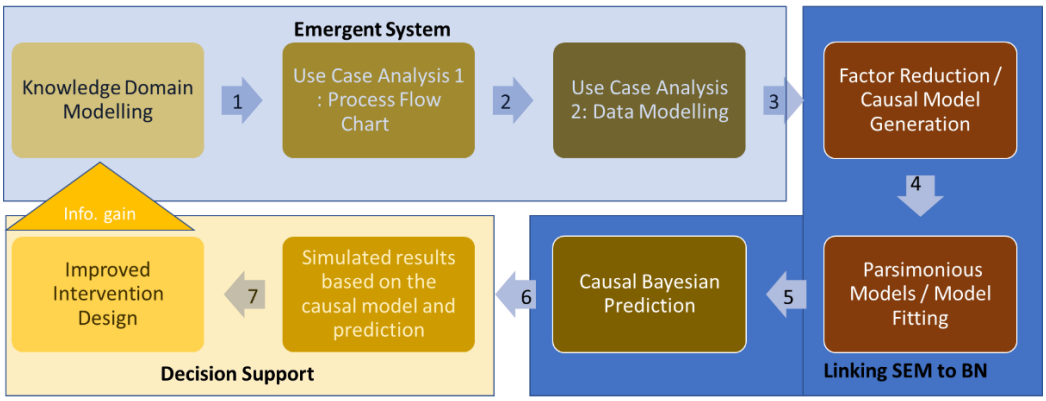


Figure 4: Overview of EASI(TM) Methodology

There are several aspects to the EASI™ methodology and system as represented in Figure 4.

3.1.1 EASI™ Methodology Modules

The system has several modules:

1. Knowledge Domain Representation
2. Modelling and Specification
3. Synthetic Data Generation – Data Model Validation ->Data Acquisition/Collection Modeling
4. Model Fitting and Accuracy
5. Determination of clinical relevance and revised Knowledge Domain Representation.
6. Model Revision upon New Domain Insights

These modules draw data from the underlying data collection system.

3.1.2 EASI™ Methodology Algorithm Explained

The EASI™ methodology allows the user to iteratively specify and acquire data for improved casual models. It then defines a set of mathematical condition under which the factor loadings and latent variables can be used in a causal BN. Improvement has two parameters simplified model

speciation and better fit for the data set. The authors introduce a concept of data efficiency that is supported by this methodology. Each iteration of the model is optimized for improved fit while requiring fewer survey items and constructs.

It is important to mention that ontologies are used to classify educational content, knowledge, practice and outcomes based on disease condition, patient profile, recovery plan, care plan, clinical quality measures and other Domain specific ontologies (Hassanzadeh et al., 2019). Here it is important to list some of the general use of ontologies:

1. Domain specific phenomenon is represented as a mapping of related concepts, concept map, in the system. Various conceptual models need to be evaluated for best fit, predictive utility and efficiency of the information system. The objective is to develop a parsimonious causal model that has the maximum predictive power. (Spirtes, Glymour, & Scheines, 2020)
2. The system facilitates the generation of semantic knowledge graphs for each of the Observed variables and latent factors. The semantic knowledge graphs include, meta data, discussions, responses to survey items, related variables, placement of a given observation in the causal pathway.
3. Domain specific ontology facilitate the identification and isolation of patterns in knowledge graphs.
4. Network topology diagram views of patterns of user inputs in well-structured in semantic knowledge models. These visualization tools help to inform the researchers selection of causal models based on theory or develop new theorical models that require empirical validation via a causal study.

5. Theory based model selection. In this step the researcher applies both inductive and deductive reasoning, reviews literature to identify closely related theoretical models. These models can be path models if observed variables are used directly and the assumptions of regression can be made or structural equation models if the researchers develop conceptual constructs that can be measured indirectly.
6. The structured semantic system can acquire data as per the specifications of the theorized mode. The specification of the theorized model can be described in the system. There are various criteria for model specification that a practitioner of structural equation model can define in the system.
7. The data acquired is then used to compute path coefficients and model fit. The model is simplified to minimize data acquisition burden while ensuring that the model is over-identified, by setting constraints on the parameters estimated.

Algorithm 1: Model Development and Validation

Result: A model that fits the acquired data

1. Define structure of the data model;
   1.1 Define structural components;
   1.2 Define structural hierarchy;
   1.3 Define structural grouping in terms of CAIM;
2. Specify Structural Equation Model in knowledge acquisition;
   2.1 Identify constructs to be validated;
   2.2 List all observed variables;
   2.3 Specify model;
   2.4 Execute the structural equation model from the existing data set from literature;
   2.5 Analyze observed variables;
   2.6 Identify the time series data and construct validation;
3. Specify time series and one time survey items in the knowledge acquisition system;

while *the model does not fit the acquired data* **do**

   5. Separate data acquired into training and test data;
   6. Use newly acquired data and perform statistical analysis with the new data set;
   7. Compare predicted outcome with actual data observed in the test data;
   8. Select appropriate model for predicting patient outcomes until a model that fits the acquired data is obtained ;

end

Figure 5: Algorithm EASI Model Development and Validation procedure

Figure 5 provides the general algorithm for model development and validation using EASI™ .

3.2 EASI™ Process Applied in Community Health Improvement Implementation

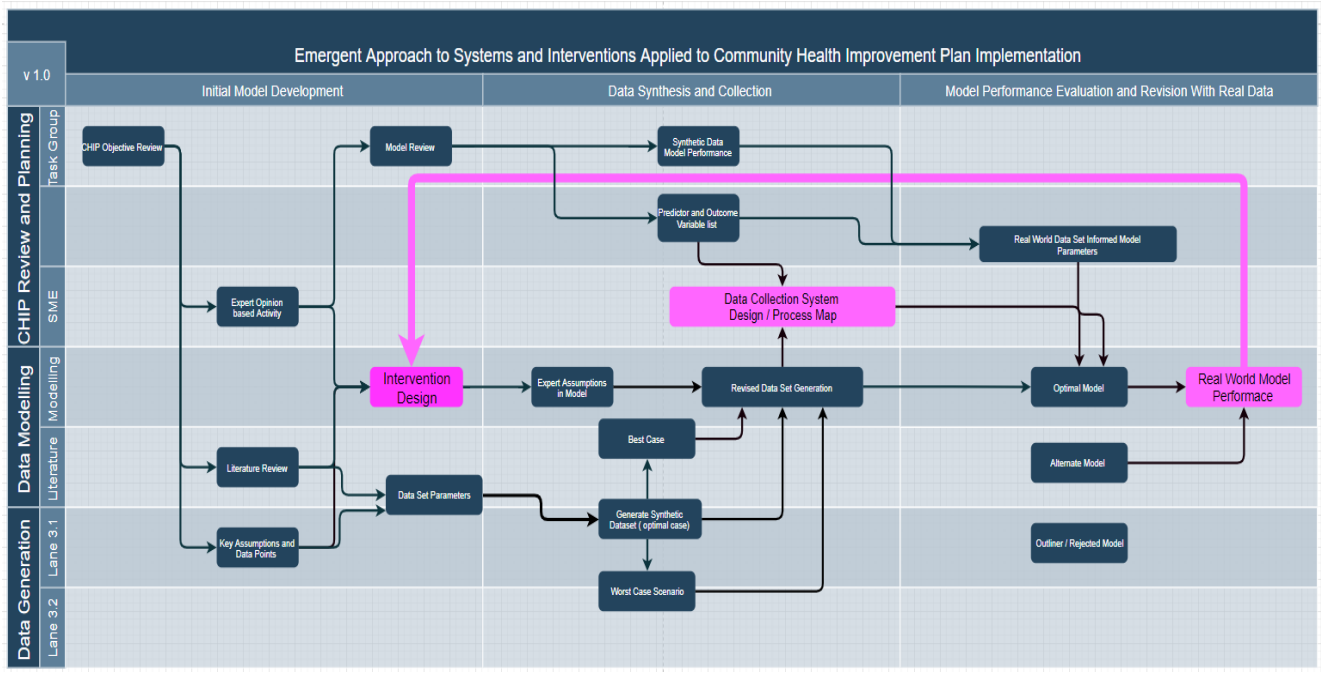


Figure 6: EASI(TM) Methodology Application in Community Health Improvement

The EASI™ methodology can be effectively applied in community health settings as described in figure 6. Community health improvement plans typically have a set of activities that collectively constitute a Community Activity Intervention Model (CAIM). The objective of the intervention model is to improve a target health condition. These health objectives are expected to be Specific, Measurable, Attainable, Relevant and Time-based (SMART). Application of the EASI™ methodology ensures that the health coalition adopts evidence-based process and supporting systems to measure progress towards the SMART health objectives. The first step is to represent the activities for improvement suggested by local subject matter experts as an intervention model. This intervention model should be supported by theory based on previous literature. Careful consideration of the data points (variables) that inform the proposed intervention model is necessary to distinguish between predictors, intermediate outcomes, and final outcomes. Review in literature may also reveal potential latent factors or constructs that are to be considered to

generate an interpretable casual model based on which resources can be allocated and performance measures. A further consideration is the ease and expense of measurement, recording and storage of the data point. For instance, while it is easy to define a meal based on a target calorific value it may prove challenging for a citizen user to calculate calories in home cooked meals or BMI and Blood pressure might be inexpensive and easy to measure when compared to triglyceride levels when assessing risk for diabetes. The EASI™ methodology allows for modeling of the intervention based on simulated data points from public or private data sets. This supports the design of the data collection process and system that can be implemented to capture real world data. In doing so the EASI™ methodology facilitates the adoption of evidence-based practices while ensuring efficient data capture where in data measured has a high degree of predictive utility in optimization of resource allocation.

3.3 EASI – Diabetic Hospitalization Health Objective Improvement model

Holmes et al., (2006) have approached the idea of developing decision support systems for helping the diabetes caregiving process by using Structural Equation Models. In this experiment the investigators have taken into consideration the following categories of variables: a) cognitive variables, b) psychosocial variables, c) diabetes care variables, and d) disease and demographic variables. A key feature of this research is the emphasis on parsimonious biopsychosocial models for the purpose of improving caregiving for adults with type 2 diabetes. Developing an efficient and parsimonious model for diabetes caregiving happens to be a key step towards synthesizing effective interventions with a higher ease of implementation. Ideally interventions should not rely on research staff instead must facilitate a smooth capture of variables associated with healthcare processes. In other words, a caregiving decision support system integrated with Electronic Health Records can be achieved through the development of parsimonious models. This can also be a

critical aspect for dealing with diseases and disorders, or other physical problems related to diabetes. Obesity, blurred vision, and kidney failures could be a few examples. The figure below represents the hierarchy of data acquired in the care planning of chronic diabetics who require Knowledge and Practice to improve outcomes. Wan 2002, demonstrated that the path from knowledge to practice is moderated by motivation and mediated attitude. The figure 7 represents data artifacts in the care planning for diabetic patients.

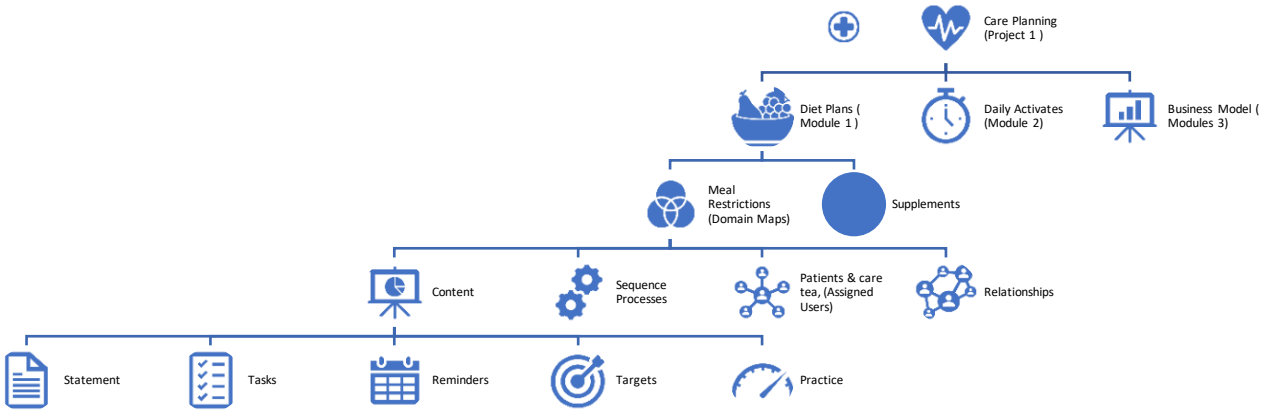


Figure 7: Data hierarchy representation in Care Planning for chronic diabetics.

Figure 7 shows the various data types that are part of a typical diet plan for diabetic patients. The diet plan may incorporate meals as well as periodic supplementations. The complexity of the data collection process can be appreciated by a cursory review of figure 4. Effective management of Chronic disease condition outside of clinical settings requires significant collaboration and active participation of the individual and the care provider network. This presents a multifactorial problem in which local factors and individual circumstances have significant impact on the outcome, in this case the probability of diabetic outcome. This makes the design and implementation of a data collection system challenging. This may offer some explanation as to

why we have an abundance of EMR and EHR systems that track clinical data but fewer community intervention level health information systems. To add further complexity to the challenge of outpatient chronic disease data collection; as can be seen in figure 4, each of the data points that can impact the health outcome might be generated when an individual visits a different provider. Health Data Privacy laws severely restrict the extent and way this data can be shared. The system architecture for an effective community level activity-intervention process to implemented must enable secure sharing of meaningful data parameters to facilitate evidence-based model generation that incorporates data from a multitude of sources while ensuring data privacy is maintained.

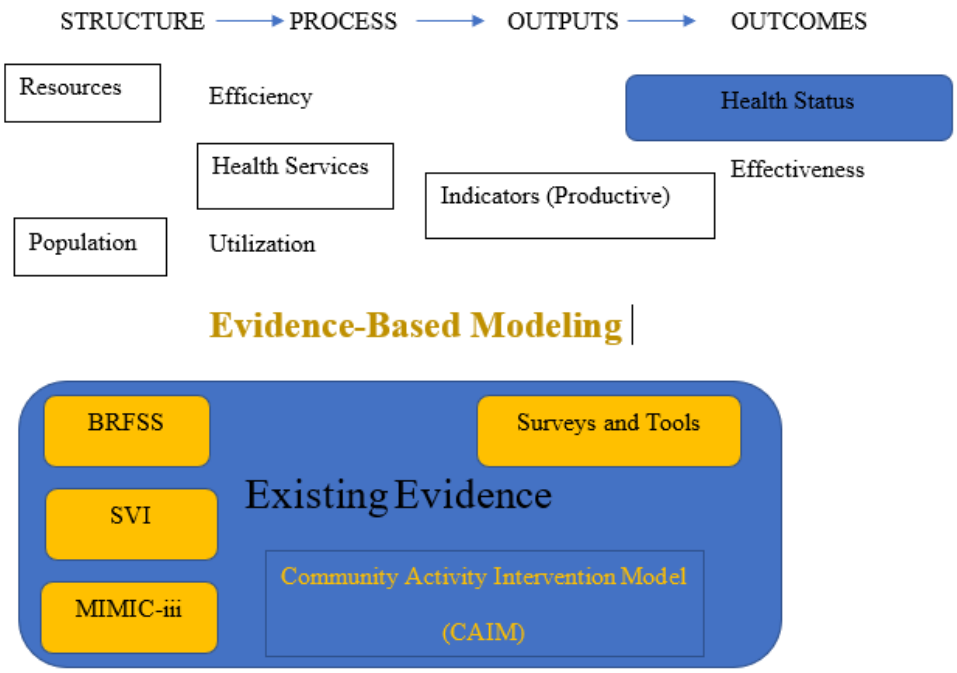


Figure 8:Figure 8: EASI(TM) Evidence Incorporation Model – Methodology

The EASI™ methodology extends the Health as a System Causal Logic Model that has been described in literature (T. T. H. Wan et al., 2022). The figure 8 is an adaptation of the logic model with the extension to the model in yellow.

At the outset of the planning phase, it is often not possible to anticipate all of the significant factors that might emerge over time. This is because at the community health level there can be several localized factors that can gain prominence at various points in time. Static data collection systems fail to pick up and measure these changes. The CDC typically has surveillance systems in place to monitor sudden dramatic changes in the health of a community; however, these systems are primarily intended for epidemiology and infectious disease outbreaks.

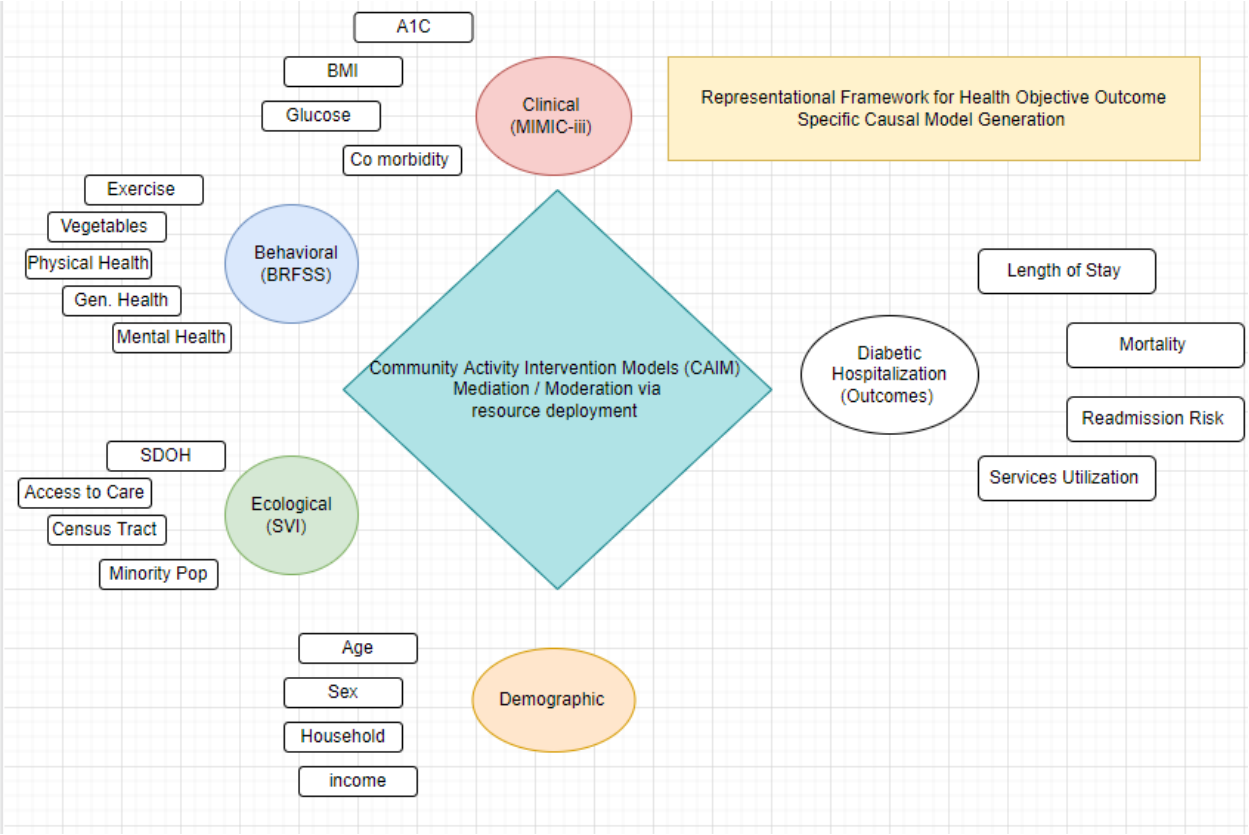


Figure 9: Representational Framework for Health Objective Specific Casual Model Generation using EASI(TM) Methodology

Ever prevalent chronic disease in a community might change significantly over time due to more subtle factors such as demographic shift such as aging residents, income drop or net immigration into a community that over time changes the model that best explains prevalence and progression

of a specific health condition. Figure 9 is a representational framework for intervention specific causal model generation for Community Activity Intervention Models (CAIM). It illustrates the data point (variable) and construct selection for specific health outcomes of interest based on the EASI™ methodology described in section 3.2 figure 6.

3.5 EASI Synthetic Data Generation For Diabetic Hospitalization Simulation

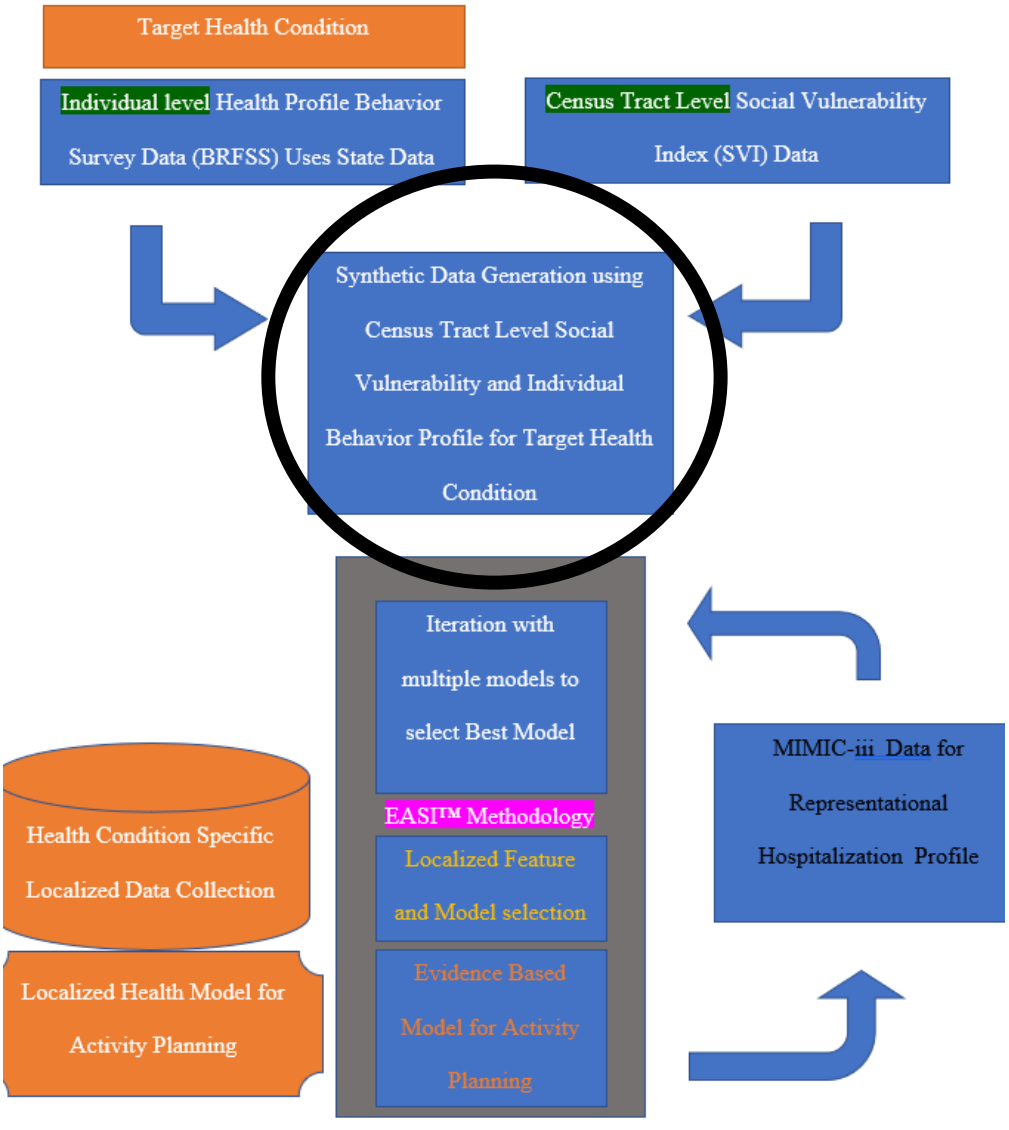


Figure 10: Synthetic Data Generation with BRFSS and SVI data for selected census tracts

Figure 10 illustrates the data envelopment analysis (DEA) for a target health objective of interest based on the EASI™ methodology described I section 3.2 figure 6. In this illustration we see the selection of data points from public data sets that can then be used to simulate outcomes based on predictive utility of key controllable variables. The figure describes the utilization of simulated data to inform the community activity intervention model and analysis of this model performance to review resource allocation. It also helps to define requirements for a real world data collection system and process that can implemented by leveraging data in existing EHR systems and collected using current process, thus facilitating the adoption of evidence based practices by community health coalitions.

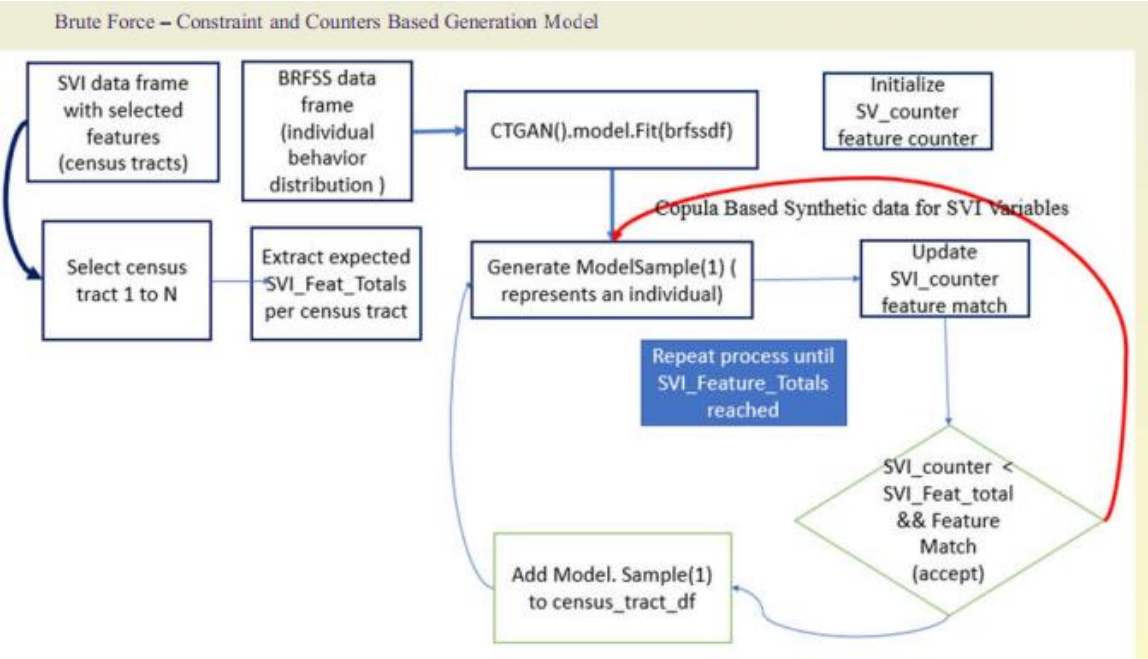


Figure 11 Brute Force - Constraint and Counters Based Generation Model

The algorithm for synthetic data generation represented in figure 11 illustrates the use of SVI data to manipulate the marginal distribution of the BRFSS data set to reflect marginal distributions of the selected variables in the local population.

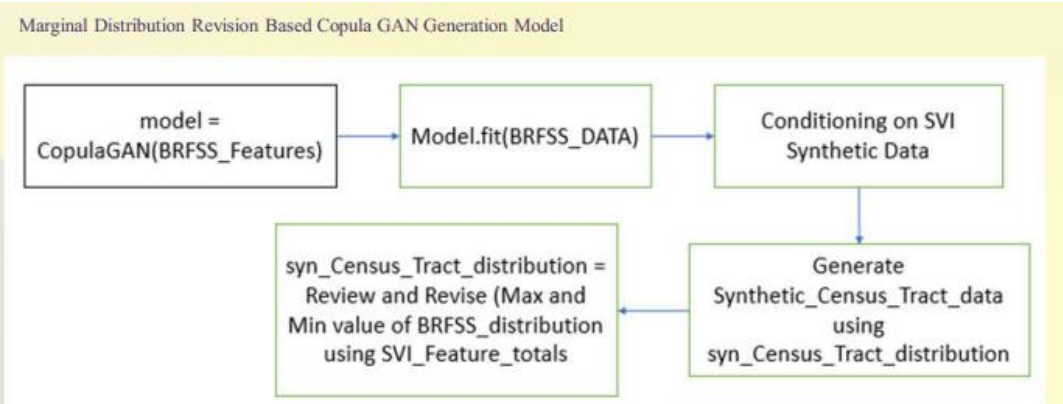


Figure 12: Marginal Distribution Based on Copula GAN Generation Model

Figure 12, describes a generation of marginal data distribution that can be used in the simulation of census tract data set while utilizing the copula from the BRFSS dataset. The BRFSS dataset is conditioned on the marginal distribution taken from the SVI data set. The local population data simulated in this process captures variations at the census tract level while retain overall distribution information in the BRFSS data set.

Table 1 : Feature Comparison and Cross Walk between BRFSS, SVI and MIMIC-iii Data Sets

BRFSS (public) Condition - specific regional behavior data **How must community partner activity induce change in individual behavior or clinical measure to reduce individual probability of hosp.**	**SVI (public) localized -census tract vulnerability data for resource allocation needs** **Census tracts within county that need to be targeted by the community partners for specific activity based social vulnerability**	**MIMIC/ EMR (participant / activity interaction data) patient data** **What is the clinical outcome or adverse event outcome I want to change and by how much**
blood pressure (high)	FID	Clinical Markers
cholesterol (high)	AFFGEOID	
smoking	TRACTCE	Status
diabetes	ST	Sex
obesity	STATE	Insurance
age	ST_ABBR	Marital Status
sex	STCNTY	Age
race	COUNTY	Location of Admission
diet	FIPS	
exercise	FID	
alcohol consumption	E_TOTPOP	
BMI	M_TOTPOP	
Household Income	E_UNEMP	
Marital Status	M_UNEMP	
Sleep	E_AGE65	
Time since last checkup	M_AGE65	
Education	E_AGE17	
Health care coverage	M_AGE17	
Mental Health	E_DISA**BL**	

Table 1 illustrates the presence of common data elements or themes across different public data sets that can be used to combine the simulated data set into one. The combined simulated data set is analyzed for its ability to inform the CAIM and is utilized as a basis to design and implement a

real world data collection system and process for a specific target health outcome as described in section 3.2 figure 6.

3.6 EASI™ Data Collection Model for Diabetic Hospitalization (Outcome 1)

The data collection model based on EASI™ methodology captures the key features based on predictive models developed using public and localized synthetic data sets as a baseline. This ensures only features that contribute to decision making and resource allocation are captured in the system.

3.6.1 Dataset Synthesis for selected Community Health Improvement Plans

The above table indicates common data themes in the BRFSS, SVI and MIMIC data sets. These common themes together with ICD-10 diagnosis and ICD-10 procedure codes form a basis to generate models that cut across clinical data collected during hospitalization that impact outcomes such as length of stay and mortality with community health features such as age, marital status, insurance, and census tract of residence.

3.6.2 System Design Criteria

In the design of EASI™ platform based on the EASI™ methodology availability of appropriate datasets is a significant issue. There is a significant challenge in the design, development, and testing of novel heath care systems regarding the utilization of existing health care datasets. There is the obvious privacy concern additionally, the data sets that are publicly available may not have all the data attributes to fully test the proposed system capabilities. The shape, volume, features, size, data distributions of existing health datasets may not adequately support intended use cases and of the proposed system.

At the outset, real world data is rarely available for community driven health interventions. This is in part because such data is rarely aggregated in one single dataset even if collected in an appropriate manner. Typically, snippets of data are collected by the various community partners who collaborate to improve a particular health objective. As a result, community-based health intervention data is usually present in an assortment of proprietary systems administered by different organizations. This imposes costs and significant barriers in the sharing of data

Furthermore, patient level data is not often collected because the organizations that provide education, exercise, personalized care, training, and nutrition assistance are in most cases not the same as the clinics that do the tests and evaluations. If collected, such data is not shared due to HIPPA concerns. Thus, data silos are rarely interconnected, typically single patient records are transferred after seeking the explicit approval of the patient for each transfer of health records between organizations. Data collection systems that support multivendor activity-based collection of patient information are not widely adopted and are the subject of research and pilot projects.

The concept of a community health records that brings together data sets from Electronic Medical Records, Electronic Health Records, Patient Portals along with relevant meta data about collaborative partners, health improvement objectives, geography of the region has been proposed in literature (King et al., 2016). The EASI™ system is an implementation of the conceptual model of the community health record. The EASI™ data collection system comprises of three modules: a) a community partner portal, b) a patient portal, and c) Health Objective Model Specification Portal. The system is designed to support a multi-stakeholder collaborative to facilitate community health objective improvement. It supports the selection of objectives, planning of interventions, tracking implementation of the interventions, modeling using real world data, evaluation, and adjustments to the intervention to achieve desired health objective improvement targets. EASI™

can be configured to capture data as specified by the intervention model for any specific health improvement objective.

This is carried out to validate the flexible emergent data collection system design. We will show that the same system without significant programmatic changes is flexible enough to collect new data based on real world changes.

3.7 EASI Data Collection Model and Design Methodology

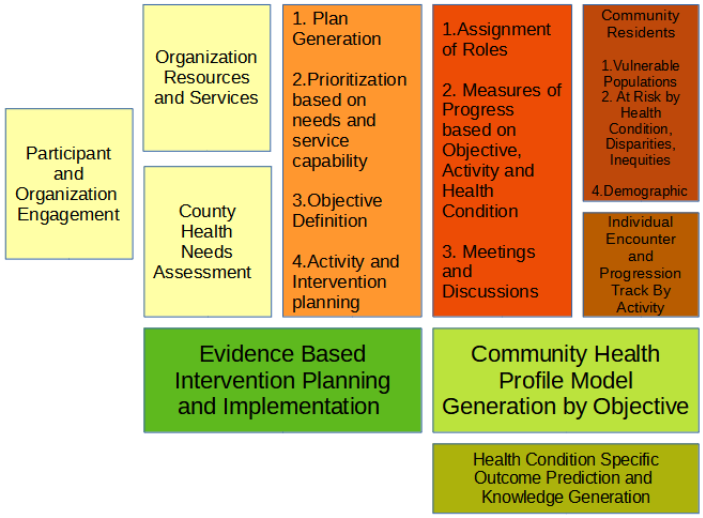


Figure 13: Community Health Activity and Engagement Record system (CHEARS) components

Figure 13 provides details of the elements of the data collection system and the data modeling systems outlined in figure 3 based on the EASI™ methodology described in section 3.2 figure 6.

3.8 EASI Methodology Summary

The EASI™ methodology is novel in that it brings to the essentially subjective process of partnership and planning of community led activity-based interventions an evidence-based approach that provides decision makers with objective feedback. It draws on recently developed simulation techniques to create a localized synthetic population level dataset and on well-

established causal / inferential techniques to model impact of community activities on targeted health outcome measures such as diabetic hospitalization. Additionally, in the specific illustrative example that is the subject of this dissertation, publicly available datasets BRFSS and SVI have been used in the generation of localized models. As illustrated in figure 7 the methodology decomposes the planning and implementation of the CHIP into modules that form the basis of a theoretically informed data collection process which is informed by the models that are appropriate for a specific health condition based on the type of data being collected. The methodology thus ties together planning, and implementation driven by subject matter experts with a data collection framework that is informed by evidence-based models.

It is important to mention that based on the arguments presented EASI™ methodology provides the flexibility to choose any appropriate modelling technique. The subject of model selection is outside the scope of this dissertation. However, we have illustrated this process by using the BRFSS data set with a few selected regression and classification models. The objective of choosing these models is to demonstrate that the data collected can be analyzed from different perspectives by the CHIP implementation partnership. The ultimate objective of model generation is to identify the most significant intervention parameters that can be manipulated by allocating resources to achieve the improvement in the health objective.

CHAPTER FOUR: DATA ANALYSIS AND RESULTS

The EASI™ methodology and the system architecture is a transdisciplinary research effort that leverages certain aspects of regression, classification, machine learning, data mining, synthetic data generation, structural equation models, Bayesian network, phenomenon of emergence, and key system design concepts to develop CHEARS architecture and health condition specific models to plan community health interventions. In this section we will analyze the MIMIC-III data set for diabetic length of stay and SVI-BRFSS data set for its ability to predict the number of diabetic hospitalizations at the census tract level.

4.1 Diabetic Hospitalization – Length of Stay MIMIC-iii Data Set

It is important to comprehend that Length of Stay, 'LOS', prediction is a well-researched healthcare application for data mining and machine learning techniques. Data mining techniques are commonly used to predict risk of adverse events in health outcomes. However, access to health data remains a challenge for several reasons including the expense of data collection, privacy issues, or laws that prevent sharing of collected data such as the Health Insurance Portability and Accountability Act (HIPAA), and limitations in ability to de-identify data while maintaining sufficient context. In emergency situations collection of extensive patient metrics might be impossible. In smaller hospitals, hospices, or clinics, access to such data might also be impossible due to lack of resources. If it is possible to determine the patient length of stay from a subset of data, resource usage and assignment can be better managed. The objective of this section is to develop and evaluate an extensible disease agnostic data analysis methodology that can readily be used to make hospital length of stay predictions.

4.1.1 Data Set Description

The length of stay (LOS) field in the dataset is a continuous variable we used as our response for our linear models. It is calculated by subtracting *admittime* from *dischtime*. It contains the length of time a person stays in the hospital. For our classification models, we further categorized *LOS* into 3 bins. The bins are: 1) 0 to 1 days 2) 2 to 5 days and 3) greater than 5 days. These bins were selected to represent 1) less severe 2) moderately severe and 3) very severe conditions. The bins were selected without consideration for any particular health condition. After further preprocessing, we reduced our original data frame to 37 predictors along with our response variable. The number of observations was reduced to 24,439 patients by removing any observations that included a null value. Also, 70% and 30% of the 24,439 patient observations were split into training and test datasets, respectively.

4.1.2 Application of Linear Regression

In this study, we utilized the best subset selection method to determine our optimal predictors. We calculated Cp, BIC, and adjusted R^2 to determine the optimal coefficients for our model. Since our initial data frame included a total of 37 possible predictors, we utilized the BIC metric to force a smaller model to predict LOS. BIC selected 23 predictors out of the original 37 as show in table 3 below.

Table 2: MIMIC-iii data linear regression model analysis

Cp	**BIC**	**Adjusted R^2**
26	23	0.439

The 23 optimal predictors were then fitted into a linear regression model along with the response, LOS, to estimate the beta coefficients that minimize RSS using a least squares approach. Our linear regression model was then used to predict LOS using unseen test data. The estimated beta

coefficients for each predictor are shown in table 4 below. The R^2 value using training data was 0.44. Using test data, the R^2 was 0.46 with an RMSE was 9.205.

Table 3: Linear regression coefficients and R2

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    34.128930   6.581929   5.185 2.18e-07 ***
first_careunit  0.474170   0.081472   5.820 5.99e-09 ***
age            -0.035401   0.006296  -5.622 1.91e-08 ***
gender         -0.614122   0.181412  -3.385 0.000713 ***
urea_n_min     -0.088998   0.017061  -5.217 1.84e-07 ***
urea_n_max      0.256593   0.009720  26.398  < 2e-16 ***
urea_n_mean    -0.259017   0.022420 -11.553  < 2e-16 ***
platelets_min  -0.052322   0.002609 -20.054  < 2e-16 ***
platelets_max   0.006176   0.001453   4.252 2.13e-05 ***
platelets_mean  0.034041   0.003514   9.688  < 2e-16 ***
magnesium_max   0.389643   0.095715   4.071 4.71e-05 ***
albumin_min    -3.396401   0.153085 -22.186  < 2e-16 ***
calcium_min     0.786550   0.124867   6.299 3.07e-10 ***
resprate_min   -0.246344   0.030312  -8.127 4.70e-16 ***
resprate_max    0.174143   0.014633  11.901  < 2e-16 ***
resprate_mean  -0.126070   0.037494  -3.362 0.000774 ***
hr_max          0.044002   0.005185   8.486  < 2e-16 ***
hr_mean        -0.065067   0.009224  -7.054 1.80e-12 ***
sysbp_mean      0.026134   0.006485   4.030 5.61e-05 ***
diasbp_min     -0.071377   0.009007  -7.924 2.43e-15 ***
diasbp_max      0.042255   0.003805  11.106  < 2e-16 ***
temp_max        2.410378   0.180139  13.381  < 2e-16 ***
temp_mean      -3.266979   0.267799 -12.199  < 2e-16 ***
status          4.713578   0.282335  16.695  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.45 on 17083 degrees of freedom
Multiple R-squared:  0.4402,    Adjusted R-squared:  0.4395
F-statistic: 584.1 on 23 and 17083 DF,  p-value: < 2.2e-16
```

4.1.3 Application of LASSO Regression

The best lambda for the LASSO regression was 0.9990724. This λ was selected using 10-fold cross-validation and comparing the resulting MSE across the different models as shown in figure 14 below. The LASSO regression model obtained an $R^2 = 0.38$ when calculated using the training data. Using the test data MSE on the LASSO model was 88.111506 giving a RMSE of 9.387. The R^2 of the LASSO execution using the test data was 0.42.

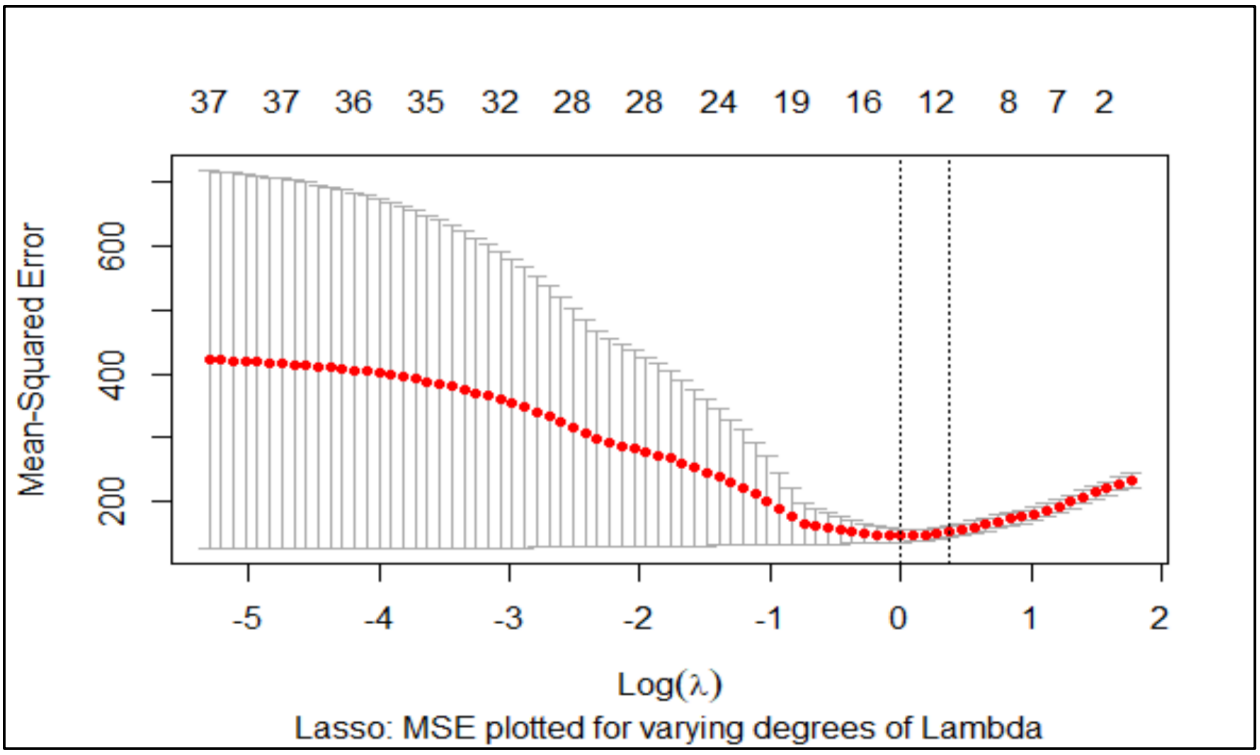


Figure 14: λ vs MSE for varying degrees of λ in Lasso modeling

LASSO reduced the coefficients of 22 out the 37 predictors to zero. Figure 15 shows the features estimated by the lasso regression. Features without values are the features which lasso estimated to have a coefficient value of zero.

```
38 x 1 sparse Matrix of class
                              s0
(Intercept)       -21.532240474
first_careunit      .
last_careunit       .
age                 .
gender              .
marital_status      .
insurance           .
urea_n_min         -0.138476739
urea_n_max          0.101684337
urea_n_mean         .
platelets_min      -0.017066975
platelets_max       0.013786778
platelets_mean      .
magnesium_max       .
albumin_min        -2.960302974
calcium_min         .
resprate_min       -0.312958889
resprate_max        0.139570566
resprate_mean       .
glucose_min        -0.004219955
glucose_max         .
glucose_mean        .
hr_min              .
hr_max              0.009467110
hr_mean             .
sysbp_min           .
sysbp_max           0.003807730
sysbp_mean          .
diasbp_min         -0.003958085
diasbp_max          0.044558963
diasbp_mean         .
temp_min           -0.142859507
temp_max            0.992907229
temp_mean           .
urine_min           .
urine_mean          .
urine_max           .
status              0.933458995
```

Figure 15: LASSO zeroed and non-zeroed coefficients

4.1.4 Application of Ridge Regression

The best lambda for the ridge regression was 81.03785. This λ was selected using 10-fold cross-validation and comparing the resulting MSE across the different models. The ridge regression model obtained an $R^2 = 0.23$ when calculated using the training data. Using the test data MSE on the ridge model was 112.073974 giving a RMSE of 10.5865. The R^2 of the ridge regression execution using the test date was 0.2646944, shown in figure 16.

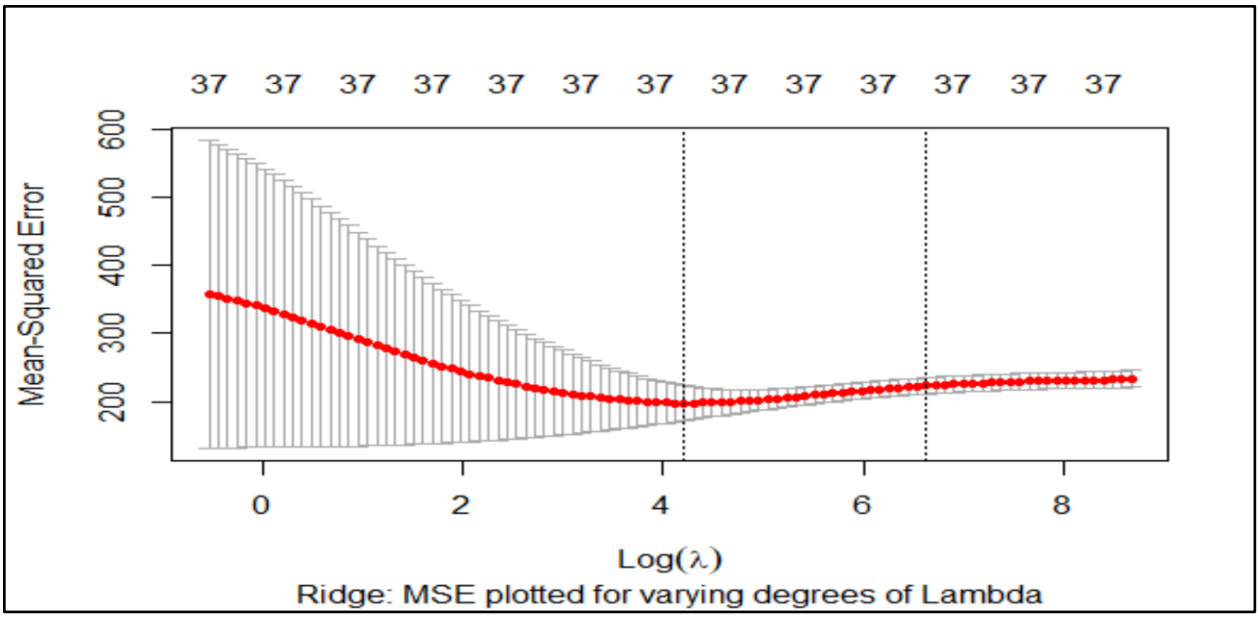


Figure 16: λ vs MSE for varying degrees of λ for Ridge modeling

The performance of our linear models using test data are summarized in the table 5 below:

Table 4: Comparison of Regression Modeling performance

	R^2	**RMSE**
Linear Regression	0.46	9.205
LASSO	0.42	9.387
Ridge	0.26	10.586

4.1.5 Application of K-Nearest Neighbors (KNN)

Using the KNN classifier with k=21 neighbors on the dataset achieved a classification accuracy of 79.10%. The relatively high accuracy of KNN showed promise that reducing the LOS prediction

to a classification problem was worthy of consideration. The confusion matrix shown in figure 17 below is for the KNN predictions using the test data.

```
> confusionMatrix_knn
          testData_labels
knn_pred     1     2     3
       1     3     1     0
       2    74   500   291
       3    84  1075  5304
```

Figure 17: KNN confusion matrix

4.1.6 Application of Multinomial Logistic Regression

Regarding our dataset, binning of our response variable was necessary to implement multinomial logistic regression. It was deemed appropriate that length of stay would be an ideal response variable. In this situation binning allows us to create 3 separate classes in our response variable which can be classified using multinomial logistic regression. The coefficients estimated by the multinomial logistic regression can be seen in table 6.

Table 5: Coefficients of multinomial logistic regression

```
Coefficients:
  (Intercept) first_careunit last_careunit        age
2   -15.85326  -0.0255602318    -0.1646757 0.00953035
3   -17.97261  -0.0007674304    -0.1859803 0.02125016
  platelets_max platelets_mean magnesium_max albumin_m
2    0.01629621   -0.018446335     0.3456066  -0.40277
3    0.02393226   -0.003790857     0.4652816  -0.82480
      hr_min     hr_max     hr_mean    sysbp_min   sys
2 0.01028864 0.01182470 -0.03188972 -0.004377842 -0.01
3 0.02060999 0.02006786 -0.04774929 -0.005754110 -0.01
    urine_min   urine_mean    urine_max   status
2 0.001586537 -0.004084611 5.764942e-05 2.192395
3 0.003976787 -0.009239663 7.094226e-04 3.738232

Std. Errors:
  (Intercept) first_careunit last_careunit        age
2 0.001246192     0.04327172    0.04333850 0.004676204
3 0.001241040     0.04362901    0.04360485 0.004838301
  platelets_max platelets_mean magnesium_max albumin_m
2   0.004362467    0.009141310    0.02559075  0.021991
3   0.004374676    0.009165547    0.02502505  0.021300
      hr_min     hr_max     hr_mean    sysbp_min    sy
2 0.004877895 0.005080243 0.008463453 0.005595558 0.00
3 0.005236753 0.005204505 0.008839441 0.005737685 0.00
    urine_min  urine_mean    urine_max     status
2 0.001366739 0.001198086 1.325022e-04 0.04566328
3 0.001482845 0.001242746 9.676289e-05 0.04353639

Residual Deviance: 12043.4
AIC: 12195.4
```

Table 7 and figure 18 below show a sample bin membership probability for 6 randomly chosen observations in our dataset as estimated by the multinomial regression. We can see for this sample that bins 2 and 3 received probabilities > 0 with bin 3 receiving the highest probability in each case. Here the multinomial regression is used to predict the probability of a particular observation to be a member of a particular classification level. It is important to note that the classification levels here are represented by the columns while the observations are represented by the rows. This means that these six observations are classified as in the third category which represent patients with a length of stay greater than 5 days. The 6 samples shown in the figure below were all correctly classified. A confusion matrix is a useful tool for evaluating the performance of a classification model. As it can be observed, our accuracy using the test data set is high at 82%.

Table 6: Bin probabilities for 6 random observations

	1	2	3
7867	0	0.03	0.97
11059	0	0.24	0.76
17069	0	0.02	0.98
26907	0	0.10	0.90
5980	0	0.00	1.00
26612	0	0.00	1.00

```
> confusionMatrix(test_pred, as.factor(testData$LOS))
Confusion Matrix and Statistics

          Reference
Prediction    1    2    3
         1   64   30    8
         2   82  819  430
         3   15  727 5157

Overall Statistics

               Accuracy : 0.8238
                 95% CI : (0.8149, 0.8324)
    No Information Rate : 0.7631
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4918

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity          0.397516   0.5197   0.9217
Specificity          0.994701   0.9110   0.5728
Pos Pred Value       0.627451   0.6153   0.8742
Neg Pred Value       0.986584   0.8739   0.6943
Prevalence           0.021959   0.2149   0.7631
Detection Rate       0.008729   0.1117   0.7034
Detection Prevalence 0.013912   0.1815   0.8046
Balanced Accuracy    0.696108   0.7154   0.7473
```

Figure 18: Multinomial logistic regression results.

4.2 Diabetic Hospitalization with BRFSS Data Analysis

In this study, we developed various linear and classification models to predict the length of stay of a patient in a hospital. We used linear regression, LASSO, as well as Ridge regression for our linear models and KNN along with multinomial logistic regression for our classification models. We utilized various predictors that stem from data collected from patient demographics and simple medical tests such as blood and urine tests. Linear regression was our best performing linear model with the highest R^2 of 0.46 and lowest RMSE of 9.205. It is known in the biological sciences that having a low value for R^2 is possible due to the existence of irreducible error and inherent uncertainties of biological systems. Since our linear regression has a low R^2 which means the model has a poor fit, we decided to group our response variable, LOS, into bins to allow the use

of classification models. Our multinomial logistic regression was our best performing classification model with the highest prediction accuracy of 82.38% for test data. The prediction accuracy we achieved is greater than the prediction accuracy reported by Wang et al. which were in the range of 68% - 69% for a bin of LOS >3 days. This analysis demonstrates that it is possible to predict LOS independent of disease information.

In conclusion, we have developed a disease agnostic framework to predict length of stay and evaluated both regression and classification models. We utilized the mimic III dataset to model real world predictors. While excluding disease conditions in our analysis, our best model utilized multinomial logistic regression to achieve a high prediction accuracy. The outcome and associated implications of our findings and our approach can potentially facilitate the addition of other predictors such as disease specific parameters. We believe that addition of this extra information will further improve the accuracy of our model. Based on literature and intuition it stands to reason that inclusion of such parameters should result in a favorable increase in prediction accuracy since it is likely that disease and severity will impact length of stay.

The purpose of this section and data analysis is to demonstrate a simplistic approach to localized diabetic hospitalization prediction model. While the results of the model are not the subject of the dissertation an illustrative example to demonstrate the EASI™ methodology serves as a practical guide on how the methodology can benefit community health improvement planners. As explained in previous sections the intuition for this analysis is that local social vulnerability and behavior impact the risk of developing diabetes. A further objective of this project is to utilize the MIMIC-iii hospitalization data for diabetics to determine the LOS. The analysis demonstrates that with appropriate selection of key feature it is possible to develop localized models that indicate the outcomes such as length of hospital stay are impacted by local social vulnerability factors. Previous

literature and studies have already indicated the same. The contribution of the EASI™ methodology is that it outlines a process of leveraging and modifying existing models to develop a theoretically informed data collection system and selection of appropriate evidence-based model to these are that are not included in typical EHR data collected in the hospital upon admission. We aim to create an integrated synthetic dataset for the purposes of CHIP objectives.

The census tract level social vulnerability index data is used as counter to draw samples from the BRFSS dataset. The synthetic data represents of the population of given census tract.

This dissertation study provides a foundational framework for utilization of evidence-based strategies in community health interventions. Previously, LOS with respect to diabetes was predicted using a combination of clinical and demographic features based on data collected in the hospital. Here it is important to focus the analysis on: a) diabetic prediction using the BRFSS data set, and b) description of the SVI data set to validate/invalidate the central hypothesis described in chapter 1.

4.2.1 Description of the Key predictors and response variable in BRFSS

Practitioners and researchers typically identify the following as important risk factors for diabetes. Below is a list of behavioral features that are extracted and cleaned from the BRFSS data set.

The response variable from the BRFSS data set is the DIABETE3.

- blood pressure (high): Adults who have been told they have high blood pressure by a doctor, nurse, or other health professional --> _RFHYPE5
- cholesterol (high): Have you EVER been told by a doctor, nurse or other health professional that your blood cholesterol is high? --> TOLDHI2

  Cholesterol check within past five years --> _CHOLCHK

Other Important Features from the BRFSS data that we consider while predicting the diabetic status of the individual include the following: age, sex, race, education, health care coverage, household income, marital status, diet, smoking, obesity, exercise, alcohol consumption, BMI, sleep, time since last checkup, and mental health (figure 19).

Many of these features are directly and significantly impacted by localized social vulnerability. Social vulnerability can be mitigated by intentional application of community resources to targeted population. This is the key significance of the EASI™ methodology.

```
# select specific columns
brfss_df_selected = brfss_2015_dataset[['DIABETE3', '_STATE',
                                        '_RFHYPE5',
                                        'TOLDHI2', '_CHOLCHK',
                                        '_BMI5',
                                        'SMOKE100',
                                        'CVDSTRK3', '_MICHD',
                                        '_TOTINDA',
                                        '_FRTLT1', '_VEGLT1',
                                        '_RFDRHV5',
                                        'HLTHPLN1', 'MEDCOST',
                                        'GENHLTH', 'MENTHLTH', 'PHYSHLTH', 'DIFFWALK',
                                        'SEX', '_AGEG5YR', 'EDUCA', 'INCOME2','EMPLOY1','CHILDREN','_RACE' ]]
```

```
brfss_df_selected.shape
```

```
(441456, 26)
```

Figure 19: BRFSS data set feature selection

Selected SVI features that might influence Behavior and resulting diabetic hospitalization as shown in figure 20.

```
105]: # select specific columns

      svi = svi[['FID','TRACTCE', 'STATE','COUNTY',
                 'FIPS', 'E_TOTPOP','M_TOTPOP','E_UNEMP',
                 'M_UNEMP', 'E_AGE65','M_AGE65','E_AGE17', 'M_AGE17',
                 'E_DISABL','M_DISABL', 'E_MINRTY','M_MINRTY',
                 'EP_POV', 'EP_LIMENG', 'EP_AGE65', 'EP_MINRTY',
                 'EP_UNEMP', 'EP_NOHSDP', 'EP_AGE17', 'EP_DISABL',
                'EP_SNGPNT', 'EP_MUNIT', 'EP_MOBILE', 'EP_CROWD', 'EP_NOVEH', 'EP_GROUPQ', 'RPL_THEMES'

              ]]
      svi.head()
```

Figure 20: SVI dataset features

The 2015 BRFSS dataset was used for illustrative purposes. A total of 26 features were initially selected from the entire data set with 441,456 rows. Missing values were dropped which resulted in 343,606 rows of data. The data shape is helpful because it gives an understanding of the data set sizes that are typically made available by the CDC. Each feature has multiple codes for different levels and no responses. To illustrate by example these codes were simplified and summarized in this analysis. In real world settings, however, a more considered, approach is to be taken while recoding variables. In our simplified model only the BMI feature is numeric while the other features are categorical, and binomial in nature. Here it is important to list a few key observations and tasks performed in the analysis:

- Binary classification of diabetes vs no diabetes that was accomplished by either joining the prediabetics with the diabetics, with the non-diabetics, or removing them entirely
- We add the prediabetic to the non-Diabetic group as these people are not yet diagnosed
- 50-50 split of non-diabetics to diabetics and prediabetics to balance the dataset
- The dataset had 35346 (diabetes)
- There dataset had 4631 (pre-diabetes) + 213703 (non-diabetic) so we can make a new 50-50 binary dataset of 218334 non-diabetic individuals, as shown in figure 21 and 22.

```
#Check Class Sizes of the heart disease column
brfss_df_selected.groupby(['DIABETE3']).size()
```

```
DIABETE3
0.0    213703
1.0      4631
2.0     35346
dtype: int64
```

Figure 21: Raw data non-diabetic, pre-diabetic and diabetic

```
: #Show the change
  brfss_binary.groupby(['Diabetes_binary']).size()
  #brfss_binary.info()

: Diabetes_binary
  0.0    218334
  1.0     35346
  dtype: int64
```

Figure 22: Merging Pre-diabetic and Non-Diabetic

To run machine learning algorithms, it is helpful to balance the data set. This is because we seek to predict the condition of diabetes based on selected features in the BRFSS dataset. From a CHIP perspective this is helpful because we can target selected groups for specific interventions and resource allocation, as shown in figure 23. Another benefit of this approach is we can build causal SEM models to understand and determine quantitively the association between the selected predictors and the target response variable. The EASI™ methodology thus has the flexibility in incorporating various models generated using low-cost public data set to determine features that have high predictive value and importance in allocation community resources.

Figure 23: The premilitary data exploration - An optimization problem in resource allocation

4.2.2 BRFSS Exploratory Data analysis

A preliminary exploratory analysis of public data indicates that this is a multi-dimensional optimization problem that has subjective as well as evidence-based components. The analysis of key features was conducted, as an example we show a bar chart of the distribution of the BMI feature, shown in figure 24.

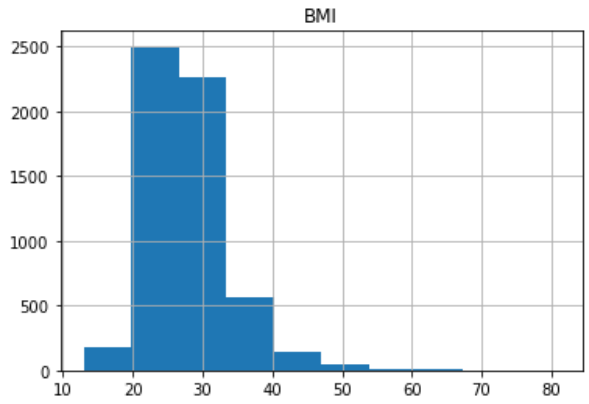


Figure 24: BMI distribution bar chart

Further exploratory analysis was conducted, as an example we show a histogram between age and number of mental health days. In figures 25, 26, 27 that follow we plot basic features to get a strong intuition on covariance structure and correlation amongst key features. This is an illustration of a simple method that can be readily adopted by community health planner while discussing specific activities and intervention plans.

```
import plotly.express as px

fig = px.histogram(brfss_florida_unbal, x="Age", color="Race",pattern_shape="MentHlth")
fig.show()
```

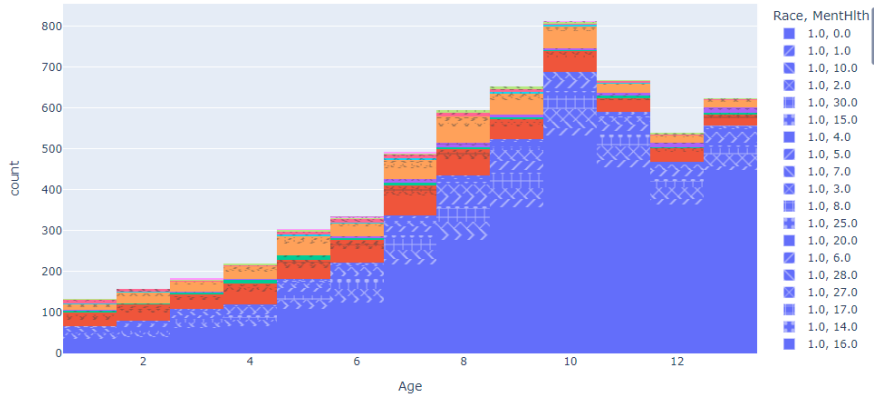


Figure 25: Age vs Mental Health days

The age variable was grouped into 12 bins. Racial mix and prevalence of mental health issues was plotted for each bin. The dataset used was not balanced. This chart gives a visual representation on key population characterizes.

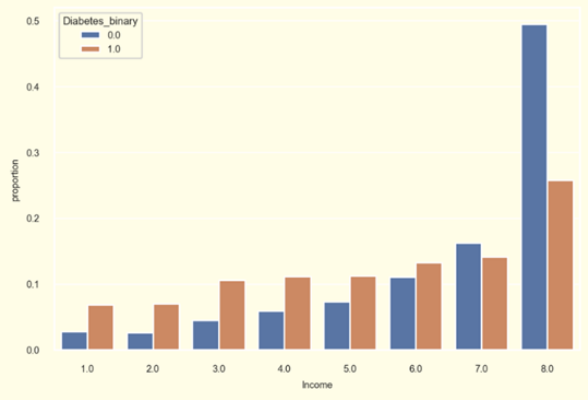


Figure 26: Income and Diabetic rate as a proportion

To understand the impact of income on diabetic rate the figure 26 was plotted using the BRFSS data. The dataset was balanced to isolate the impact of income distribution. It shows that at higher income level the proportion of diabetes significantly reduced. Incidentally, there seems to be a

income threshold of $75,000 per year for this effect. At very low-income levels under $50,000 the rate of diabetic is greater that the rate of non-diabetics.

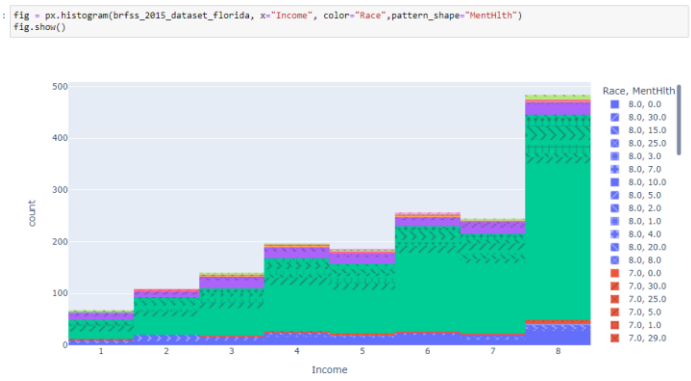
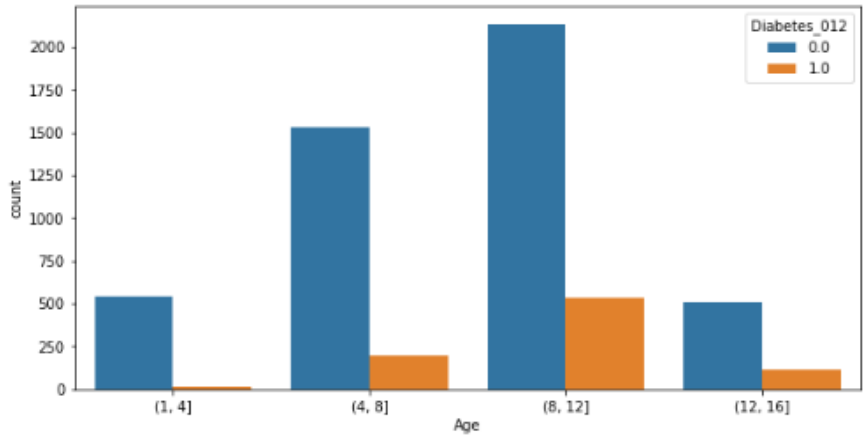


Figure 27: Income and Mental health unbalanced data

Figure 27 confirms general finding in literature that as a proportion of the population people with higher incomes have lower incidents with mental health.

```
plt.figure(figsize=(10,5))
bins=[1,4,8,12,16]
sns.countplot(x=pd.cut(brfss_florida_unbal.Age,bins=bins),hue=brfss_florida_unbal.Diabetes_012)
plt.show()
```




```
plt.figure(figsize=(10,5))
bins=[1,4,8,12,16]
sns.countplot(x=pd.cut(brfss_florida_unbal.Age,bins=bins),hue=brfss_florida_unbal.HighBP)
plt.show()
```

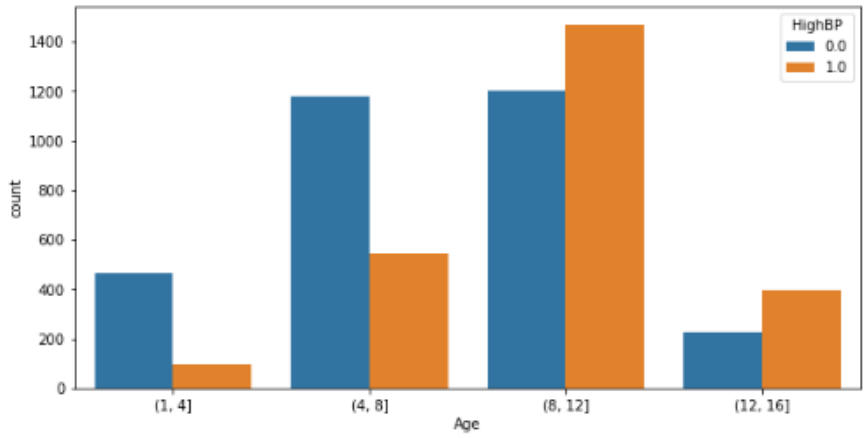
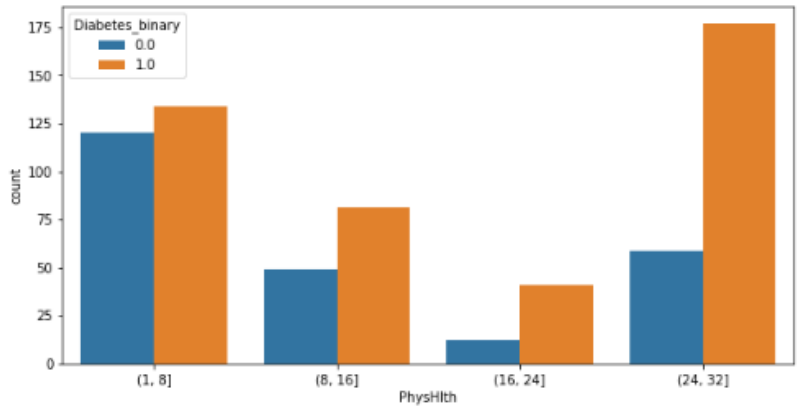


Figure 28:(a) Age and diagnosis of diabetic diagnosis (b) Age and diagnosis of High BP in the Florida BRFSS data set

Figure 28 delineates that high blood pressure is a highly correlated with age. While the likely hood of diabetes increases with age it is not the association is not as strong as in the case of high blood pressure.

```
plt.figure(figsize=(10,5))
bins=[1,8,16,24,32]
sns.countplot(x=pd.cut(brfss_2015_dataset_florida.PhysHlth,bins=bins),hue=brfss_2015_dataset_florida.Diabetes_binary)
plt.show()
```




```
plt.figure(figsize=(10,5))
bins=[1,10,20,30,40,50]
sns.countplot(x=pd.cut(brfss_2015_dataset_florida.BMI,bins=bins),hue=brfss_2015_dataset_florida.Diabetes_binary)
plt.show()
```

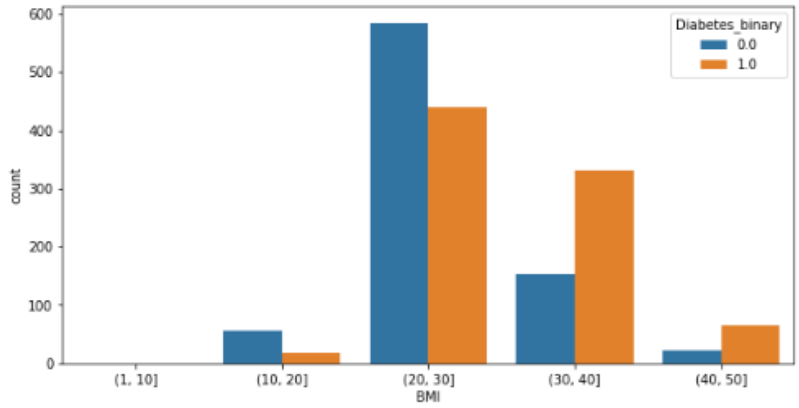


Figure 29: BMI and Physical Health showing increase in diabetes when BMI is high and physical health is poor

Figure 29 illustrates that BMI and Physical Health are strong predictors of diabetes. To get an overall sense of correlations among the predictor variables a correlation plot was generate and is shown in the figure below.

```
plt.figure(figsize=(15,12))
sns.heatmap(brfss_2015_dataset_florida.corr(), annot=True,cmap='Accent_r')
plt.show()
```

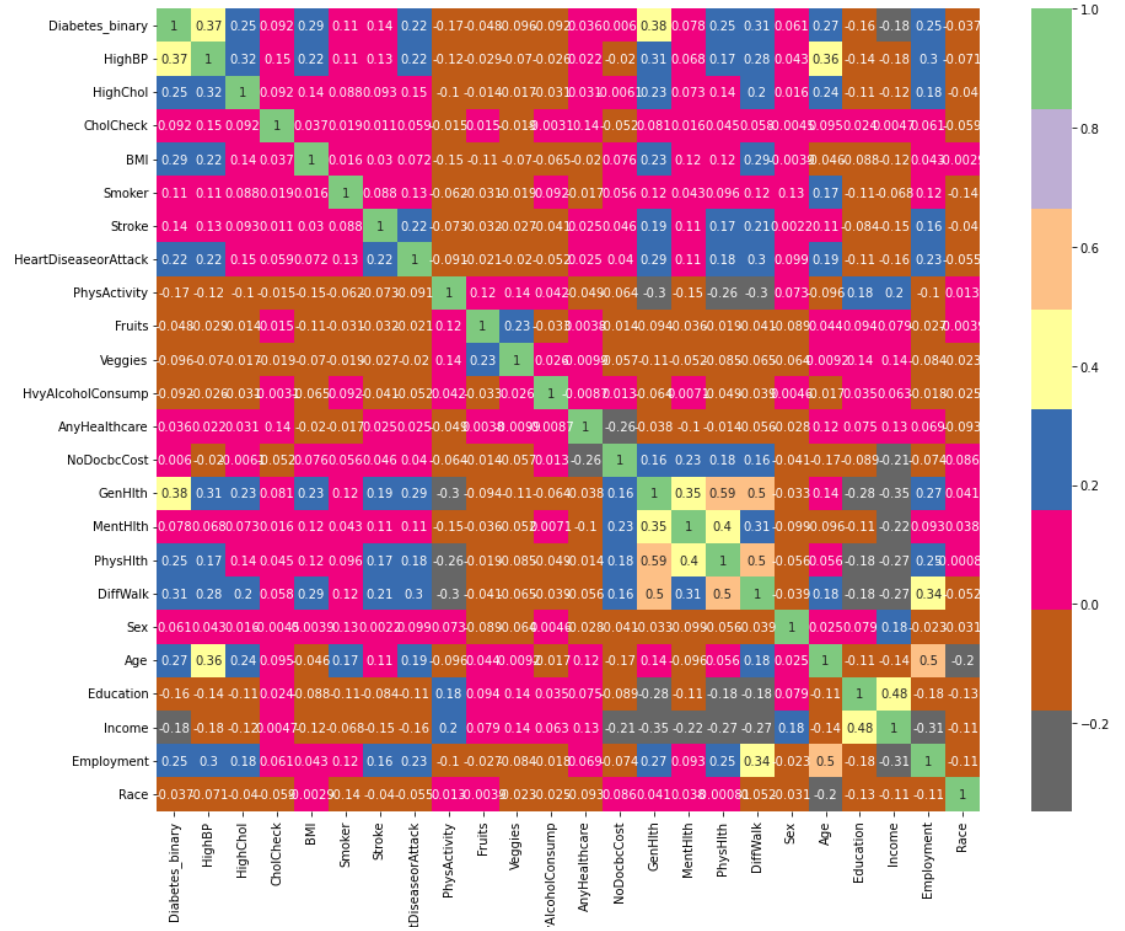


Figure 30: Correlation plot of observed data (predictors and response variable)

The correlation plot includes numeric, categorial and binomial variables. The correlation plot, (figure 30) indicates high blood pressure, general health, a history of heart disease, physical health, BMI and income have strong correlation with a positive diabetic outcome. The data appears to have a multicollinearity problem and it was checked as shown in figure 31 below. While several features showed an extremely high degree of multicollinearity some were selected to be dropped, specifically education level and if a person has checked their cholesterol. This is because this information is likely to be captured by the other variables.

```
calc_vif(x)
```

	variables	VIF
0	HighBP	1.733624
1	HighChol	1.631744
2	CholCheck	16.460351
3	BMI	17.545114
4	Smoker	1.731445
5	Stroke	1.103532
6	PhysActivity	5.180710
7	Fruits	2.995858
8	Veggies	6.309420
9	HvyAlcoholConsump	1.073032
10	AnyHealthcare	15.452221
11	NoDocbcCost	1.287795
12	GenHlth	9.759439
13	MentHlth	1.511619
14	PhysHlth	1.893655
15	DiffWalk	1.723138
16	Sex	1.919867
17	Age	8.559954
18	Education	32.973123
19	Income	20.350436
20	Employment	2.876433
21	Children	4.424651
22	Race	4.210999
23	HeartDiseaseorAttack	1.193317

```
#### If the VIF is above 5 consider removal of the variable based on the correlation lot based on this we consider dropping CHOL
```

Figure 31: Checking for Multicollinearity

```
sns.set(style="whitegrid")
sns.pairplot(x_numeric_norm, kind='reg', diag_kind='kde')
plt.show()
```

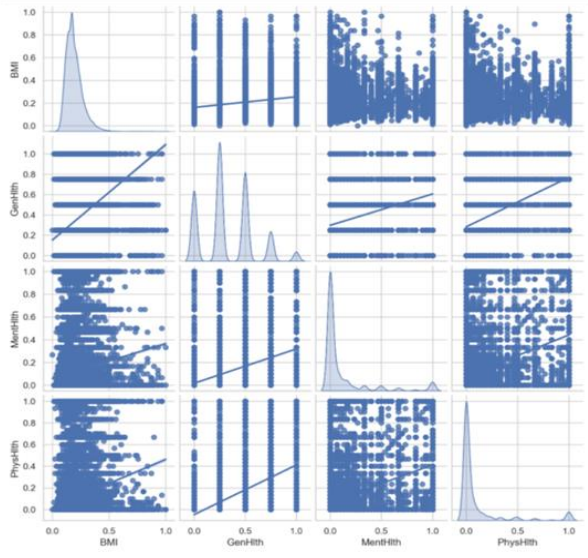


Figure 32: Pair plot for the numeric features

Figure 32 above show pair plots and the correlation matrix suggest that many of the features might have similar level of impact on the outcome with a few key features having the maximum impact. A Principal Component Analysis (PCA) was conducted to identify how many components are

needed to account for the variance. The PCA in figure 33, 34, 35 below shows that a few features account for most of the variations.

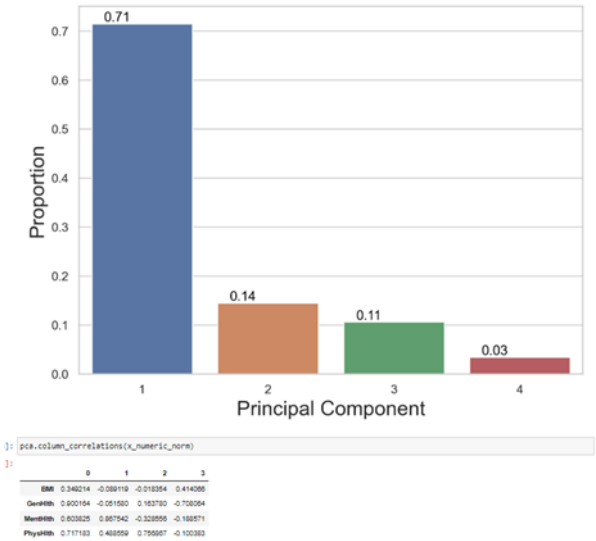


Figure 33: PCA for the numeric data

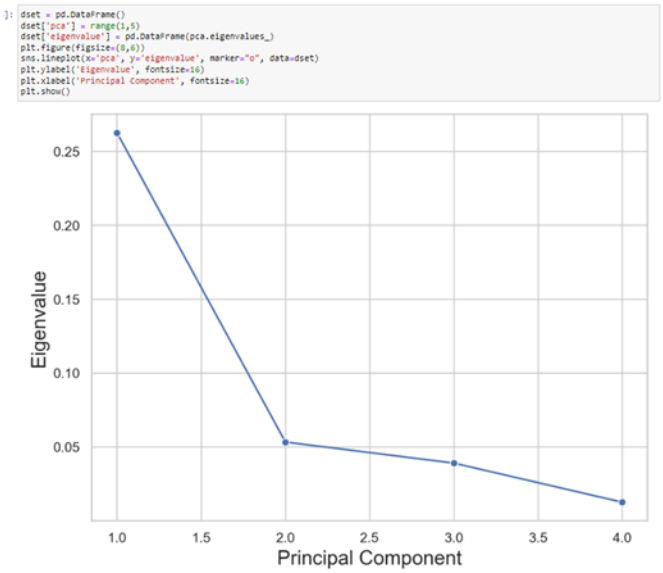


Figure 34: Elbow graph of the PCA for the numeric features

Literature suggests that there is a high causal effect between BMI, General Health status and diabetes so the numeric data was separated to see if just the numeric data can be used in the prediction.

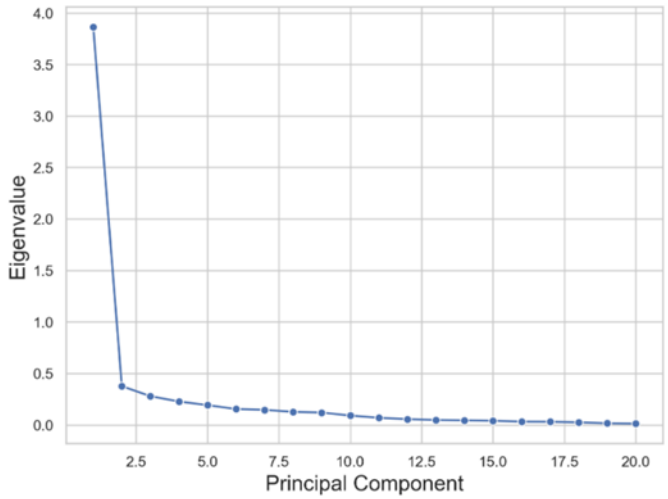


Figure 35: PCA Elbow graph representation using all features

PCA suggest that just a few components account for most of the variability in the data set. This is an opportunity for significant dimensionality reduction. With 3 to 5 components the data can modelled to fit well. This is a key finding and is of interest when applying the EASI™ methodology step 4, figure 4.

4.2.3 Modeling with the BRFSS data set – An illustrative example

Train, Test data sets (row selection), k-fold cross validation table. Split into categorical and numerical table (by variable types). The data was standardized. PCA was performed and unsupervised machine learning algorithms were run. Key results are listed in the section below. Gaussian mixture models were compared with K-Means for the complete data set. K-Prototypes models was run to better handle the mixture of categorical and numeric features.

Area Under Curve (AUC) & Receiver Operating Characteristic (ROC) curves were compared for selected supervised machine learning models.

4.2.3.1 Unsupervised Machine Learning Models

Unsupervised learning describes the structure of the data. It is useful in clustering similar data elements (rows). It also helps to reduce the dimensionality of the data. PCA is a technique used in reducing the dimensionality of the data, as previous shown.

a. Gaussian Mixture Model vs K-Means for All features

This chart shows the Gaussian Mixture Model (GMM) can work with both categorial and numeric features while the K-Means works with numeric features. K-means groups data pointed based on distances computed from a centroid for the group whereas GMM performs a probabilistic assignment of data point to a group. As can be seen in figure 36 below, the silhouette analysis for optimal K the GMM outperforms the K- Means. This is expected because the data has categorial and numerical data and therefore probabilistic assignment will outperform Euclidean distance measures.

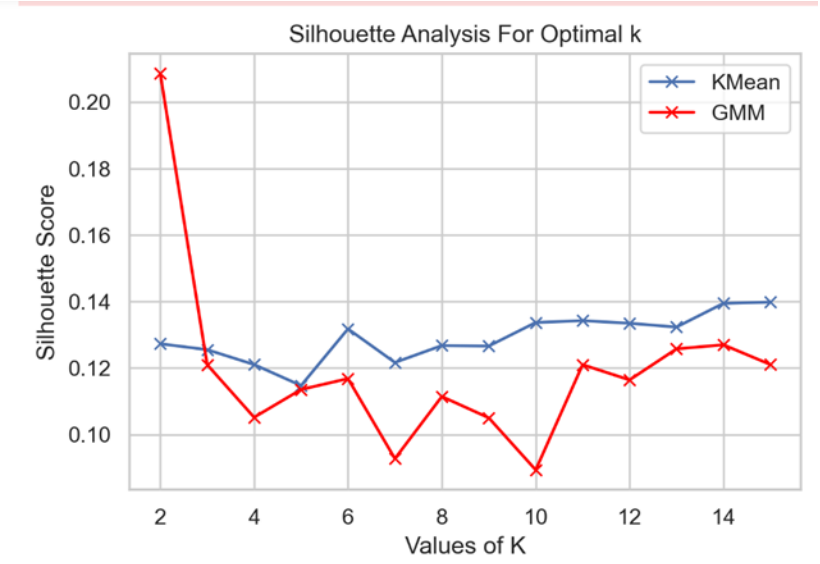


Figure 36: GMM vs K-Means for categorical and numeric features

Figure 36, demonstrates that the GMM clustering technique works well with an optimal value of 4.

b. T-SNE scatter plot

Given the mix of categorical and numerical features a T-SNE scatter plot was generated to get a visual representation of clustering, shown in figure 37 below.

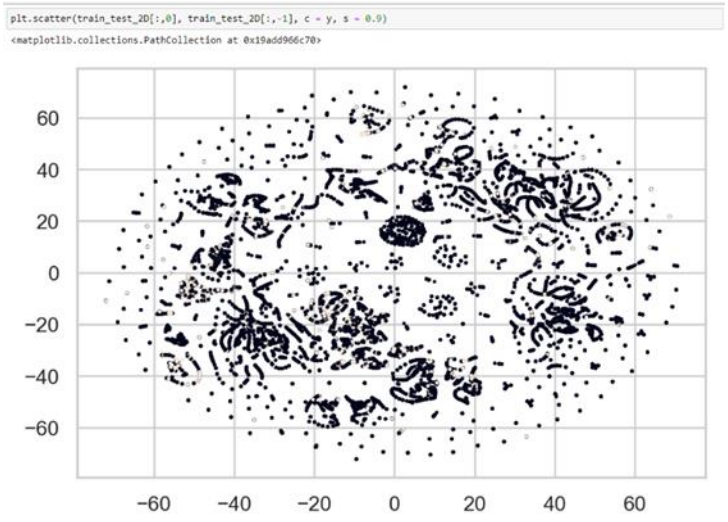


Figure 37: T-SNE scatter plot

T-distributed stochastic embedding (t-SNE) algorithm is useful in non-linear dimensionality reduction. This plot is significant because we have had several non-linear associations. It is a high dimensional data set wherein the separation between clusters can occur in several planes and dimensions. We chose a large number for perplexity and given the large data set the algorithm is computationally intensive. The T-SNE performs a binary search for the value of sigma that produces a probability distribution with a user specified perplexity of 30.

c. K-Prototype for both Categorial and Numerical Data.

To further analyze the mix of categorial and numerical data a K – prototype analysis was performed to overcome the limitations of the K-Means classifier. The results are shown in figure 38 below.

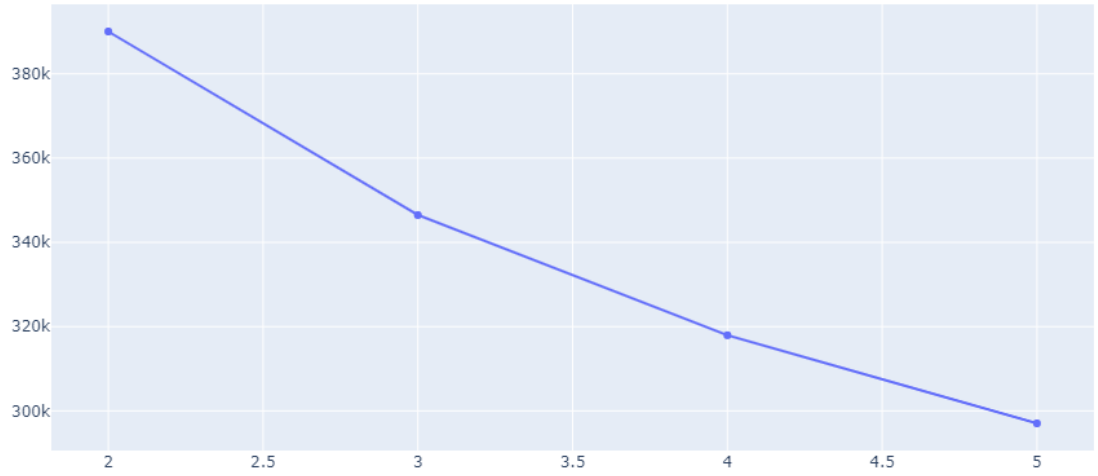


Figure 38: K-prototypes after 10 runs with tqdm range (2 and 6)

The K-Prototypes algorithm works with both categorical and numeric data. It measures the distance between numerical features using Euclidean distance, while also measuring the distance between categorical features using the number of matching categories. The k prototypes are computationally intensive. Given the PCA and TSNE analysis 4 clusters will be optimal for the dataset.

4.3.2.2 Supervised Machine Learning Model

```
1]: print(classification_report(y_test,y_pred))
              precision    recall  f1-score   support

         0.0       0.93      0.99      0.96     17599
         1.0       0.36      0.08      0.14      1471

    accuracy                           0.92     19070
   macro avg       0.65      0.54      0.55     19070
weighted avg       0.88      0.92      0.89     19070

2]: print("{} NN Score: {:.2f}%".format(n_neighbors, knn.score(x_test, y_test)*100))
2 NN Score: 91.80%
```

Figure 39: KNN model metrics

Figure 40: The KNN performs well with more than 2 clusters

The K nearest neighbors (KNN) performs well with two clusters and there is no appreciable improvement beyond five clusters as shown in figure 39,40 above.

a. The Decision Tree Model

```
print(classification_report(y_pred_tree,y_test))
              precision    recall  f1-score   support

         0.0       0.92      0.94      0.93     17272
         1.0       0.28      0.23      0.25      1798

    accuracy                           0.87     19070
   macro avg       0.60      0.58      0.59     19070
weighted avg       0.86      0.87      0.87     19070
```

b. The Random Forest Model

```
]: print(classification_report(y_pred,y_test))
```

	precision	recall	f1-score	support
0.0	1.00	0.92	0.96	19065
1.0	0.00	0.60	0.00	5
accuracy			0.92	19070
macro avg	0.50	0.76	0.48	19070
weighted avg	1.00	0.92	0.96	19070

c. LGBM

```
:
#Append Model Name
model_name.append("Lgbm")
```

```
: print(classification_report(y_pred,y_test))
```

	precision	recall	f1-score	support
0.0	0.99	0.93	0.96	18602
1.0	0.16	0.51	0.25	468
accuracy			0.92	19070
macro avg	0.57	0.72	0.60	19070
weighted avg	0.97	0.92	0.94	19070

d. Logistic Regression Model

```
3]: print(classification_report(y_pred,y_test))
```

	precision	recall	f1-score	support
0.0	0.99	0.93	0.96	18631
1.0	0.17	0.56	0.26	439
accuracy			0.93	19070
macro avg	0.58	0.75	0.61	19070
weighted avg	0.97	0.93	0.94	19070

All the above models have low precision. The logistic regression model outperforms these models as expected.

e. Model Comparison

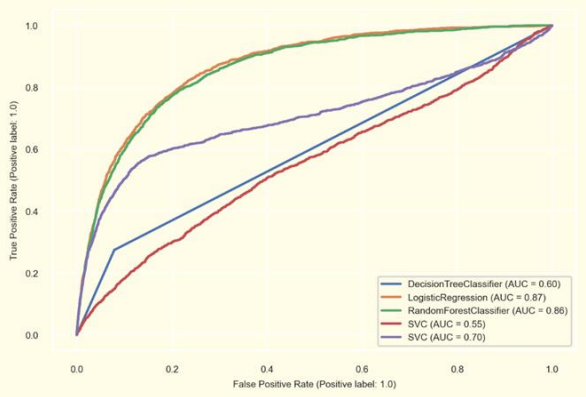


Figure 41: Comparison of Supervised Machine Learning Models

The AUC suggest that the logistic regression model is most likely to be the best model for the dataset shown in figure 41 above. This finding is consistent with findings in literature that reports the utility of logistic regression as a prediction model in health outcomes.

A full Logistic Regression Analysis was performed and results are shown in figure 42 below.

```
Call:
glm(formula = Diabetes_binary ~ ., family = binomial, data = train)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-2.4217  -0.3437  -0.2069  -0.1323   3.5349

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -8.567349   0.273027 -31.379  < 2e-16 ***
HighBP                0.788012   0.044379  17.756  < 2e-16 ***
HighChol              0.761255   0.042151  18.060  < 2e-16 ***
CholCheck             1.464092   0.197856   7.400 1.36e-13 ***
BMI                   0.048459   0.002470  19.621  < 2e-16 ***
Smoker                0.057943   0.042136   1.375 0.169085
Stroke                0.208393   0.096117   2.168 0.030150 *
HeartDiseaseorAttack  0.221552   0.069775   3.175 0.001497 **
PhysActivity         -0.018136   0.045377  -0.400 0.689397
Fruits               -0.040922   0.042517  -0.962 0.335805
Veggies               0.057748   0.050989   1.133 0.257399
HvyAlcoholConsump    -0.613320   0.123587  -4.963 6.95e-07 ***
AnyHealthcare         0.267848   0.081298   3.295 0.000985 ***
NoDocbcCost           0.101583   0.057384   1.770 0.076688 .
GenHlth               0.630038   0.025870  24.354  < 2e-16 ***
MentHlth             -0.005621   0.002500  -2.249 0.024525 *
PhysHlth             -0.007750   0.002617  -2.962 0.003061 **
DiffWalk              0.069442   0.060454   1.149 0.250690
Sex                   0.135297   0.042742   3.165 0.001548 **
Age                   0.197096   0.009899  19.911  < 2e-16 ***
Education            -0.057547   0.021606  -2.664 0.007733 **
Income               -0.040603   0.012010  -3.381 0.000723 ***
Employment            0.021228   0.009178   2.313 0.020725 *
Children             -0.003194   0.020055  -0.159 0.873458
Race                 -0.392410   0.044488  -8.821  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24342  on 44396  degrees of freedom
Residual deviance: 17782  on 44372  degrees of freedom
AIC: 17832
```

```
# gives 95% confidence interval of coefficients

confint(model_logistic_full)
```

```
Waiting for profiling to be done...
                            2.5 %        97.5 %
(Intercept)          -9.115730153 -8.0437006545
HighBP                0.701101195  0.8750771860
HighChol              0.678698383  0.8439394088
CholCheck             1.096116980  1.8745248837
BMI                   0.043602856  0.0532857030
Smoker               -0.024750688  0.1404291308
Stroke                0.018820137  0.3956907011
HeartDiseaseorAttack  0.084233950  0.3577774433
PhysActivity         -0.106799786  0.0710847980
Fruits               -0.124150722  0.0425238167
Veggies              -0.041750899  0.1581394876
HvyAlcoholConsump    -0.862492928 -0.3774906451
AnyHealthcare         0.110164575  0.4289379838
NoDocbcCost          -0.011451461  0.2135124723
GenHlth               0.579476752  0.6808916581
MentHlth             -0.010545695 -0.0007469094
PhysHlth             -0.012898960 -0.0026401773
DiffWalk             -0.049355617  0.1876353734
Sex                   0.051510577  0.2190671025
Age                   0.177725335  0.2165302710
Education            -0.099830134 -0.0151312977
Income               -0.064100514 -0.0170202962
Employment            0.003200844  0.0391786791
Children             -0.042723973  0.0358978858
Race                 -0.479483628 -0.3050837499
```

Figure 42: (a) Full Logistic Model with All Features (b) 95% confidence intervals of the coefficients

The logistic model is appropriate because of the binary outcome (presence or absence of diabetes) that is being predicted. Given the pervious findings a stepwise forward selection was performed to select key features as shown in figure 43 below.

```
## generate the step wise model from Full Model
```{r}
model_logistic_step <- model_logistic_full%>%  stepAIC(trace = FALSE)
# Summarize the final selected model
coef(model_logistic_step)


```
```

(Intercept)	HighBP	HighChol	CholCheck	BMI
-8.599877743	0.789992114	0.761316959	1.459435780	0.048864697
Smoker	Stroke	HeartDiseaseorAttack	HvyAlcoholConsump	AnyHealthcare
0.063256326	0.212515972	0.224434541	-0.611666229	0.272134442
NoDocbcCost	GenHlth	MentHlth	PhysHlth	Sex
0.105117484	0.634989440	-0.005426052	-0.006909643	0.133385894
Age	Education	Income	Employment	Race
0.198266899	-0.056950856	-0.041450987	0.023181544	-0.388358724

Figure 43: A step wise feature selection

This confirms previous analysis and identifies High BP, High Cholesterol and BMI as key features of interest in predicting diabetes. From an intervention design perspective such rigorous analysis on localized data set is critical in identifying activities for resource allocation. For example, from our analysis a possible recommendation is that monitoring High BP, BMI for people with low-income levels is likely to identify people who are at high risk of developing diabetes or risk of readmission. Mitigation strategies can be designed implemented and evaluated using the same models to quantify improvement. However, only monitoring BMI is not likely to help in identification of target at risk population.

To test this theory a logistic regression model with only BMI as generated as show in figure below.

```
Call:  glm(formula = Diabetes_binary ~ BMI, family = binomial(link = "logit"),
    data = brfss_binomial_logistic)

Coefficients:
(Intercept)          BMI
   -4.72459      0.07514

Degrees of Freedom: 63564 Total (i.e. Null);  63563 Residual
Null Deviance:      35110
Residual Deviance: 33230        AIC: 33230
```

Figure 44: Prediction of diabetes with only BMI

Confirming literature, just BMI alone is not a good predictor of diabetes. The coefficient is low 0.07514 as shown in figure 44 above. The full model shows HighBP and HighColestrol have a greater utility as predictors of diabetes than BMI and many of the features. This is finding is consistent with exploratory analysis and literature. This establishes the need for careful identification of controllable parameter and forms the basis for combining BRFSS and SVI dataset via the GAN in future work.

The supervised machine learning models used an unbalanced data set, and this can result in poor precision, recall and f-1 scores. In the context of the overall population the proportion of diabetics is low.

Due to imbalanced data the accuracy predition about diabates are low. SMOTE + ENN to fix

```
[199]: ### SMOTE --> Synthetic Minority Oversampling Technique

[200]: sm = SMOTEENN()
       x_resampled, y_resampled = sm.fit_resample(x,y)

[201]: xre_train,xre_test,yre_train,yre_test = train_test_split(x_resampled, y_resampled, test_size=0.3, random_state=42)

[202]: knn_smote = KNeighborsClassifier(n_neighbors = 5)
       knn_smote.fit(xre_train,yre_train)

[202]: KNeighborsClassifier()

[203]: yre_pred = knn_smote.predict(xre_test)

[204]: print(classification_report(yre_test,yre_pred, labels=[0,1]))
```

	precision	recall	f1-score	support
0	1.00	0.89	0.94	12453
1	0.93	1.00	0.96	17417
accuracy			0.96	29870
macro avg	0.96	0.95	0.95	29870
weighted avg	0.96	0.96	0.96	29870

Figure 45: Increase in precision, recall and F-1 score of KNN with SMOTE ENN

Running the KNN after adjusting with SMOTE ENN produced the best results with high precision in correctly predicting both diabetics and non-diabetics as can be seen in figure 45 above.

These models show various modeling techniques are helpful in the analysis of a particular dataset. The results can be impacted by several factors including data coding, collection and preparation. The selection of appropriate machine learning models for specific type health data is outside the scope of this dissertation. These models and the BRFSS data features selection provide an illustrative example of how Community Activity Intervention Model (CAIM) for specific target health condition can be informed by data and theory that is facilitated by the CHEARS™ data collection model. The EASI™ methodology provides guidance to community health practitioners to adopt evidence-based strategies for data collection and model evaluation.

Summary of Key findings from Prediction Models in illustrative example with BRFSS data is as follows:

1. Logistic model is an appropriate model for prediction using BRFSS data. There here is opportunity of significant dimensionality reduction. This is helpful in design of data collection systems and in health interventions to reduce the rate of diabetes. Easy to measure and report features can be used instead of more expensive variables.
2. The data set has a combination of categorial and numeric variables. The structure of each of these can be modeled separately K – Means (Numeric) and GMM (Categorial and Numeric) and collectively using K-Prototype. Unsupervised Machine Learning algorithms provide useful insight into grouping similar patients / people. This insight is helpful in the targeting of potential diabetics and diabetics based on their current features.
3. K-Prototype clustering is useful as an unsupervised ML technique in understanding the structure of the data.
4. KNN with SMOTE ENN and Logistic Regression models are the best predictive models for the dataset. KNN-SMOTE ENN is a useful technique in predicting a rare occurrence in a population.

4.3. Data Collection System Design

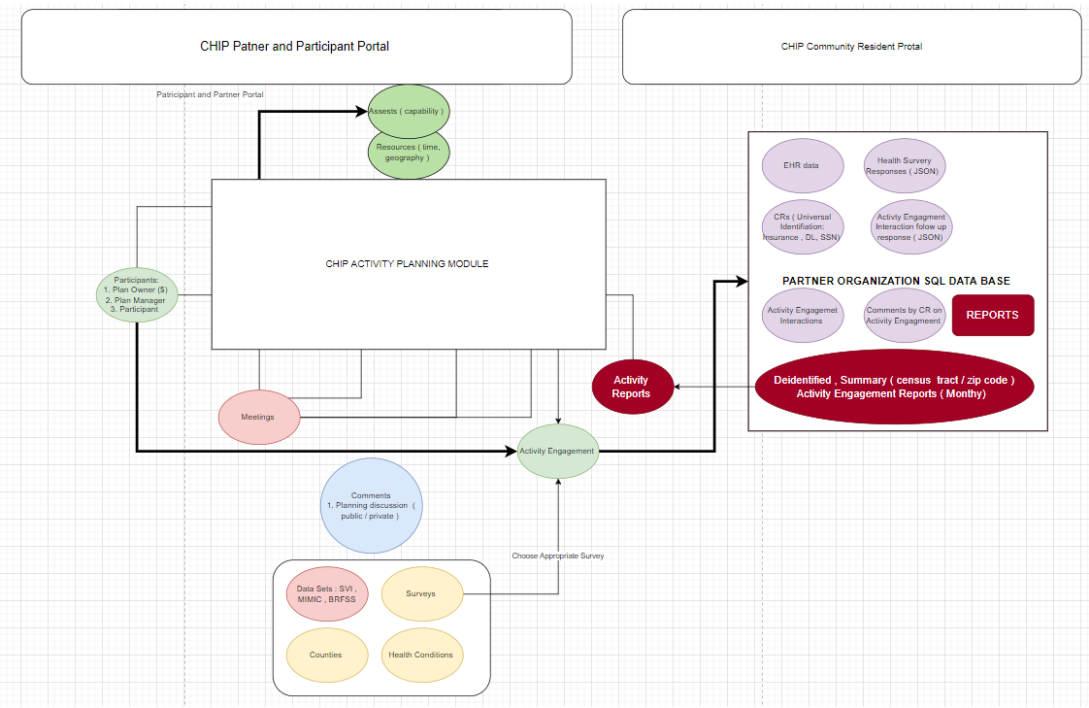


Figure 46: The Community Health Activity and Engagement Record Systems (CHEARS) Architecture

Figure 46 shows a high-level entity relational model for CHEARS™. The key distinguishing feature of this system is in the utilization of a microservices based architecture that separates out individual level data that is protected by health privacy laws from community level data such as health objective selection planning engagement and activity mapping data points. Further it accommodates publicly available datasets and other non-confidential health data sources that can be used to facilitate the CAIM development. The system can utilize standard validated surveys, successful activities, and well-established report structure for progress measurement. The data sets and survey tools can be made available via knowledge library to any health coalition that utilizes the EASI™ methodology and the CHEARS™ data model to facilitate data collection and progress measurement models.

4.4. Data Collection System Implementation as a Relational Model

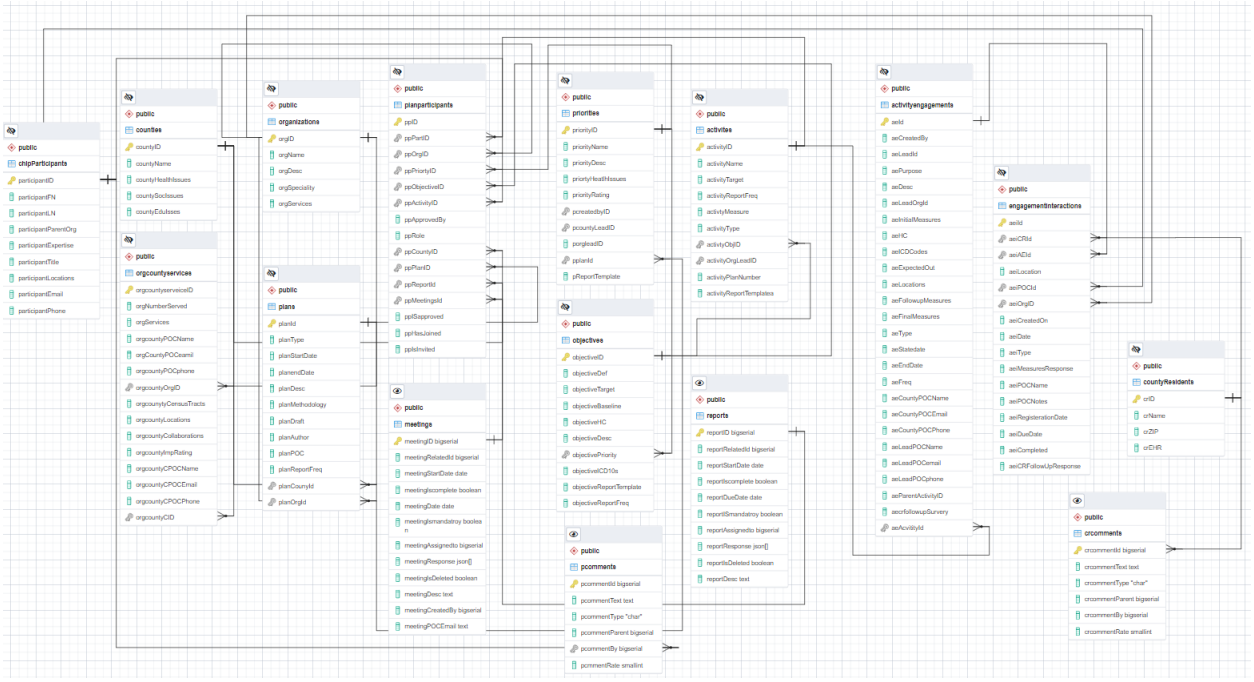


Figure 47: The Community Health Activity and Engagement Record Systems (CHEARS) Architecture

The EASI™ methodology applied in community health improvement planning revealed four essential stages off the planning and implementation process – Collaboration, Participation, Planning, and Implementation. The CHEARS™ data collection system, shown in figure 47, supports each of these process in community health planning for any target health objective.

As can be seen in the CHIP planning process flow chart community health improvement starts at the collaboration phase. This phase brings together willing participants who represent local organizations that are willing to allocate certain resources for the improvement of health objectives that are in the collective interest of the community under the guidance of the county Community Health Departments (CHDs). These participants engage in a series of meetings to identify the collective interest in specific health objectives. Diabetic hospitalizations are often a target health condition for community health interventions. In the Participation phase the health coalition comes

together to prioritize and selected a group of target health conditions that are in the collective best interests of the community. In this dissertation we have used diabetic hospitalization as an example to illustrate this process. Once health objectives are selected specific CAIM are discussed in the Planning phase oftentimes led by local subject matter experts. The CHEARS™ data model allows the collaborative to capture and share data related to the organizations, resources and expert opinions during the Collaboration, Participation and Planning phase, key conversations during the meeting ratings of different health objectives and activities during the participation phase and finally in the planning phase participants come together to determine specific measures for each activity and reporting requirements such that progress towards the target health objective can be measured on a periodic basis. The data collection during the

It is in the implementation phase where the benefits of investment in the CHEARS™ data model are fully realized. Typically, data is collected at various points in time by different organizations and this data cannot be shared due to health data privacy laws. By adopting a microservices based architecture organization specific data is not made available to the rest of the coalition whereas activity planning objective and priority data he's made available to the general community. The individual level health data at the organization can then be deidentified by generating a synthetic data set and utilized in future planning activities. In the EASI™ methodology the mimic data set is used as an illustrative example of such organizational data. Thus, the CHEARS™ system is a relational data model for a rules based system that facilitates evidence based community health improvement planning implementation and progress monitoring

CHAPTER FIVE DISCUSSSION, CONCLUSIONS AND FUTURE WORK

5.1 Key Contributions

Overall, we believe that the proposed work can provide much needed integration of individual methods in developing a critically needed methodology for data modelling and prediction. Its use with rural clinics will provide a rostrum for solving some of the existing problems with rural healthcare in addition to the advancement of science presented by the proposed methodology.

Finally, we can state that this transdisciplinary and transformative proposed contribution to data science and system design presents the following key contributions:

1. Data Efficiency:
    a. Relatively small sample set will suffice when compared to deep learning, unsupervised learning techniques
    b. The number of survey items can be reduced to include on those items that truly add predictive power to the resultant causal BN
2. Intervention Design
    a. Interventions can be designed based on validation of theorized causal structure.
    b. System workflow and requirements can be developed based on validated causal reasoning.
3. Application and Adoption in Practice
    a. This dissertation has resulted in the validation of the proposed methodology, EASI™ .
    b. EASI™ methodology was utilized to develop the CHEARS™ data model
    c. The methodology and flexible data architecture ensure data privacy and also facilitate evidence-based collaboration.

4. Scientific Contribution
    a. A modeling framework to generate and validate causal models for community-based health care interventions.
    b. A methodology to generate simulated / synthetic data based on localized datasets to validate effectiveness of targeted resource allocation for specific health conditions.

5.1.1 Contribution to Modeling

From a perspective of the discipline of modeling and simulation the easy methodology helps to leverage various simulation techniques such as GANs and modeling techniques including linear regression classification supervised unsupervised learning, in a way that is easy to implement the framework so that these techniques can be utilized in the improvement of community health. The CHEARS™ data architecture supports the capture of new data elements or reduced set of data elements based on acceptability of the resultant classification or prediction model. It allows the data model to evolve over time based on dimensionality reduction techniques such as Liner Discriminant Analysis (LDA) for supervised learning and PCA in unsupervised learning. PCA generates new constructs or principal components that result from linear transformation of the original data set features. These components are oriented to maximize the capture of variation in the original features using the fewest components. As expected, the prediction often improves when the principal components are also included in neural network based PCA. In context of adaptive dimensionality reduction flexible emergent data architecture proposed in the CHEARS™ model can retain the principal components from previous models and utilize these components in newer emergent models. (Migenda, Möller, & Schenck, 2021)

In healthcare, there is an abundance of unlabeled data and as a result self-taught learning models that can assign labels can help in classification of data. The emergent data architecture model can capture the new constructs and support development of self-taught learning models that can use both labeled and unstructured text data. (Raina, Battle, Lee, Packer, & Ng, 2007)

5.1.1.1 Contribution 1: Flexible Data Architecture

As a modeling methodology EASI™ advances, the discipline by incorporating seamlessly emergent data shapes thus offering a flexible approach to practitioners who seek to improve target outcomes overtime. From a scientific perspective, the EASI™ methodology demonstrates the utility of synthetic tabular health data in planning health care resource allocation.

The EASI™ provides a framework for model selection, feature reduction, model generation, prediction, and utility of the model in health care planning. The flexible data architecture ensures that only the data required to inform Models that are generated based on theory is captured. The CHEARS™ data model provides a lightweight, inter-organization data capture system to inform collaborative CAIMs.

5.1.1.2 Contribution 2: Adoption in Practice

The proposed EASI™ methodology is flexible and can applied based on the expertise of the data modeler and the performance desired. Health data can be diverse and include categorical, numeric, text, speech, and images. As such there is no restriction on the data types that might inform a community activity intervention model. Each organization is at liberty to capture the data it considers essential to track its services during a specific health intervention activity. The data modeler must determine appropriate techniques to apply based on the target outcome improvement desired. The EASI™ methodology imposes no restrictions on the techniques that can be applied.

It instead provides guidance on the steps that a modeler should follow to plan for emergent data shapes. In doing so it motivates the adoption of evidence based casual reasoning, continual review of model performance and intervention design by community health collaboratives.

5.1.2 Contribution to Simulation

Sharing of health data across organizations is a challenge due to data privacy concerns. There are several techniques to de-identify data. However, manual techniques are burdensome and unsustainable in practice. Furthermore, it is unlikely that a hospital, or county health department will accept the risk of HIPPA violations.

5.1.2.1 Contribution 1: Utilization of Synthetic Data for Intervention Performance Modeling

The EASI™ methodology informs a practical use case for the application of Generative Adversarial Networks (GANs) to generate simulated health data sets based on activity data captured by partners in the health coalition. The activity interactions result when a patient access services. Data to measure type, quantity and mode of service is recorded along with patient details. The visit or "interaction" record is typically recorded in the EHR system. The EASI™ methodology helps to identify specific data points that are needed to inform the inter-organizational Community Activity Intervention Model for a specific target health objective. The activity data and patient records across organizations pertinent to the intervention are used to simulate a synthetic data set. This synthetic data set can then be used to review the performance of the community health intervention.

5.1.2.2 Contribution 2: Utilization of Public Data Sets such as BRFSS and SVI in Intervention Planning

The EASI™ method leverages public data sets specifically, BRFSS and SVI dataset and uses GAN based synthetic data generation to simulate census tract level population data. This data set can be

used in the planning of CAIM and resource allocation for specific target health objectives. Simulation of localized population level data with a focus on specific data features will gives planners the ability to simulate best case and worst-case scenarios, educate and mobilize community leadership and citizens. Simulation of population data at the census tract level allows for the application of small area analysis techniques. Geospatial analysis of simulated data can help with logistics and positioning of allocated resources to maximize access.

5.1.3 EASI Methodology contribution to localized Diabetic Hospitalization Intervention Models

Diabetic hospitalization is dependent on several individual clinical factors which are well expounded in literature. Hospitalization is also dependent on local community factors such as access to care, insurance status, social determinants of health, general mental health of the population, demographic features and social capital of the individual in question. This makes it's challenging to allocate community resources to improve the overall health of the community. The easy methodology combines behavioral data with social vulnerability data and leverages hospitalization data at the local level to determine which features are likely to impact hospital admission and in the event of admission length of stay and mortality. Furthermore, the easy methodology helps planners identify target groups that are likely to be uninsured and thereby a burden on the hospital and the community if admitted for emergency care. The methodology thus aligns with the philosophy of community health intervention planning that seeks to improve the health of socially vulnerable groups who if admitted frequently in the hospital increase overall costs for the entire community. In the context of diabetic hospitalization, the easy methodology provides an evidence based structured approach for targeted allocation of resources to specific vulnerable groups targeted activities by community partners to minimize progression towards

diabetic hospitalizations for those target individuals. This approach if implemented can transform the monitoring of resource allocation while aligning with existing mode of community health improvement planning and improving a community is ability to measure how effectively are deploying their resources.

5.2 Limitations

This dissertation intentionally simplifies the extremely complex approaches to health care intervention modeling and data analysis. It is important to note that this work contributes to the methodology of community health improvement planning. This dissertation outlines how a practitioner who is planning to allocate resources at the community level can leverage low cost publicly available data to improve intervention design and data collection. However, this dissertation does not provide validated causal models or regression models all classification models that might be readily utilized in intervention design for a specific county or community. Intentionally, this work illustrates and approach and does not focus on validation in a real-world setting.

In this work, we have recoded several variables in overly simplified manner. The data analysis and the exploration of public datasets provide an intuition on design for data collection system however this work is not a system design project in and of itself therefore the high level architecture described in this dissertation attempts to address some of the key concerns in community health data collection namely maintaining privacy while acquiring data on a multitude of features that might come from different providers. To design and implement able system this dissertation provides an initial high level entity model which then needs to be reviewed with community health planners to identify specific data fields and a process model by which those data points can be acquired in a real-life setting.

This dissertation leverage is the concept of synthetic data generation using GANs as a means of generating population level data given marginal distributions from the social vulnerability index data at the census track level and the distribution joint distribution of features from the BRFSS data however the synthesized data it's not reflective of the true population characteristics because we have taken a simplistic approach to illustrate how synthetic data might be leveraged in generating population level datasets for use in healthcare planning . It is not the intention of this work to generate a realistic data set and optimize for a particular location.

5.2.1 Limitation 1: Methodology

The dataset along with the features selected represent an academic exercise to illustrate a part of the EASI™ methodology. The data set and the features selected may not correspond with data points captured to model a specific community activity intervention model. Another significant limitation of the data analysis conducted in this dissertation is that longitudinal data was not utilized. To overcome this limitation a rigorous data analysis with a 5-year longitudinal data set can be used to show that the factor loadings and very structure of the causal model may change over time. This will require careful consideration while making model comparisons based on the ultimate utility of the resultant models.

5.2.2 Limitation 2: Synthetic Data Generation

The algorithm described in this dissertation is yet to be validated. The implementation of the Coupla GAN to generate the synthetic dataset and graph data models to ingest the data and build machine learning pipelines to run selected machine learning models have not been attempted in this dissertation. This is a subject of an ongoing research effort.

5.2.3 Limitation 3: Data Modeling

All the features that might be of significance in prediction may not have been considered. The data modeling of diabetic patients from the MIMIC dataset is limited. A number of co-morbidities have not been considered; these can have a significant impact on outcome. Furthermore, longitudinal data analysis has not been conducted. This is in part because it is hard to find emergent datasets due to the limitations in current EHR and other healthcare data collection systems.

5.2.4 Limitation 4: Interpretability and Intervention Design

Causal models for diabetic prediction were not developed as part of this work. This is because there exist well researched and accepted casual models for most chronic disease health conditions such as diabetes. Predictive supervised models that utilize health data exist for most chronic diseases in literature. These models were not utilized since a simplified data set was used to illustrate the methodology. The objective of this dissertation is limited to illustration of how the EASI™ methodology may be utilized by collaborative community health coalitions.

5.3 Comparison with most closely related work

5.3.1 Comparison with Health as a System Model

In this paper the researchers propose an optimization method based on two objective functions productive efficiency (PE) and quality effectiveness (QE). These functions as simultaneously utilized in the computation of system performance. However, this one step method to evaluate goal attainment (GA) doesn't consider the phenomenon of emergence and resultant changes in the ecology within which the data envelopment analysis (DEA) was initially performed. The EASI™ methodology make two significant additions to the Health as a System Model. It accommodates changes to the ecology over time and reframes goal attainment (GA) as an emergent, iterative process. Each cycle of analysis requires a revised DEA and this can result in the identification of

new features hitherto considered insignificant, alter the relative weights and the resultant frontier scores for effectiveness and efficiency and necessitate the revision of the feature set considered. A second addition to the Health as a System Model is the utilization of public datasets such as BRFSS and SVI to develop a baseline model that the community intervention can use to define expected goal attainment in the development of community activity intervention models targets.

5.3.2 Comparison with GAIN: Missing Data Imputation using Generative Adversarial Networks

Missing data in health care data set can in some cases be a significant limitation in planning resource allocation. There are several reasons for missing data such as non-collection, incomplete records, insufficient data in dataset that can be complimented with auxiliary dataset. There are many techniques in literature for imputation of missing data, such as missing completely at random (MCAR), observed data missing at random (MAR) and data missing not at random (MNAR). Literature has several discriminative and generative methods to impute missing data. The GAIN algorithm advances field of MCAR data imputation. It uses generative adversarial networks in which the generator is trained to maximize the discriminators misclassification rate , while the discriminator attempts to minimize classification loss. An interesting aspect of this algorithm is that the discriminator is provided with a "hint matix" based on true population distribution that which in turn forces the generator to generate samples according to the true the underlying data distribution.

The community health data synthesis GAN algorithm proposed this in this dissertation extends the "hint matrix" concept outlined in the GAIN algorithm and uses the conditional CTGAN.(L. Xu, Skoularidou, Cuesta-Infante, & Veeramachaneni, 2019) Instead of looking at the problem as an imputation problem the algorithm considers the scenario wherein two separate data set with different features of interest are to be combined. The BRFSS data set with individual responses is

the parent data set with true population distribution over a large geographic area i.e a national survey. However, community health interventions are localized activities. The SVI dataset provides census tract level totals for certain overlapping or closely related features while incorporating variables that are not included in the BRFSS dataset. Both data sets are drawn from the same global population however, they have entirely different structures. The health data synthesis GAN algorithm retains the copula of the population distribution by training the generator on the BRFSS data set and forces localized variations by utilizing the marginal distribution based on the census tract level totals from the SVI dataset. Instead of the "hint matrix" corresponding the true population distribution as in the GAIN algorithm we provide the CTGAN with a marginal distribution matrix from the localized SVI dataset to induce localized variations to the population level distribution in the BRFSS dataset. This is anticipated to generate localized population level synthetic datasets variables pertinent to Community Activity Intervention Model for specific health objectives of interest to the health coalition.

5.4 Direction of Future Work

The future objective of the CHIP is to ensure intervention strategies to improve health outcomes have a positive cost to benefit ratio. Assuming variables follow the normal distribution the central limit theorem or the bootstrapping methods can be used to estimate costs of intervention vs readmission. The EASI methodology thus ensures revision to the causal structure as a result of emergence and utilization of the casual structure in making data driven tradeoffs to improve outcomes while cost of interventions are optimized. (Nixon, Wonderling, & Grieve, 2010).

More importantly, synthetic data generation algorithm and code to modify the marginal distribution using Coupla GANs has been developed. These methods can be used to generate high fidelity multidimensional representational observational health data sets which can then be used

to validate theoretically (literature/expert opinion) informed causal model. This model is then used to define and develop an efficient data collection system. As an immediate extension to this dissertation a representational sample generated using GANs is being developed along with the implementation of the data architecture in appropriate database structure. The synthetic data generate using the GAN will be imported into the data structure using standard data engineering Extract, Transform and Load (ETL) procedures. The machine learning models described in this dissertation along with other appropriate models will be used to evaluate the predictive potential of the BRFSS and SVI features that impact outcomes for the specific target health condition, in this case diabetic hospitalizations. MIMIC-iii hospitalization data sets will be utilized to evaluate the impact of community level interventions on potential hospitalization reduction.

This dissertation provides a methodology and illustrates the same with the example of diabetic hospitalizations. It can be extended to any target health condition and be adapted based on Community Specific Activity Intervention Models to improve overall health of the community by targeted allocation of resources to mitigate specific, localized vulnerabilities in the community. When the CHEARS™ for data collection is designed and implemented as described in this dissertation using the EASI™ methodology it has the potential to facilitate the adoption of evidence-based practices by community health departments and CHIP coalitions without the need of extensive data engineering, modeling and analysis expertise. It can strengthen the heath coalition by facilitation of secure data sharing with no risk to private health data.

Finally, EASI™ methodology can be extended to other domains where in new data shapes emerge over time.

REFERENCES

Anand, R. S., Stey, P., Jain, S., Biron, D. R., Bhatt, H., Monteiro, K., . . . Chen, E. S. (2018). Predicting Mortality in Diabetic ICU Patients Using Machine Learning and Severity Indices. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, 2017*, 310-319. Retrieved from https://pubmed.ncbi.nlm.nih.gov/29888089

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5961793/

Anderson, R. D., & Vastag, G. (2004). Causal modeling alternatives in operations research: Overview and application. *European Journal of Operational Research, 156*(1), 92-109. doi:https://doi.org/10.1016/S0377-2217(02)00904-9

Arling, G., Blaser, M., Cailas, M. D., Canar, J. R., Cooper, B., Flax-Hatch, J., . . . Sambanis, A. (2021). A Data Driven Approach for Prioritizing COVID-19 Vaccinations in the Midwestern United States. *Online journal of public health informatics, 13*(1), e5-e5. doi:10.5210/ojphi.v13i1.11621

Baptista, D. R., Wiens, A., Pontarolo, R., Regis, L., Reis, W. C. T., & Correr, C. J. (2016). The chronic care model for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome, 8*(1), 7. doi:10.1186/s13098-015-0119-z

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol, 51*(6), 1173-1182. doi:10.1037//0022-3514.51.6.1173

Beretta, S., Castelli, M., Gonçalves, I., Henriques, R., & Ramazzotti, D. (2018). Learning the Structure of Bayesian Networks: A Quantitative Assessment of the Effect of Different Algorithmic Schemes. *Complexity, 2018*, 1591878. doi:10.1155/2018/1591878

Blodgett, J. G., & Anderson, R. D. (2000). A Bayesian Network Model of the Consumer Complaint Process. *Journal of Service Research, 2*(4), 321-338. doi:10.1177/109467050024002

Braithwaite, J. (2018). Changing how we think about healthcare improvement. *BMJ, 361*, k2014. doi:10.1136/bmj.k2014

Britannica, T. E. o. E. (2017). https://www.britannica.com/science/emergence-science. In.

CDC, C. f. D. C. a. P. (2015). *Centers for Disease Control. Access the survey data and documentation for any BRFSS survey year.*

. Retrieved from: https://www.cdc.gov/brfss/annual_data/annual_data.htm

Chan, A. P. C., Wong, F. K. W., Hon, C. K. H., & Choi, T. N. Y. (2018). A Bayesian Network Model for Reducing Accident Rates of Electrical and Mechanical (E&M) Work. *International journal of environmental research and public health, 15*(11), 2496. doi:10.3390/ijerph15112496

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering, 5*(6), 493-497. doi:10.1038/s41551-021-00751-8

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W., & Sun, J. (2017). *Generating Multi-label Discrete Patient Records using Generative Adversarial Networks.* Paper presented at the MLHC.

Davis, J., Lim, E., Taira, D. A., & Chen, J. (2019). Healthcare network analysis of patients with diabetes and their physicians. *The American journal of managed care, 25*(7), e192-e197. Retrieved from https://pubmed.ncbi.nlm.nih.gov/31318509

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6999614/
Dinh, A., Miertschin, S., Young, A., & Mohanty, S. D. (2019). A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Medical Informatics and Decision Making, 19*(1), 211. doi:10.1186/s12911-019-0918-5
Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. arXiv:1706.02633. Retrieved from https://ui.adsabs.harvard.edu/abs/2017arXiv170602633E
FDOH, F. D. o. H. (2021). Retrieved from http://www.flhealthcharts.com/charts/default.aspx
Fernandez, M. E., Ruiter, R. A. C., Markham, C. M., & Kok, G. (2019). Intervention Mapping: Theory- and Evidence-Based Health Promotion Program Planning: Perspective and Examples. *Frontiers in public health, 7*, 209-209. doi:10.3389/fpubh.2019.00209
Flanagan, B. E., Hallisey, E. J., Adams, E., & Lavery, A. (2018). Measuring Community Vulnerability to Natural and Anthropogenic Hazards: The Centers for Disease Control and Prevention's Social Vulnerability Index. *Journal of environmental health, 80*(10), 34-36. Retrieved from https://pubmed.ncbi.nlm.nih.gov/32327766

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7179070/
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative Adversarial Networks. *ArXiv, abs/1406.2661*.
Grant, S. W., Collins, G. S., & Nashef, S. A. M. (2018). Statistical Primer: developing and validating a risk prediction model†. *European Journal of Cardio-Thoracic Surgery, 54*(2), 203-208. doi:10.1093/ejcts/ezy180
Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., & Pagano, E. (2011). Regression models for analyzing costs and their determinants in health care: an introductory review. *International Journal for Quality in Health Care, 23*(3), 331-341. doi:10.1093/intqhc/mzr010
Gupta, S., & Kim, H. W. (2008). Linking structural equation modeling to Bayesian networks: Decision support for customer retention in virtual communities. *European Journal of Operational Research, 190*(3), 818-833. doi:https://doi.org/10.1016/j.ejor.2007.05.054
Hassanzadeh, O., Bhattacharjya, D., Feblowitz, M., Srinivas, K., Perrone, M., Sohrabi, S., & Katz, M. (2019). *Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts*.
Hayashi, T., Sakaji, H., Matsushima, H., Fukami, Y., Shimizu, T., & Ohsawa, Y. (2021). Data Combination for Problem-Solving: A Case of an Open Data Exchange Platform. *The Review of Socionetwork Strategies, 15*(2), 521-534. doi:10.1007/s12626-021-00083-8
Hays, R. D., Revicki, D., & Coyne, K. S. (2005). Application of structural equation modeling to health outcomes research. *Eval Health Prof, 28*(3), 295-309. doi:10.1177/0163278705278277
Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning, 20*(3), 197-243. doi:10.1007/BF00994016
Hewitt, A. M., & Dykstra, D. Assessing Population Health: Community Health Needs Assessments. In A. M. Hewitt, J. L. Mascari, & S. L. Wagner (Eds.), (pp. 39-54). New York: Springer Publishing Company.
Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., . . . Haire-Joshu, D. (2020). Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care, 44*(1), 258-279. doi:10.2337/dci20-0053

Holden, R. J., McDougald Scott, A. M., Hoonakker, P. L. T., Hundt, A. S., & Carayon, P. (2015). Data collection challenges in community settings: insights from two field studies of patients with chronic disease. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 24*(5), 1043-1055. doi:10.1007/s11136-014-0780-y

Holmes, C. S., Chen, R., Streisand, R., Marschall, D. E., Souter, S., Swift, E. E., & Peterson, C. C. (2005). Predictors of Youth Diabetes Care Behaviors and Metabolic Control: A Structural Equation Modeling Approach. *Journal of Pediatric Psychology, 31*(8), 770-784. doi:10.1093/jpepsy/jsj083

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Jia, H., & Lubetkin, E. I. (2009). Time trends and seasonal patterns of health-related quality of life among U.S. adults. *Public Health Rep, 124*(5), 692-701. doi:10.1177/003335490912400511

King, R. J., Garrett, N., Kriseman, J., Crum, M., Rafalski, E. M., Sweat, D., . . . Cutts, T. (2016). A Community Health Record: Improving Health Through Multisector Collaboration, Information Sharing, and Technology. *Prev Chronic Dis, 13*, E122. doi:10.5888/pcd13.160101

Lara-Garcia, O. E., Retamales, V. A., Suarez, O. M., Parajuli, P., Hingle, S., & Robinson, R. (2020). Application of Social Vulnerability Index to Identify High- risk Population of Contracting COVID-19 Infection: a state-level study. *medRxiv,* 2020.2008.2003.20166983. doi:10.1101/2020.08.03.20166983

Layde, P. M., Christiansen, A. L., Peterson, D. J., Guse, C. E., Maurana, C. A., & Brandenburg, T. (2012). A model to translate evidence-based interventions into community practice. *American journal of public health, 102*(4), 617-624. doi:10.2105/AJPH.2011.300468

Lichtenstein, B. (2015). The SAGE Handbook of Action Research. In H. Badbury (Ed.), (Third Edition ed., pp. 446-452). doi:10.4135/9781473921290

Linder, S., Marko, D., Tian, Y., & Wisniewski, T. (2018). A Population-Based Approach to Mapping Vulnerability to Diabetes. *International journal of environmental research and public health, 15*(10), 2167. doi:10.3390/ijerph15102167

McLeroy, K. R., Norton, B. L., Kegler, M. C., Burdine, J. N., & Sumaya, C. V. (2003). Community-based interventions. *American journal of public health, 93*(4), 529-533. doi:10.2105/ajph.93.4.529

Mehrotra, A., Wang, M. C., Lave, J. R., Adams, J. L., & McGlynn, E. A. (2008). Retail clinics, primary care physicians, and emergency departments: a comparison of patients' visits. *Health affairs (Project Hope), 27*(5), 1272-1282. doi:10.1377/hlthaff.27.5.1272

Migenda, N., Möller, R., & Schenck, W. (2021). Adaptive dimensionality reduction for neural network-based online principal component analysis. *PloS one, 16*(3), e0248896. doi:10.1371/journal.pone.0248896

Mossio, M., Bich, L., & Moreno, A. (2013). Emergence, Closure and Inter-level Causation in Biological Systems. *Erkenntnis*. doi:10.1007/s10670-013-9507-7

Neelon, B., Mutiso, F., Mueller, N. T., Pearce, J. L., & Benjamin-Neelon, S. E. (2021). Spatial and temporal trends in social vulnerability and COVID-19 incidence and death rates in the United States. *PloS one, 16*(3), e0248702. doi:10.1371/journal.pone.0248702

Nixon, R., Wonderling, D., & Grieve, R. (2010). Non-parametric methods for cost-effectiveness analysis: the Central Limit theorem and the bootstrap compared. *Health economics, 19*, 316-333. doi:10.1002/hec.1477

O'Connor, P. J., Bodkin, N. L., Fradkin, J., Glasgow, R. E., Greenfield, S., Gregg, E., . . . Wysham, C. H. (2011). Diabetes Performance Measures: Current Status and Future Directions. *Diabetes Care, 34*(7), 1651-1659. doi:10.2337/dc11-0735

Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., & Omar, R. Z. (2015). How to develop a more accurate risk prediction model when there are few events. *BMJ, 351*, h3868. doi:10.1136/bmj.h3868

Philis-Tsimikas, A., & Gallo, L. C. (2014). Implementing community-based diabetes programs: the scripps whittier diabetes institute experience. *Current diabetes reports, 14*(2), 462-462. doi:10.1007/s11892-013-0462-0

Pollack, A. H., Simon, T. D., Snyder, J., & Pratt, W. (2019). Creating synthetic patient data to support the design and evaluation of novel health information technology. *Journal of Biomedical Informatics, 95*, 103201. doi:https://doi.org/10.1016/j.jbi.2019.103201

Powers, D. M., Bowen, D. J., Arao, R. F., Vredevoogd, M., Russo, J., Grover, T., & Unützer, J. (2020). Rural clinics implementing collaborative care for low-income patients can achieve comparable or better depression outcomes. *Fam Syst Health, 38*(3), 242-254. doi:10.1037/fsh0000522

Raghu, V. K., Poon, A., & Benos, P. V. (2018). Evaluation of Causal Structure Learning Methods on Mixed Data Types. *Proceedings of machine learning research, 92*, 48-65. Retrieved from https://pubmed.ncbi.nlm.nih.gov/31080946

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6510516/

Rahaman, K. S., Majdzadeh, R., Holakouie Naieni, K., & Raza, O. (2017). Knowledge, Attitude and Practices (KAP) Regarding Chronic Complications of Diabetes among Patients with Type 2 Diabetes in Dhaka. *International journal of endocrinology and metabolism, 15*(3), e12555-e12555. doi:10.5812/ijem.12555

Raina, R., Battle, A., Lee, H., Packer, B., & Ng, A. Y. (2007). *Self-taught learning: transfer learning from unlabeled data*. Paper presented at the Proceedings of the 24th international conference on Machine learning, Corvalis, Oregon, USA. https://doi.org/10.1145/1273496.1273592

Ravaut, M., Sadeghi, H., Leung, K. K., Volkovs, M., Kornas, K., Harish, V., . . . Rosella, L. (2021). Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. *npj Digital Medicine, 4*(1), 24-24. doi:10.1038/s41746-021-00394-8

Regmi, M. R., Tandan, N., Parajuli, P., Bhattarai, M., Maini, R., Kulkarni, A., & Robinson, R. (2021). Social Vulnerability Indices as a Risk Factor for Heart Failure Readmissions. *Clinical medicine & research, 19*(3), 116-122. doi:10.3121/cmr.2021.1603

Roversi, C., Tavazzi, E., Vettoretti, M., & Camillo, B. D. (2021, 27-30 July 2021). *A Dynamic Bayesian Network model for simulating the progression to diabetes onset in the ageing population.* Paper presented at the 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI).

Rueger, A. (2000). Physical Emergence, Diachronic And Synchronic. *Synthese, 124*(3), 297-322. doi:10.1023/A:1005249907425

Sacerdote, C., Ricceri, F., Rolandsson, O., Baldi, I., Chirlaque, M.-D., Feskens, E., . . . Wareham, N. (2012). Lower educational level is a predictor of incident type 2 diabetes in European

countries: The EPIC-InterAct study. *International Journal of Epidemiology, 41*(4), 1162-1173. doi:10.1093/ije/dys091

Sartenaer, O. (2015). Synchronic vs. diachronic emergence: a reappraisal. *European Journal for Philosophy of Science, 5*(1), 31-54. doi:10.1007/s13194-014-0097-2

Scutari, M. (2017). Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimized Implementations in the bnlearn R Package. *Journal of Statistical Software, 77*. doi:10.18637/jss.v077.i02

Seligman, B., Ferranna, M., & Bloom, D. E. (2021). Social determinants of mortality from COVID-19: A simulation study using NHANES. *PLOS Medicine, 18*(1), e1003490. doi:10.1371/journal.pmed.1003490

Shigaki, C., Kruse, R. L., Mehr, D., Sheldon, K. M., Bin, G., Moore, C., & Lemaster, J. (2010). Motivation and diabetes self-management. *Chronic Illn, 6*(3), 202-214. doi:10.1177/1742395310375630

Spirtes, P., Glymour, C., & Scheines, R. (2020). Prediction and Search.

Stoto, M. A., Davis, M. V., & Atkins, A. (2019). Making Better Use of Population Health Data for Community Health Needs Assessments. *EGEMS (Washington, DC), 7*(1), 44-44. doi:10.5334/egems.305

Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., & Chaimontree, S. (2021). Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand. *Informatics in Medicine Unlocked, 26*, 100769. doi:https://doi.org/10.1016/j.imu.2021.100769

Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digital Medicine, 3*(1), 147. doi:10.1038/s41746-020-00353-9

Vaona, A., Del Zotti, F., Girotto, S., Marafetti, C., Rigon, G., & Marcon, A. (2017). Data collection of patients with diabetes in family medicine: a study in north-eastern Italy. *BMC health services research, 17*(1), 565-565. doi:10.1186/s12913-017-2508-5

Wan, T. T. H. (2002). *Evidence-based health care management : multivariate modeling approaches*. Boston: Kluwer Academic Publishers.

Wan, T. T. H., Matthews, S., Luh, H., Zeng, Y., Wang, Z., & Yang, L. (2022). A Proposed Multi-Criteria Optimization Approach to Enhance Clinical Outcomes Evaluation for Diabetes Care: A Commentary. *Health Serv Res Manag Epidemiol, 9*, 23333928221089125. doi:10.1177/23333928221089125

Wan, T. T. H., Terry, A., McKee, B., & Kattan, W. (2017). KMAP-O framework for care management research of patients with type 2 diabetes. *World journal of diabetes, 8*(4), 165-171. doi:10.4239/wjd.v8.i4.165

Weinick, R. M., Caglia, J. M., Friedman, E., & Flaherty, K. (2007). Measuring racial and ethnic health care disparities in Massachusetts. *Health Aff (Millwood), 26*(5), 1293-1302. doi:10.1377/hlthaff.26.5.1293

Wells, K. B., Staunton, A., Norris, K. C., Bluthenthal, R., Chung, B., Gelberg, L., . . . Wong, M. (2006). Building an academic-community partnered network for clinical services research: the Community Health Improvement Collaborative (CHIC). *Ethn Dis, 16*(1 Suppl 1), S3-17.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019). *Modeling Tabular data using Conditional GAN*. Paper presented at the NeurIPS.

Xu, X.-f., Sun, J., Nie, H.-t., Yuan, D.-k., & Tao, J.-h. (2016). Linking structural equation modeling with Bayesian network and its application to coastal phytoplankton dynamics in the Bohai Bay. *China Ocean Engineering, 30*(5), 733-748. doi:10.1007/s13344-016-0047-1

Yoon, J., Jordon, J., & Schaar, M. (2018). *RadialGAN: Leveraging multiple datasets to improve target-specific predictive models using Generative Adversarial Networks.* Paper presented at the International Conference on Machine Learning.

Zhang, T., Fu, X., Ma, S., Xiao, G., Wong, L., Kwoh, C. K., . . . Hung, T. (2012). Evaluating temporal factors in combined interventions of workforce shift and school closure for mitigating the spread of influenza. *PloS one, 7*(3), e32203-e32203. doi:10.1371/journal.pone.0032203