ESSAYS ON DIGITAL HEALTH AND PREVENTIVE CARE ANALYTICS

by

Karthik Srinivasan

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2019

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Karthik Srinivasan, titled Essays on Digital Health and Preventive Care Analytics and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Sudha Ram — Date: 04/23/2019

Susan Brown — Date: 04/23/2019

Wei Chen — Date: 04/23/2019

Faiz Currim — Date: 04/23/2019

Chiu-Hsieh (Paul) Hsu — Date: 04/23/2019

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: 04/23/2019

Dissertation Director:
Sudha Ram
Anheuser-Busch Endowed Chair of MIS, Entrepreneurship & Innovation
Director, INSITE Center for Business Intelligence & Analytics
Department of MIS, Eller College of Management, University of Arizona

ACKNOWLEDGEMENTS

I wish to express my utmost appreciation to my dissertation committee members, Drs. Sudha Ram, Susan Brown, Wei Chen, Faiz Currim, and Chiu-Hsieh (Paul) Hsu, for their encouragement, inspiration and guidance. I am especially grateful to my advisor, Dr. Sudha Ram, for her constant supervision and support throughout my tenure as a doctoral student at the University of Arizona. I consider myself most fortunate to have such an experienced and understanding mentor.

I am indebted to the University of Arizona and all my instructors for the irreplaceable learning experience I have had during the last five years. I am thankful to my colleagues in the INSITE lab, peers in the PhD program, my research collaborators, and my wife Danya for their critical feedback and constructive inputs. I owe my deepest gratitude to my family, and close friends for their patience and psychological support during the ups and downs of my journey as a PhD student. I gratefully acknowledge that the studies in this dissertation have been partly supported by the United States General Services Administration grant # GS-00-H-14-AA-C-0094.

DEDICATION

This dissertation is dedicated to my parents, and the human spirit of curiosity and selflessness.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Analytics is an integral component of health information systems (IS), showing promise in various areas such as disease risk modeling, clinical intelligence, pharmacovigilance, precision medicine, hospitalization process optimization, digital health, preventive care, etc. In my dissertation, I focus on analytics in two important application areas of health IS, namely digital health and preventive care. Digital health analytics focuses on enhancing individual wellbeing via continuous tracking of health indicators, while preventive care analytics is the science of extracting insights from electronic health records to assist clinical decision-making towards preventing illness or diseases. With rapid development in healthcare big data and IoT technologies, research in digital health and preventive care (DHPC) analytics is increasing in importance and complexity. Limited predictors, incomplete data, non-linear input-outcome relationships, the presence of multiple outcomes, and heterogeneity in effects are some of the key challenges in DHPC analytics. My dissertation consists of three essays that introduces a collection of novel quantitative methods to address these challenges. The first essay presents a new feature engineering method that uses network science to predict high-cost patients at the point of admission in hospitals with limited information. The second essay describes a novel method to analyze incomplete data containing block-wise missing patterns using a reduced modeling approach. The third essay leverages a wearable devices-based study and introduces three new quantitative methods to model the effects of sound level on an individual's physiological wellbeing in the workplace. The set of predictive and explanatory modeling methods proposed in these essays not only address important modeling challenges in DHPC analytics, but also more broadly contribute to business analytics, design science, and health IS research.

1. INTRODUCTION

The U.S. healthcare system is known to have the highest costs but lowest performance globally (cms.gov 2016; The Commonwealth Fund 2014). The potential of information technology and analytics to address challenges in healthcare is well-established in both medicine and information systems (IS) literature (Agarwal et al. 2010; Bates et al. 2014). IS researchers have been at the forefront of highlighting issues in healthcare (Kohli and Tan 2016), studying the emergence and effectiveness of new healthcare technologies (Agarwal and Dhar 2014), and providing novel solutions in health analytics (Ahsen et al. 2018). Health analytics is one of the fastest growing disciplines today with a wide range of applications including disease risk modeling, clinical intelligence, pharmacovigilance, precision medicine, hospitalization process optimization, digital health, preventive care, etc.

In this dissertation, I inquire into health analytics in the dual context of preventive care and digital health. Preventive care analytics is defined as the science of extracting insights from electronic health records to assist clinical decision-making towards preventing illness or diseases. Digital health analytics is defined as the analytics of digital information (e.g., internet, online social media, wearable/remote sensors, telemedicine, mobile health, etc.) for actionable intelligence towards improving the health and wellbeing of individuals. While preventive care is concerned towards data-centric disease prognosis, digital health is geared towards developing a deeper understanding of human wellbeing using digital technology and preventing diseases by supporting and encouraging healthy activities. Big data technology, Internet of Things (IoT), and favorable health policies in recent years have led to a steady increase in the creation and accumulation of rich and fine-grained data, offering boundless opportunities in digital health and preventive care (DHPC) analytics. The data generated by DHPC applications can be used to answer interesting research questions, but at the same time, they pose multiple challenges to analytics due to its

complexity. In my dissertation, I identify five such challenges pertaining to analyzing DHPC data: making predictions using limited predictors, handling incomplete data with block-wise missing patterns, representing curvilinear relationships, modeling multiple outcomes, and accounting for heterogeneity in effects across populations. My dissertation consists of three essays proposing a collection of new quantitative methods to address these challenges. Brief descriptions of the three essays are given below.

Essay 1: Predicting High Cost Patients at Point of Admission using Network Science

Data mining models for high-cost patient encounter prediction at the point-of-admission (HPEPP) in inpatient wards are scarce in literature. This is due to lack of availability of relevant features at such an early stage of treatment. In this study, we create a disease co-occurrence network (DCN) using a subset of Arizona's inpatient database. We explore this network for community formation and structural properties to create new input features for HPEPP models. Tree-based data mining models are trained using input feature sets that include these new network features, and distinct disease communities in the DCN are identified. We propose community membership and high-cost propensity scores as two network-based features for HPEPP modeling. We compare the performance of models with different input feature sets and find that the new features significantly improve the accuracy sensitivity and specificity of prediction models.

Essay 2: Analyzing incomplete data with block-wise missing patterns

Important research questions in IS often require combining inputs from multiple data sources. Such integration often leads to datasets with blocks of missing values. Traditional methods such as imputation are not effective for processing such datasets. We present a new method that utilizes block-wise missing patterns to build an ensemble of models that require minimum imputation of data. We use a two-phase process: (i) training several candidate models over overlapping subsets of original data that contain only populated values; and (ii) mapping test

instances at run time to corresponding candidate models to make predictions. We apply our method to the problem of predicting hourly demand of rental bikes in a city-wide bike sharing (CBS) program. We compare the predictive performance of our method with existing methods by simulating 25%, 50%, and 75% block-wise missing values in the dataset. We find that our method improves predictive performance for regression models over existing methods between 4% and 16%. In addition, it improves predictive performance for tree-based data mining models between 5% and 50%. We show that our method scales very well and is significantly faster than existing imputation methods. Important for researchers, even with 50% of feature values removed at random in a simulated version of the CBS dataset, we were able to determine effects consistent with those found in a model trained over complete data. We further verified the scalability and external validity of our method by testing it on a healthcare problem of determining factors related to patient visit costs, using more than 13 million observations with incomplete data. While existing imputation methods are inefficient for handling such data, we show that our method is useful to draw inferences without discarding valuable information.

Essay 3: Sound-wellbeing modeling using wearables

In today's world, people spend a significant proportion of their active hours in enclosed workplaces. Previous studies have established that workplace environment is closely tied to an individual's wellbeing. We conduct a field study using wearable devices on 231 federal office workers that measure the impact of indoor environment on individual wellbeing. This paper reports our results with respect to the effect of sound level amplitude. For modeling the dynamics of the sound-wellbeing relationship, we propose new methods for representing curvilinear effects, simultaneous modeling of multiple outcomes, and identifying factors contributing to heterogeneity in effects. We show that our methods have better model fit as well as predictive performance than existing methods for each of the three modeling problems. We observed that the relationship

between sound level and two physiological wellbeing measures (i.e., SDNN, normalized-HF) in a regular office workplace (with sound levels between 40 dBA and 80 dBA) is curvilinear and varies across individuals. We find that an individual's physiological wellbeing is optimal when sound level in the workplace is around 50 dBA. For sound amplitudes lower than 50 dBA, a 10 dBA increase in sound level is related to a 3.6% increase in physiological wellbeing; whereas for amplitudes above 50 dBA, a 10 dBA increase in sound level is related to decrease in physiological wellbeing by 1.3%. Age, body-mass-index, high blood pressure, anxiety, and computer use intensive work are person level factors contributing to heterogeneity in effects of sound level on physiological wellbeing across individuals. Workers with higher blood pressure are more negatively affected by increases in sound levels than others. Workers with computer intensive work are more negatively affected by extremities of low sound (i.e., quietude) or high sound (i.e., noise) than others. This study informs policies and practices that affect the health and wellbeing of office workers worldwide.

Table 1.1 summarizes the research question, analytics problem, areas of application and modeling methods/approaches of each study.

Table 1.1: Summary of dissertation essays

Essay	**Research question**	**Area**	**Modeling methods/approaches**
Essay 1: Predicting High Cost Patients at Point of Admission using Network Science	How can we effectively identify high cost patients at point of admission in hospitals?	Preventive care analytics	Network analysis, Data mining
Essay 2: Analyzing incomplete data with block-wise missing patterns	How can we analyze data and make predictions with incomplete data containing block-wise missing patterns?	Preventive care analytics, digital health analytics	Set theory, Clustering, Reduced modeling
Essay 3: Sound-wellbeing modeling using wearables	How is workplace sound level related to physiological wellbeing of office-workers?	Digital health analytics	Multilevel modeling, Hierarchical Bayes modeling

2. ESSAY 1: PREDICTING HIGH COST PATIENTS AT POINT OF ADMISSION USING NETWORK SCIENCE

2.1. Introduction

The 2015-2025 projections of National Health Expenditures Data (cms.gov 2016) estimate that healthcare spending in the US will increase from $3.197 trillion in 2015 to $5.631 trillion by 2025. The U.S. per capita healthcare spending, about $10,000, is the highest in the world. From a healthcare management perspective, of significant concern is that approximately five percent of patients account for about 50 percent of the total health care spending (Bates et al. 2014). One approach to reducing overall costs for healthcare systems is to identify such patients early and have case managers work with them to improve their care (Bates et al. 2014). The earlier high-cost and high-risk patients can be identified, the better. Even if a fraction of the total high-cost patients were successfully identified, reducing only 0.01% of the health costs due to pre-emptive patient care, the annual savings could extend to the range of hundreds of millions dollars (Hayes et al. 2016).

Our contributions include: (i) Addressing the Point of Admission (PoA) prediction problem which, to our knowledge, has not been done before; (ii) introducing a novel big data approach to feature engineering using network science; (iii) applying knowledge discovery using disease co-occurrence networks (DCN) including community formation and structural measures; and (iv) developing high-cost propensity scores for diseases. We introduce new features that reflect the direct relationship of the diseases present at the PoA with high-cost diseases. We demonstrate that our model incorporating network information has better overall accuracy sensitivity and specificity than models without network data.

2.2. Background

Retrospective analyses using electronic health records (EHR) examine the history of patient encounters to model a visit as high-cost or not (Bertsimas et al. 2008; Chechulin et al. 2014). This is useful for ex post facto identification of factors contributing to patient encounter costs. However, a different approach is needed for an early high-cost encounter detection system owing to limited data availability at the PoA. Typically, at the PoA, the inpatient ward has limited data, e.g., patient reports from the emergency department (or outpatient specialists) and some primary data such as basic patient information (Chechulin et al. 2014). Therefore, we require innovative methods that utilize information available at the PoA to make early predictions. We propose a big data approach of integrating network science with data mining models to effectively predict high cost patients at the PoA.

The phenomenon of the high-cost patient population (Zook and Moore 1980) is well-known in literature. The high-cost patient prediction issue is typically modeled as a classification problem using large public health datasets for training (Ash et al. 2001; Bertsimas et al. 2008; Chechulin et al. 2014; Moturu et al. 2007). Some of these studies have looked at aggregate health costs for patients and predicted high-cost patients in the population (Ash et al. 2001; Moturu et al. 2007). Other studies have broken up aggregated patient encounter data into time intervals, such as six or twelve months, to predict the likelihood of an individual becoming a high-cost patient in the future (Bertsimas et al. 2008; Chechulin et al. 2014; Sushmita et al. 2015). In terms of prediction techniques, studies have compared several modeling methods and determined that tree-based methods have the best predictive performance for patient health cost predictions (Moturu et al. 2008; Sushmita et al. 2015). Previous research has also identified past costs incurred by patients, patient demographics, and risk-indices such as diagnostic cost group (DCG) as important predictors of a patient becoming a high-cost candidate (Ash et al. 2001; Chechulin et al. 2014).

Research also shows that complex predictive models are not currently implemented in practice (Califf and Pencina 2013) due to a variety of reasons. Of these, two key reasons are: (i) fragmented efforts at developing predictive models with different variables make it difficult to establish a uniform or generally accepted list of important features, and (ii) models with a fixed set of features do not consider feature change history during patient treatment. None of the previous efforts develop a model suitable for use at the PoA. Our work addresses these gaps by developing a predictive model that can be trained at the PoA and be dynamically updated with new features as treatment progresses.

One of the contributions of our paper is integrating network science features into our predictive model. The field of network medicine has adapted network science to discover knowledge contained in the inter-relationships among medical entities (Barabasi and Frangos 2014). Naturally occurring networks, such as gene regulatory and protein-protein interaction networks, have been widely investigated. Phenotypes, i.e., observable characteristics, have also been used in network analysis. A disease co-occurrence network (DCN), also known as phenotype-disease network, was first conceptualized by Hidalgo et al. (2009). Exploratory studies using DCN with focus on specific diseases such as chronic obstructive pulmonary disease (Divo et al. 2015), diabetes (Klimek et al. 2015), and hypertension (Liu et al. 2016) have also been conducted. Visualization and ordered-listing (i.e., top-n analysis) are common methods for knowledge discovery in these studies. While network science for studying disease co-occurrence has been an active area of research, a gap in past work that we address is considering the results of network analysis as an input into classification algorithms. One of the contributions of our paper is using features generated through a network science analysis into our predictive model.

2.3. Methods

2.3.1. Data

In this study, we used the State Inpatient Discharge (SID) public dataset of Arizona to generate the DCN (Bureau of Public Health Statistics 2017). The dataset contains patients who are transferred from outpatient and emergency departments for treatment. Our Arizona SID dataset covers the period from 2012 through 2014, with 2,269,512 patient visit records and 10,809 distinct ICD9 diagnosis codes. Each patient encounter can have up to 25 diagnoses codes, i.e., 25 co-occurring diseases. Figure 2.1 is a linear quantile plot of the unadjusted cost of encounters. The X-axis shows the range of quantiles from 0.5 to 1.0, and the Y-axis show the corresponding value of unadjusted cost at a given quantile. A sharp increase in the gradient of quantiles can be seen at the 95th percentile of unadjusted cost of encounter, which illustrates the sharp rise in encounter cost at the 95th percentile, consistent with previous literature (Bates et al. 2014).

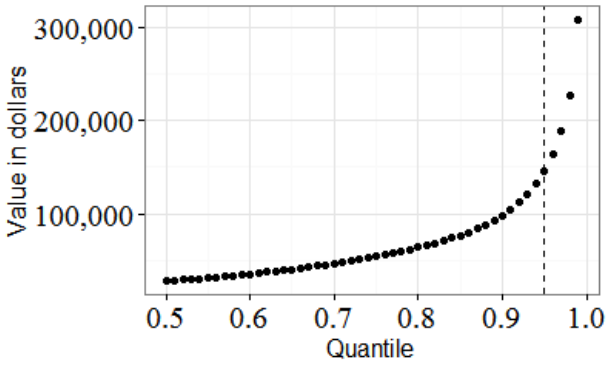


Figure 2.1: A linear quantile plot of the unadjusted cost of patient encounters

At the PoA, features available for training were age-group, race, sex and diagnoses reported until the PoA (i.e., PoA diagnoses). There were over 10,000 unique diagnoses codes in the dataset; hence, the PoA diagnoses fields could not be directly used as categorical features. In the past, group codes such as clinical classification software (CCS), were used instead of diagnoses codes directly as features (Ash et al. 2001; Bertsimas et al. 2008; Chechulin et al. 2014; Moturu et al. 2007). Therefore, we included the CCS multilevel group code of primary diagnosis as an additional

baseline feature. Table 2.1 shows the descriptive statistics of existing features in the dataset. In order to account for disease co-occurrence effects, comorbidity indices like cumulative risk scores are included as features in predictive models for mortality. One could argue that these measures may also be used for predicting high cost patients at PoA. Twelve Elixhauser comorbidity indicator variables (such as metastatic cancer, paralysis, renal failure, etc.) (Elixhauser et al. 1998) and the Charlson Comorbidity Index (CCI) (Charlson et al. 1987) were therefore included in the input feature set of a separate model (i.e., Baseline + CI, further described in later sub-section) to benchmark the performance of the proposed features in this study. The comorbidity variables are binary categorical features, each of which has a less than 15% positive parity (i.e., "is present" in fewer than 15% of the records) in the dataset. To examine the utility of the network features introduced in this study, we trained and tested predictive models on the Arizona SID dataset for the year 2015 (739,798 encounters). Since the disease co-occurrence network was developed using data from 2012-14, these years were excluded.

Table 2.1: Descriptive statistics of input features

Feature	**Statistics**					
Age-groups	*0-9 years*	*10-19 years*	*20-29 years*	*30-39 years*	*40-49 years*	
	14.30	3.67	11.05	9.97	8.06	
	50-59 years	*60-69 years*	*70-79 years*	*80-89 years*	*90+ years*	
	11.83	14.69	14.17	9.60	2.65	
Race	White	African American	Native American	Asian	Others	
	86.55	5.07	4.11	2.40	1.76	
Sex	Male			Female		
	43.98			56.02		
CCS groups	C1	C2	C3	C4	C5	C6
	20.25	16.10	8.68	7.39	7.23	4.91
	C7	C8	C9	C10	C11	Others

	2.81	2.75	2.38	1.59	1.23	19.74
CCS groups codes lookup: C1 - Nutritional/metabolic/immunity, C2 - Circulatory system, C3 - Nervous system and sense organs, C4 - Digestive system, C5 - Respiratory system, C6 - Musculoskeletal, C7 - Injury and Poisoning, C8 - Liveborn, C9 - Blood and blood-forming organs, C10 - Genitourinary, C11 - Neoplasms.						

2.3.2. Disease co-occurrence networks (DCN)

DCNs are constructed based on repeated evidence of two or more diseases. The nodes of these networks are ICD9/10 diagnosis codes and the edges represent co-occurrence relationship between pair-wise disease diagnoses. Figure 2.2 shows a disease co-occurrence network for a hypothetical example of three related conditions—diabetes, hypertension and obesity. The pairwise co-occurrence coefficients between the three are derived using number of co-occurrences and prevalence of each disease. Even though there is evidence of co-occurrence between each disease pair in this network, the strength of the co-occurrence relationship is not straightforward, and therefore implicit. Multiple measures have been proposed in the past to capture the co-occurrence between disease pairs as edges weights as shown in Table 2.2.

Table 2.2: Edge weights for disease co-occurrence networks

No	Formula	Study	Application
1	$RR_{xy} = NC_{xy}/P_xP_y$	Hidalgo et al., 2009 (Hidalgo et al. 2009)	Propose the DCN
2	$\phi_{xy} = \frac{NC_{xy} - P_xP_y}{P_xP_y(N-P_x)(N-P_y)}$	Hidalgo et al., 2009 (Hidalgo et al. 2009)	Propose the DCN
3	$SC_{xy} = C_{xy}/(P_x + P_y)$	Steinhaeuser and Chawla 2009 (Steinhaeuser and Chawla 2009)	Compute patient similarity
4	$\phi'_{xy} = \ln_2\left(\frac{C_{xy}+1}{\frac{P_xP_y}{N}+1}\right)$	Roque et al. 2011 (Roque et al. 2011)	Stratify patient cohorts
5	$RR_{xy} = NC_{xy}/P_xP_y$	Klimek et al. 2015 (Klimek et al. 2015)	Identify diabetes comorbidity risks
6	C_{xy}	Liu et al. 2016 (Liu et al. 2016)	Comorbidities analysis of hypertension
N is the size of the total number of observations in the dataset, C_{xy} is the co-occurrence of disease x and y across patient encounters and $\{P_x, P_y\}$ are prevalence of diseases $\{x, y\}$.			

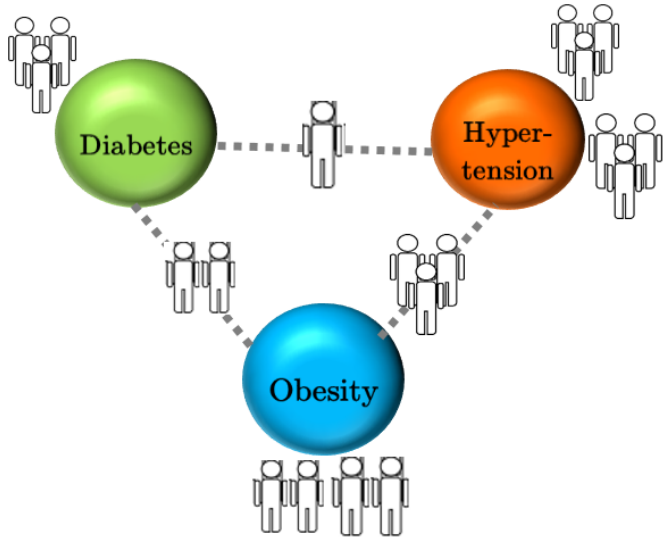


Figure 2.2: An illustrative representation of a three-node disease co-occurrence network

Existing approaches to assign DCN edge weights have limitations such as bias towards rare diseases and intractable ranges (Roque et al. 2011). We propose a new measure for edge weights called co-occurrence coefficient (CC) that is suitable for analyzing disease co-occurrence networks as follows:

$$CC_{xy} = \frac{\sqrt{2}C_{xy}}{\sqrt{P_x^2 + P_y^2}} \tag{2.1}$$

where CC_{xy} is the co-occurrence of disease x and y across patient encounters, and P_x and P_y are prevalence of diseases x and y respectively.

CC_{xy} is symmetric, reflexive and constrained to the range [0,1] and therefore, it is semi-metric (Shi et al. 2012) and does not suffer the limitations of previously proposed measures. A two-sided t-test is used to retain significant edges in the network with the following test statistic:

$$t_{n-1} = \sqrt{\frac{(n-2)CC_{xy}^2}{(1-CC_{xy}^2)}}, n = \max(P_x, P_y) \tag{2.2}$$

The resultant DCN for Arizona SID dataset had 2025 diagnoses as nodes, 38,812 pair-wise co-occurrence relationships as edges and a network density of 0.019. The distribution of different edge weight measures for the DCN created using the AZ SID data is given in Table 2.3. We observed in the DCN that RR_{xy} and SC_{xy} have bias towards pairs of rare and highly prevalent diseases, ϕ_{xy} and ϕ'_{xy} have bias towards pairs of highly prevalent diseases, and measures other than CC_{xy} and SC_{xy} do not have tractable ranges, for they depend on the total size of data.

Table 2.3: Distribution of edge weights of DCN created for Arizona SID data

No	Edge weight	Min.	25th Per-centile	Median	75th Per-centile	Max.
1	C_{xy}	1	1	2	7	272831
2	RR_{xy}	0.0011	1.0161	2.1312	6.4054	154739
3	ϕ_{xy}	-0.0240	0.0001	0.0011	0.0032	32.017
4	ϕ'_{xy}	-8.873	0.0162	0.6371	0.9884	14.462

5	SC_{xy}	0.0000	0.0001	0.0004	0.0010	0.9858
6	CC_{xy}	0.0024	0.0116	0.0180	0.0320	0.8392

The DCN is useful for making early predictions, for it captures the implicit inter-relationship of diseases that may or may not be apparent at the PoA. Consider a case where the patient is diagnosed with Diabetes with ketoacidosis (250.1) at the PoA. Follow-up diagnoses show acute respiratory failure (518.81), cerebral edema (348.5), subendocardial infarction (410.7) and anoxic brain damage (437.9). Such co-occurrences are easily observed and quantified in a DCN, making it a good source for PoA diagnoses related feature extraction. Therefore, we constructed and explored a large DCN to find communities of diseases and derived measures that were then used to create input features for the HPEPP model.

2.3.3. Community detection from disease co-occurrence networks (DCN)

Communities are parts of a graph with ties to the rest of the system. Often, they are considered separate entities with unique properties. Health-conditions and diseases may be related due to common causes, symptoms, side-effects, and other latent characteristics (Hassan et al. 2014). We conducted community detection over the DCN to identify distinct clusters of co-occurring diseases. We compared community detection algorithms (Csárdi and Nepusz 2006) in terms of their modularity scores (Newman and Girvan 2003) and determined the Louvain method (Blondel et al. 2008) to be optimal. We ensured the community formation was robust by running the algorithms for different random seeds and selecting the pattern that was repeatedly observed across several iterations. One hundred and twenty distinct non-overlapping disease communities were identified, of which the none major disease groups are shown in Figure 2.3. This visualization is created using the *OpenOrd* layout in gephi, a network analysis tool. The colors used for the communities are as follows – purple: general; light green: liver related; blue: delivery related (mother); yellow: delivery related (infant); red: nerve and bone damage related; light pink:

cardiovascular related; green: cancer related; brown: cerebrovascular related; light blue: accident related; grey: others. Eigenvector centrality is a measure of influence of a node in a network (Easley and Kleinberg 2010). The Eigenvector centrality score of member diseases in the community subgraph networks are considered as membership strength of respective diseases. Table 2.4 shows the top three members of each community ordered in decreasing order of Eigenvector centrality scores.

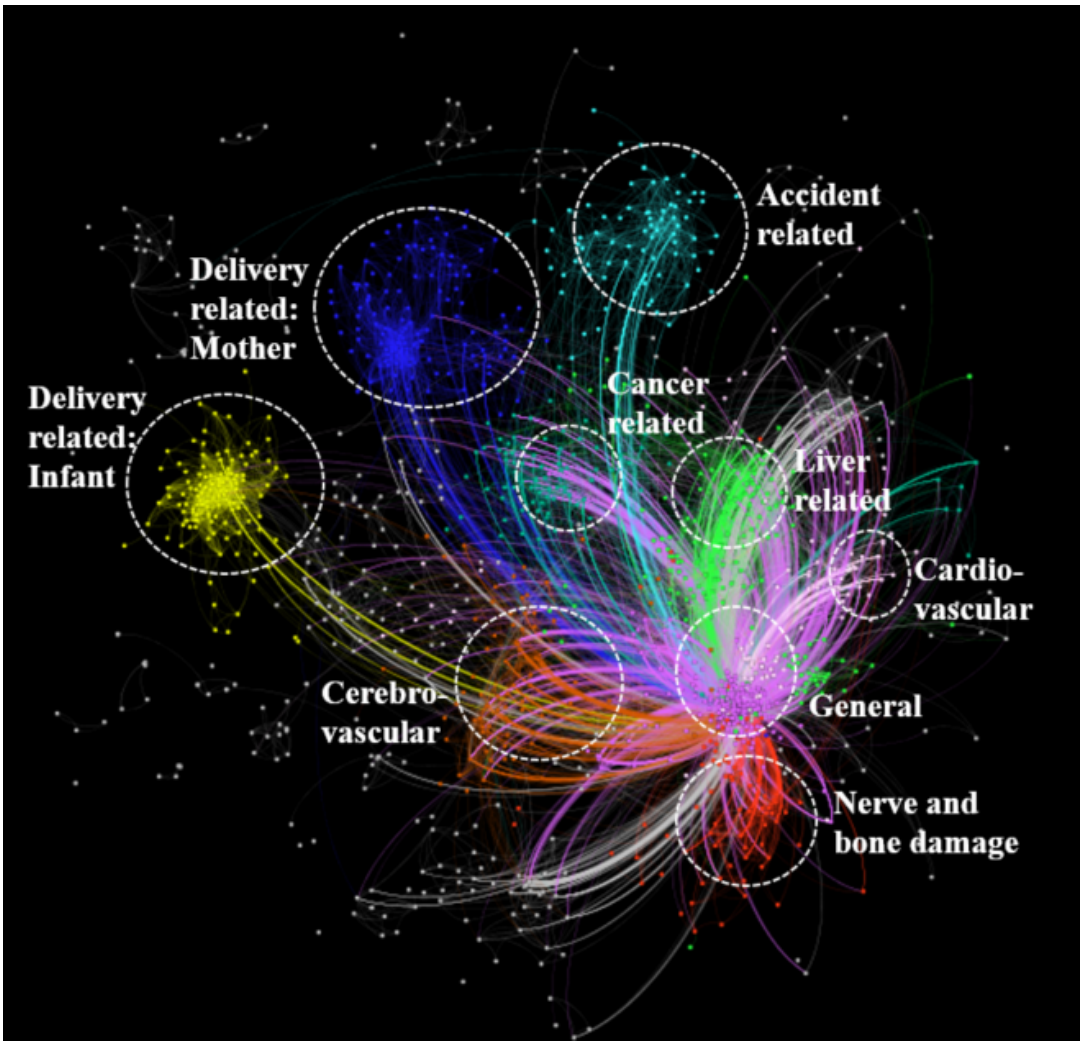


Figure 2.3: Nine distinct disease communities identified using the Louvain community detection method

Table 2.4: Top three diseases in each community

Community	Size (%)	Disease 1	Disease 2	Disease 3
General	19.39	Acute kidney failure	Hypertensive chronic kidney disease	Congestive heart failure
Liver related	8.95	Portal hypertension	Alcoholic cirrhosis of liver	Other ascites
Delivery related (Mother)	8.57	Other current conditions of mother	Obesity complicating pregnancy	Abnormality in fetal heart rate
Delivery related (Infant)	7.18	Anemia of prematurity	Neonatal jaundice	Respiratory distress syndrome
Nerve and bone damage related	5.89	Bone involvement in diseases	Polyneuropathy in diabetes	Ulcer of other part of foot
Cardiovascular related	5.84	Cardiac arrest	Anoxic brain damage	Cardiogenic shock
Cancer related	5.31	Malignant neoplasm of liver	Malignant neoplasm of bone and bone marrow	Malignant neoplasm of lung
Cerebrovascular related	4.60	Dysphagia	Cognitive deficits	Dysarthria
Accident related	4.55	Contusion of lung	Traumatic pneumothorax	Closed fracture of nasal bones

2.3.4. Feature engineering

Feature engineering is the process of creating new features from implicit information in data to improve predictive performance of the model. Using structural properties of the PoA diagnoses in the DCN, we have created two input features: community membership score, and high-cost propensity.

Community membership score: Membership scores for the PoA diagnoses $d \in D$ in the identified communities are defined as follows:

$$MembershipScore(m) = \sum_{d \in D} E_d(m) \quad (2.3)$$

where $E_d(m)$ is the Eigenvector centrality score of disease d in community $m \in M$.

High-cost propensity: The network created using relative risk information between diseases is useful in generating features that contain information about the structural properties of the underlying network. However, to predict for high cost, it will be beneficial to identify a feature that not only takes advantage of network structural properties, but also captures information about propensity of the PoA diagnoses for increasing treatment costs. In other words, the feature should capture the relationship of the PoA diagnoses with high-cost diseases that have not yet been diagnosed for the current patient, but for which a relationship exists in general, for these may potentially be diagnosed in the future. Consequently, we have defined a new feature as follows.

Create an isomorphic directed network of DCN with edge weights given as:

$$CC'_{xy} = CC_{xy}\Omega_y \tag{2.4}$$

$$\Omega_y = \frac{N_y^{highcost}}{N_y} \tag{2.5}$$

where CC_{xy} is the cooccurrence coefficient, previously defined and Ω_y is the fraction of patient encounters recorded with diagnosis y that were high-cost (i.e., cost exceeding the 95th percentile of encounter costs). Accordingly, if disease y occurred 100 times in the training data, of which 10 occurrences have a cost greater than 95th percentile of costs, then Ω_y is 0.1.

Compute high-cost propensity for disease x as follows:

$$HCP_x = \Omega_x + \sum_{y \in Y} CC'_{xy} \tag{2.6}$$

That is, the propensity of a disease x to be a high-cost disease is the sum of Ω_y and CC'_{xy} over it's out-links to diseases $y \in Y$. The first term relates to its independent probability of being present in high-cost encounters, and the second term relates to co-occurrence probability with other diseases

associated with high-cost. The high-cost propensity for the PoA diagnoses $d \in D$ is defined as: $\underset{d \in D}{\text{argmax}}\ (HCP_d)$.

For the 2,025 diseases in the DCN, the mean, median, and standard deviation of high-cost propensity scores are 0.305, 0.168 and 0.387, respectively. Figure 2.4 shows an example of an ego network for *Heat exhaustion, unspecified (992.5)* in the DCN. The high-cost propensity score of Heat exhaustion depends on its weighted directed relationship with high-cost diseases.

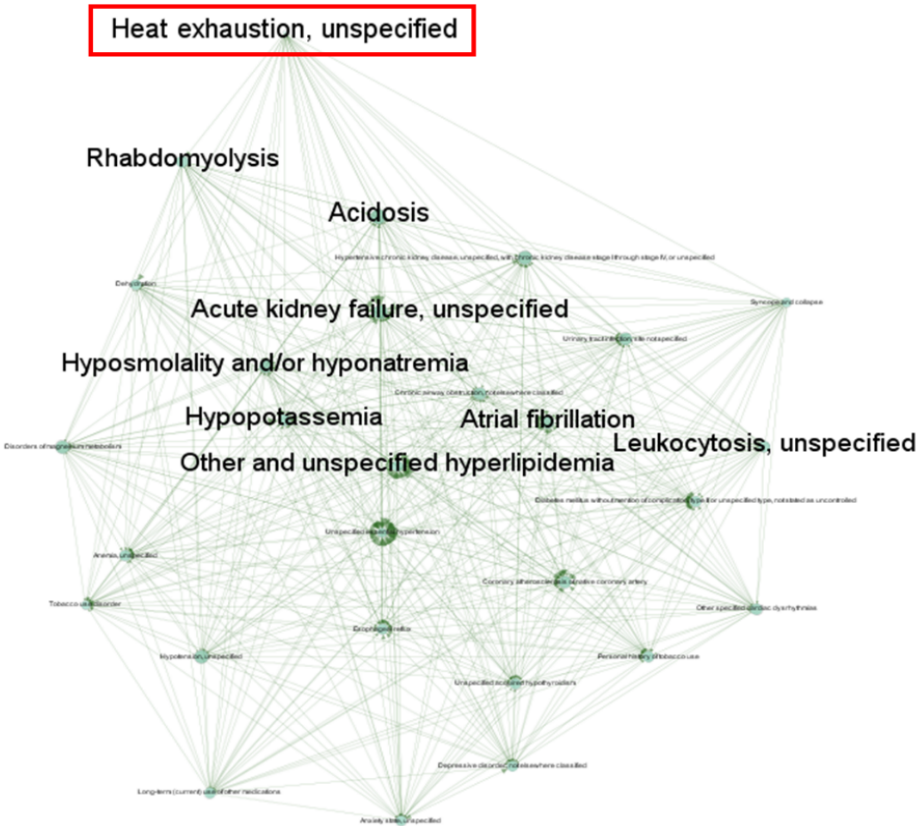


Figure 2.4: High-cost propensity of Heat exhaustion due to its relationship with other diagnoses

2.3.5. Model

We developed four different models with different input features to compare their performance. These models have a combination of the following input feature sets: Baseline, Baseline + CI, Baseline + Network, and Baseline + CI + Network. The baseline features are indicated in Table 1.1. CI stands for the set of twelve Elixhauser comorbidity indicators and the

Charlson Comorbidity Index. Network are the set of features generated using DCN as described previously. With each of the input feature sets, we fitted three tree-based models: Bagged Trees (treeBag), Random Forests (RF) and Gradient Boosting Machines (GBM) (Hastie et al. 2009). We used 10-fold cross validation for parameter tuning and model selection. We discretized unadjusted charges for each patient encounter to two levels: high-cost and not high-cost by splitting at the 95th percentile of the variable (Chechulin et al. 2014).

We carried out data pre-processing such as missing value imputation, Box-cox transformation, centering, scaling, and removal of non-zero variance for all models. Unbalanced outcome classes can pose a problem for predictive models (Moturu et al. 2008). We used a recent clustering based under-sampling method (Y. Wang et al. 2015) to balance the outcome classes of high-cost and low-cost patient encounters. Parameter tuning for models was done using grid search with five parallel cores (Hastie et al. 2009). The input hyper-parameter set for the grid search were number of bagging iterations ={2,3,5,7} for treeBag models, (number of variables randomly sampled as candidates at each split X number of trees grown) = ({4,8,12}X{100}) for RF models and (interaction depth X shrinkage X number of trees grown) = ({2,3}X{0.01,0.1,0.3}X{100,200}) for GBM models. The data set was separated randomly into a training set and test (holdout/validation) set in the ratio of 75:25. All models were trained using the caret package (Kuhn 2015) in R in a 64GB RAM, 2.60GHz processor machine.

2.4. Results

Accuracy, recall, and precision for the high cost class prediction over the test data are shown in Figure 2.5. We see that predictive performance is the best for the model with the input feature superset comprising all features. Sensitivity or recall drops somewhat when network features and comorbidity indicators are added to the derived features, but both specificity and

accuracy improve significantly. Accuracy of models with Baseline + Network is better than Baseline + CI, but including CI features into the model improves the accuracy further.

We also conducted experiments to test the contribution of the proposed network-based features when included in existing models for high cost patient prediction. We fit models using features specified in two recent studies conducted by Sushmita et al. (2015) (abbreviated in figures as: Sush) and Chechulin et al. (2014) (abbreviated as: Chec), We compared model fit when network features were included to existing feature set used by Sushmita et al. (Sush + Network) as well as when network features were included to existing feature set used by Chechulin et al. (Chec + Network). As seen in Figure 2.6 (and mirroring the pattern see in Figure 2.5), we find that network features significantly improve the specificity and overall accuracy of existing models.

The variable importance scores (Hastie et al. 2009) of top ten features for the tree-based models are shown in Table 2.5. The importance of a variable is proportional to the change in out-of-bag prediction accuracy when it is excluded from the input feature set for random forest models (Breiman 1999). In contrast, for bagging and boosting models, it is the average reduction in the loss function at each split of the variable (Loh and Shih 1997). Consequently, the order of the important features is slightly different across the tree-based models due to the non-linear fit of the input feature space of these models. We observe that the Central-nervous system related community (a group of 29 disease diagnoses) has a high variable importance score in the HPEPP model. Other than a few CCS, Elixhauser codes and CCI measures, most of the features in the list are network-based measures. High-cost propensity, membership score for the *general* disease community, CCI and membership score for cardiovascular related disease community are among the top-five important features in all three models.

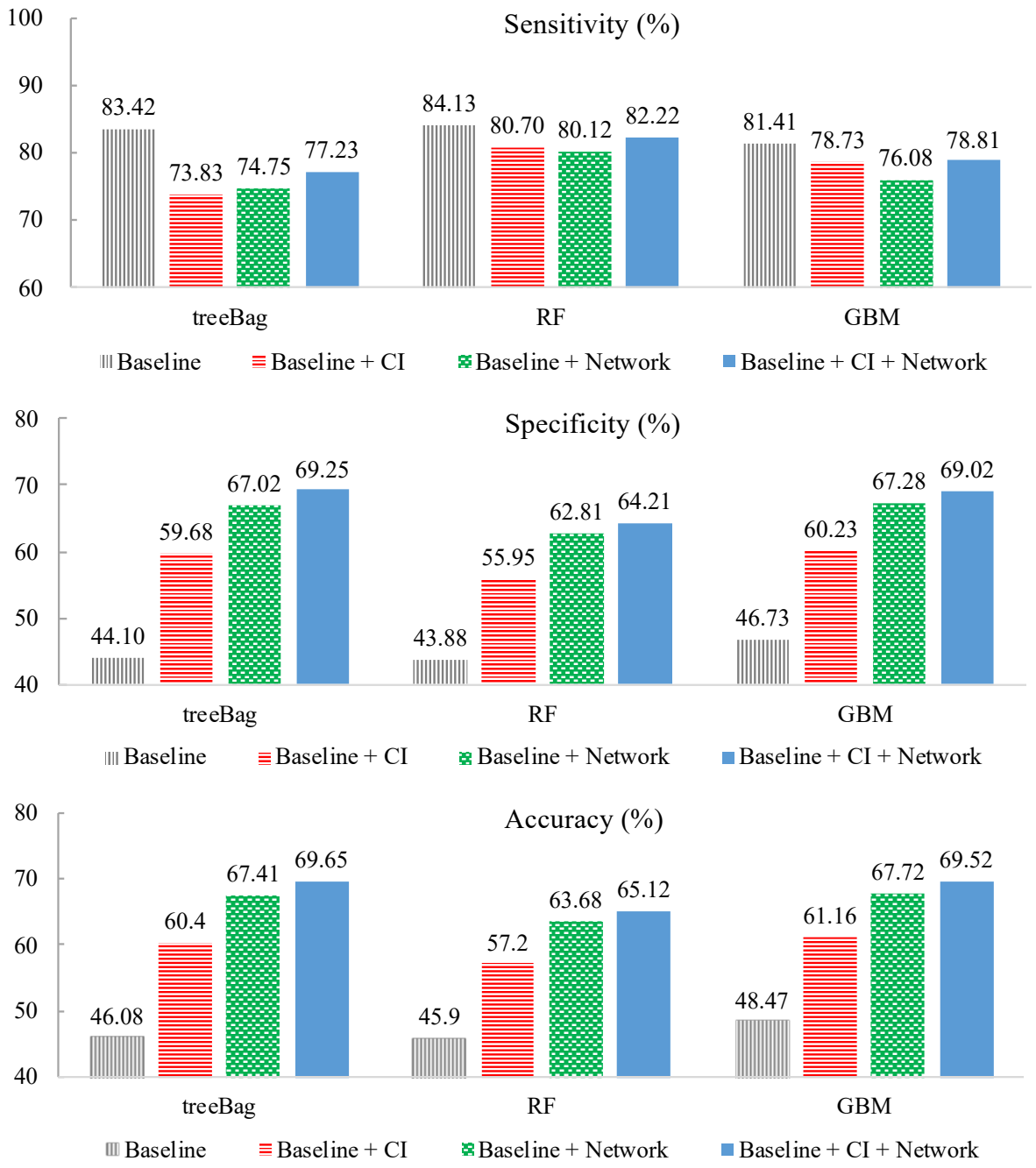


Figure 2.5: Sensitivity, Specificity and Accuracy for models with four different input feature sets: (a) Baseline, (b) Baseline + CI, (c) Baseline + Network, and (d) Baseline + CI + Network

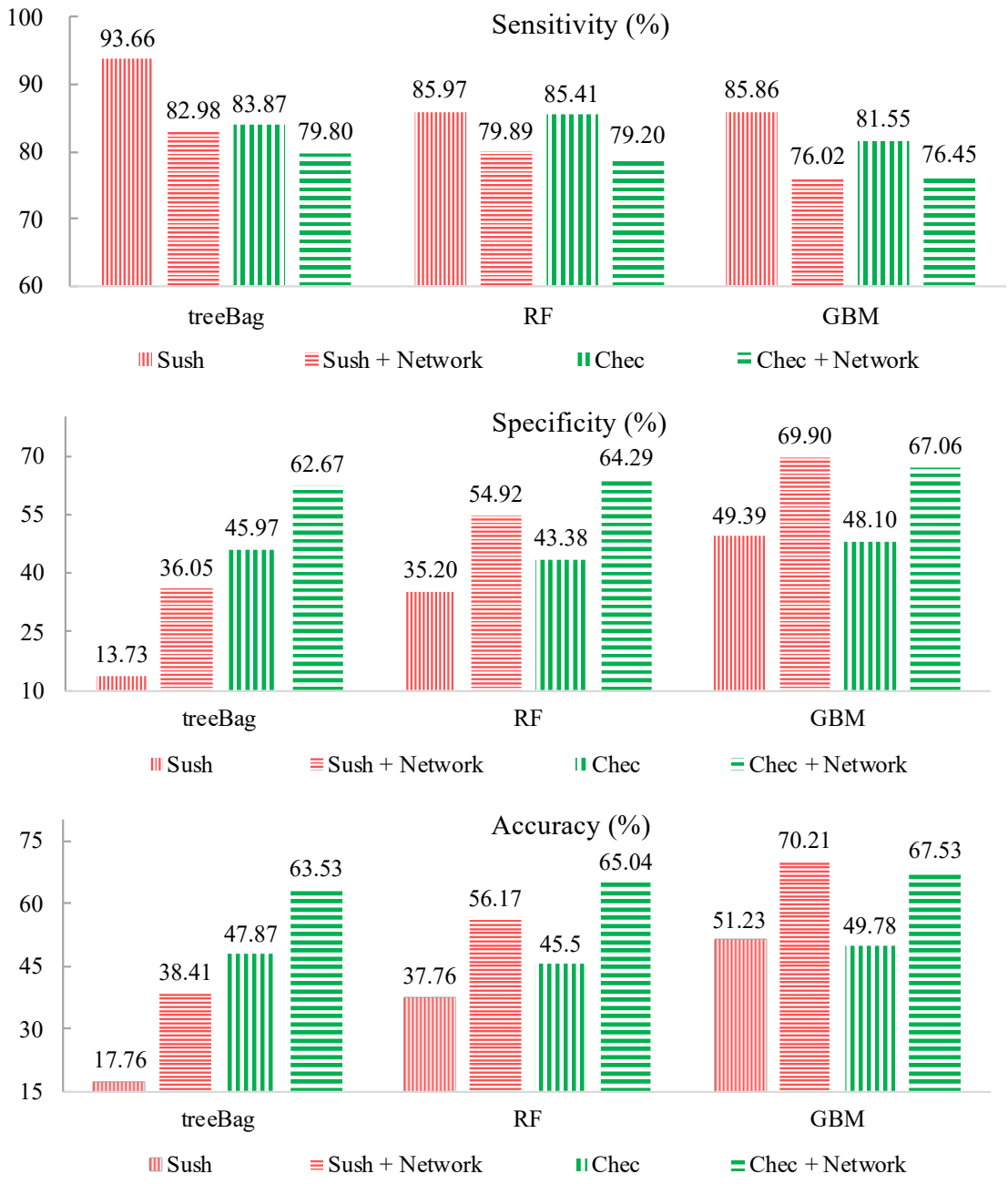


Figure 2.6: Sensitivity, Specificity and Accuracy for models used in previous studies with and without proposed network features

Table 2.5: Top ten features in models

TreeBag	RF	GBM
High cost propensity	***High cost propensity***	***General***
Central-nervous system related	***General***	***High cost propensity***

General	***Cardiovascular related***	CCI
Cardiovascular related	CCI	***Cardiovascular related***
CCI	***Central-nervous system related***	***Liver related***
Elixhauser - Valvular disease	***Liver related***	***Central-nervous system related***
CCS - Injury and poisoning	***Nerve and bone damage related***	CCS - Complication of pregnancy, childbirth
Cerebrovascular related	***Cerebrovascular related***	***Cerebrovascular related***
Elixhauser - Weight loss	Elixhauser - Weight loss	***Nerve and bone damage related***
Delivery related: Mother	***Cancer related***	***Cancer related***

The group means of *high cost* and *non-high cost* class data are significantly different for network-based data and are presented in Table 2.6 with Cohen's d effect sizes. The Cohen's d for membership score for general and cardiovascular disease communities is the highest, consistent with the variable importance list in Table 2.5.

Table 2.6: Mean differences of network-based features between high-cost and non high-cost encounters

Feature	Mean - High cost	Mean - Non-high cost	Difference	Cohen's d
High-cost propensity	1.6343	1.2206	0.4138	0.24
General	2.8058	1.8839	0.9219	0.15
Liver related	0.0878	0.0428	0.045	0.06
Delivery related: Mother	0.0063	0.0806	-0.0743	0.07
Delivery related: Infant	0.0684	0.0109	0.0575	0.12
Nerve and bone damage related	0.1589	0.0851	0.0738	0.06
Cardiovascular related	0.0582	0.0084	0.0497	0.15
Cancer related	0.0507	0.0303	0.0204	0.03
Cerebrovascular related	0.0040	0.0023	0.0017	0.03
Accident related	0.0383	0.0068	0.0315	0.12

2.5. Discussion

For the high cost prediction problem, we primarily focus on increased sensitivity (i.e., proportion of high-cost patients correctly identified) for a model to be useful. However, the specificity and accuracy need to be checked as auxiliary measures. In the extreme case of a model predicting every encounter as high-cost, sensitivity may be 100% for high-cost outcomes, but the model would be useless because it may have very low overall accuracy. Classification models with a 95 percentile cut-off for high-cost patients have been observed to have low accuracy in the past (Chechulin et al. 2014). Figure 2.5 shows that models with network features included have uniformly better performance in terms of overall accuracy.

The HPEPP model can reduce US healthcare costs by improving sensitivity or recall of high-cost patients as well as reducing the very high false positive rate. Suppose we estimate that health costs reduce by 0.1% due to pre-emptive patient care on 1% of the high-cost patient encounters (Bates et al. 2014). On considering the ~$35 million inpatient visits and overall expenditure of ~$3 trillion in the U.S. for FY 2013 (cms.gov 2016; HCUP 2013), cumulative savings due to this reduction is approximately $3 billion (Bates et al. 2014). Owing to the sheer volume of patient encounters per year in U.S. healthcare, early prediction of high cost patient encounters, even with a small performance improvement, can result in significant improvements in patient care targeting and in-turn, a reduction in overall healthcare costs. The training and implementation cost of the HPEPP model is negligible as the feature engineering process is a one-time effort. Even though the sensitivity or recall of high cost patients does not improve significantly due to inclusion of network-based features, we do get a better predictive performance in terms of accuracy and specificity, which are tied to a reduction in false positive rates. A drop in false positives due to improvement in specificity using HPEPP model can greatly reduce annual case management costs of high-risk patients. The network created using the Arizona SID dataset

is generalizable to other U.S. states due to the fair number of observations considered in the dataset as well as the fact that Arizona's population is representative of the diversity found in a national sample (census.gov 2016).

We have used the Arizona SID dataset for creating the network as well as for training and testing the HPEPP model. The Arizona SID dataset is a limited public use dataset that does not have patient identifier or patient history. As patient history is a significant predictor in patient encounter predictive models (Bertsimas et al. 2008; Chechulin et al. 2014; Sushmita et al. 2015), the performance of the HPEPP model may improve using such information. The HPEPP model is yet to be tested in the field. For field testing, our model would need to be used for predicting high-cost patient status at the PoA followed up with high-cost patient status at the time of discharge. This will give us a realistic estimate of the results of using our model.

2.6. Conclusion

With the increase in centralized healthcare information systems, there is great potential for big data approaches using innovative methods to identify at-risk patients. In this study, we introduce the novel idea of exploring disease co-occurrence networks to generate new input features for a model to predict high-cost patient encounters at the Point-of-Admission. To our knowledge, our paper is the first study to use hospital discharge data to build a disease co-occurrence network. We use this network in predicting high cost patients at the early and critical stage of PoA, so timely intervention can be carried out to improve patient care and reduce costs. We trained and validated the baseline as well as updated models using state-of-the art data mining techniques. We found evidence of a steady performance improvement by including network-based features.

Our big data method combining network analysis and data mining is useful for developing applications for predicting high-cost patients at PoA, so they can receive the care necessary to

improve their healthcare outcomes. The novel technique of incorporating network science into predictive models introduced in this study can be applied to model a variety of big data healthcare problems. This study contributes to knowledge discovery via exploratory analysis such as community detection as well as introduces a new empirical index, cost propensity score, which predicts high cost diseases. Our ongoing work includes investigating historical information about discharged patients to predict future readmission and high-cost encounters. Proposed areas for future work include testing our predictive model for dynamic feature updates during treatment progression, fusion of information from different networks, and predicting scores for related outcomes such as mortality, length of stay, and re-admission probability.

3. ESSAY 2: ANALYZING INCOMPLETE DATA WITH BLOCK-WISE MISSING PATTERNS

3.1. Introduction

With the growing use of big data, data collection and processing have not only become more important in the data analysis life cycle but have also become more challenging. User-generated textual content (e.g., tweets, online reviews, etc.), streaming data (e.g., sensor-data pertaining to the environment, heart rate, geo-location, etc.), and cross-sectional facts (such as census data, hospital administrative information, population health indicators, etc.) are rich sources of information for IS research (Bardhan et al. 2015; Staats et al. 2017; Yin et al. 2014). Researchers often combine them in a logical way to answer a wider range of questions (Jiang et al. 2007). However, there are challenges associated with combining and processing data, termed data plumbing or data munging, which take up considerable time and effort of a data analyst (Parssian et al. 2004). A common problem in multi-source data integration is the generation of an incomplete resultant dataset. For instance, when combining data from multiple sensors, or from sensors and online streaming resources, there are instances in the combined dataset when one has no recorded values from one or more sources. Such gaps in a dataset can be termed as *block-wise missing* as shown in Table 3.1. Occurrence of block-wise missing values is not only restricted to heterogeneous data source integration, but also extends to other common scenarios. For example, Figure 3.1 illustrates block-wise missing values generated by doing an outer join (two tables as sources) in relational databases.

There are three traditional approaches to process and analyze incomplete data: (i) pruning, defined as reviewing and systematically removing sparse rows/columns; (ii) choosing special models, which means selecting models that can inherently handle missing values such as Classification And Regression Trees (Breiman et al. 1984), Bayesian Networks (Friedman 1997),

etc.; and (iii) treatment or using imputation methods to replace missing values in a dataset with estimated values (Schafer and Graham 2002). The first approach is useful when the ratio of missing values to populated values in the dataset is low (i.e., there is minimal loss of data), and the pruned dataset is still representative of the initial dataset. The second approach precludes the use of several commonly adopted models (such as regression models, tree-based models, neural networks, etc.) that require complete data. In the third approach, incomplete data is transformed to a "complete" dataset by imputing values using univariate/multivariate machine learning or statistical methods.

Table 3.1: Example of block-wise missing data

Obs.	Pressure	Humidity	CO	CO2
1		56.09	0.35	
2		56.1	0.39	
3		56.1	0.43	
4	55.5			755.38
5	45.43			655.7
6	43.54			
7	44.93			
8	45.24			
9	50.33			
10	42.21	58.33	0.52	664.2
11	56.99	55.82	0.47	709.5

Patient ID	Hemoglobin levels	Mean Heart rate
101	300	70
102	350	110
103	200	120
104	345	90
105	320	80

⋈

Patient ID	Mean sitting bout (%)	Caffeine consumption/hour
103	80.0	0.2
104	74.5	1.4
105	88.5	0.1
106	81.1	
107	63.2	0.8

Patient ID	Hemoglobin levels	Mean Heart rate	Mean sitting bout (%)	Caffeine consumption/hour
101	300	70		
102	350	110		
103	200	120	80.0	0.2
104	345	90	74.5	1.4
105	320	80	88.5	0.1
106			81.1	
107			63.2	0.8

Figure 3.1: Outer joins in relational databases resulting in datasets with block-wise missing patterns

An alternative (fourth) approach proposed for predictive modeling with incomplete data is to use reduced (features) models (Friedman et al. 1996; Schuurmans and Greiner 1997), which involves employing only those features that are known for a given test data tuple. For each pattern of missing features, a different model is to be trained. The naïve implementation of this approach is not commonly used in practice due to the cost of storing all possible models or computing models at run time. We propose a new method called Block-wise reduced modeling (BRM) method for handling incomplete data created through multi-source data integration that builds on the idea of reduced modeling. Our method can be used to analyze incomplete datasets that have block-wise missing patterns. BRM has two phases for handling incomplete training data and test data, respectively. In the first phase, our algorithm automatically groups the training data into overlapping subsets with different combinations of feature columns that contain only populated

values. We then train candidate models over each subset. In the second phase, our algorithm assigns each instance of the test dataset to one of the candidate models using a similarity scoring mechanism. BRM is an improvement over the naïve reduced modeling approach in the following two ways. One, it is scalable as training is done in advance instead of at run time as originally proposed, and two, it can be used for explanatory modeling (e.g., linear regression, logistic regression, etc.) as well as predictive modeling (e.g., tree-based models, neural networks, etc.) of incomplete data with block-wise missing patterns.

We apply our method to the problem of predicting hourly demand of rental bikes in a city-wide bike sharing program. Complete data is available for this dataset; therefore, the impact of dealing with missing values can be compared against having full information. We validate our method by simulating 25%, 50%, and 75% block-wise missing values in the dataset and comparing model fit with the complete dataset. We further demonstrate the utility, cross-domain applicability and scalability of our method by applying it to a healthcare problem of modeling patient visit costs using incomplete electronic health records (EHR) data. The bike-sharing dataset is suitable for conducting simulations of block-wise missing patterns and making performance comparisons of our method with existing methods. On the other hand, the healthcare dataset with 13 million records is useful to demonstrate successful application of our method to a big data problem with incomplete data, where it is infeasible to adopt existing methods. We show that our method supports prediction and can provide overall feature scores even in the presence of a relatively large proportion of missing values.

The remainder of this paper is structured as follows. In Section 3.2, we discuss related work. In Section 3.3, we describe our method in detail. We then evaluate it using the bike sharing dataset and demonstrate its application to a healthcare problem in Section 3.4. Section 3.5 contains the discussions and limitations. Finally, the conclusions and directions for future work are in

Section 3.6.

3.2. Related Work

It is well-known that incomplete data or the presence of missing values in a dataset can pose challenges for data analysis. Not addressing the missing values issue can lead to faulty conclusions about predictor-response effects as well as deterioration in predictive performance of the model (Schafer and Graham 2002). Block-wise missing values commonly occur in scenarios such as multi-source data integration, outer joins, and non-overlapping start times of sensor data (e.g., in an experiment where a CO2 sensor may start streaming data from day 1, but an NO2 sensor may start streaming data only from day 4). In the case of single domain (simple) datasets, missing values occur due to several reasons including non-response, measured features being out of range, random signal interference, or removal of noisy data. They are dispersed across the data randomly such that there is no systematic pattern of "missingness". However, in the case of block-wise missing values, data in a subset of feature columns is missing for contiguous observations, as shown in Table 3.1.

As mentioned in Section 3.1, the three traditional approaches to process incomplete data are pruning, choice of special models, and treatment. The first approach is to prune or discard columns or rows that are sparse. This approach results in loss of information and becomes sub-optimal if more than 5% of the observations have missing values (Buuren 2012). The second approach is to select models that do not require complete data for training. Data mining models such as classification and regression trees (CART), handle missing values during creation of classification trees using surrogate splits (Friedman et al. 2001); that is, when considering a predictor for a split, only the observations for which that predictor is not missing are used for training. These models are known to be biased towards features with fewer missing values in the dataset; hence, the models can be misleading when comparing feature importance in the presence

of incomplete data. The second approach also precludes the use of several commonly adopted models (such as regression models, tree-based models, neural networks, etc.) that require complete data. The third approach involves imputation or replacing a missing value with a representative value using univariate or multivariate estimation methods. It is the most common treatment to generate complete data from incomplete data (Little and Rubin 2002).

There are several well-developed statistical and machine learning-based univariate and multivariate imputation methods such as single value imputation (Stekhoven and Buhlmann 2012), expectation maximization (Graham 2012), nearest neighbor (Batista and Monard 2003), and multiple imputation (Melville and McQuaid 2012; Sterne et al. 2009). Simple imputation methods such as mean/median value replacement can understate variability in the imputed features (Buuren 2012). However, sophisticated imputation methods (Van Buuren and Groothuis-Oudshoorn 2011; Stekhoven and Buhlmann 2012) pose other problems. They are not only computationally intensive, but also distort the data by over-synthesizing artificial replacements for missing values as the proportion of missing values increases in the dataset (Buuren 2012). They also become computationally infeasible with large datasets having high proportions of missing values.

The example data shown in Table 3.1 is a representative sample of a real incomplete dataset containing block-wise missing patterns with 11 observations and four features. We see approximately 27%, 55%, 55%, and 64% of values for the features Pressure, Humidity, CO, and CO_2 are missing respectively. Let us consider this example to compare existing methods for processing incomplete data. Pruning rows with missing values (i.e., list-wise deletion) will leave us with only two out of 11 observations (i.e., rows 10 and 11). Column-wise deletion is also not a feasible option with multiple columns containing block-wise missing patterns, and the exclusion of multiple features would bias the model. Naïve treatment approaches such as mean value substitution will understate the variability as three or more observations in each feature will have

the same values. On the other hand, sophisticated imputation methods assign values for missing cells using iterative procedures, generating a resultant dataset that may be very different from the original data (Buuren 2012). As a result, traditional approaches are not suitable for processing incomplete data with block-wise missing patterns, especially as dataset size and the extent of missing values increases.

There is a need for new methods to handle large incomplete datasets with block-wise missing values that not only minimize inference errors, but are also fast, efficient, and scalable. Two recent studies (Srinivasan, Currim, et al. 2016; Xiang et al. 2013) focused on data analysis with incomplete data with block-wise missing patterns. Xiang et al. (2013) develop a feature selection method for high-dimensional medical data that applies a group penalty on features with similar missing value patterns. Post feature selection, they train a random forest classifier on complete data. Their method addresses the problem of feature selection using a sparse learning framework, but it is not suitable for predictive modeling where test instances are incomplete. Srinivasan et al. (2016) propose an ensemble-based method that combines predictions from models trained over subsets of original data, where test instances are handled using traditional imputation approaches. Their method is not optimized for operational performance for large datasets since imputation is still required for test instances.

The reduced modeling approach was proposed by Schuurmans and Greiner (1997) two decades ago. The basic idea was to build individual predictive models for features that have values populated in a test instance. This approach required training a different reduced candidate predictive model for every test instance. Since this was computationally intensive and time-consuming, an improvement was suggested for classification trees as a hybrid scheme of reduced modeling and imputation for each test instance (Saar-Tsechansky and Provost 2007). In the hybrid scheme, the number of reduced candidate models to be stored is determined by a heuristic that

maximizes a utility function proportional to the differences in the accuracy of reduced models and accuracy of model using imputation across all missing patterns in data. This method is only suitable for classification trees and becomes computationally intensive for datasets with five or more features having missing values. Therefore, the naïve approach and the hybrid scheme of reduced modeling cannot be directly used for analyzing large datasets with block-wise missing patterns.

In the big data era, multi-source data integration is common, often resulting in block-wise missing patterns in resultant data. There is a need to develop a new method, as existing methods used for processing incomplete data are not optimal for analyzing data with block-wise missing patterns. We propose a new method called the block-wise reduced modeling (BRM) method, that is a scalable improvement of the naïve reduced modeling approach. Our research contribution is important for information systems research and applications where there is a need for integrating data from several disparate sources.

3.3. The Block-wise reduced modeling method

The BRM method has two phases: (i) training, which generates overlapping subsets and training candidate models, and (ii) prediction, which allocates test instances to candidate models. In the training phase, overlapping subsets with different sets of feature columns are generated from the training data. Separate/candidate models are then developed for each training subset. In the prediction phase, for each test instance, the candidate model with the feature set having the highest similarity with the test instance feature vector is selected for prediction.

3.3.1. Phase 1: Training

The four steps in the algorithm are creating non-overlapping subsets, creating overlapping subsets, training candidate models over overlapping subsets, and computing the global coefficient scores.

3.3.1.1. Creating non-overlapping subsets

To create non-overlapping subsets, we first generate a missing value indicator matrix for a given input feature matrix. That is, for a set of input features represented as a matrix D_{train}, we construct a missing value indicator matrix as $D'_{train} = \{x' \in \{0,1\} \mid (x = \emptyset \rightarrow x' = 1) \wedge (x \neq \emptyset \rightarrow x' = 0)\}$. A cell in the missing value indicator matrix contains 1 if there is a missing value in corresponding cell in the input data matrix, and 0 otherwise. Using standard clustering methods (e.g., k-means algorithm) over the missing value indicator matrix, we identify subsets of data with different combinations of the feature columns on which we can fit independent models. These subsets are non-overlapping because each observation in the training data belongs to only one of the subsets determined by the clustering algorithm. For each subset, we discard feature columns with more than 95% missing values. If less than 5% of data in the subsets are missing, we impute them using single value imputation methods (Buuren 2012). If subsets have features with a percentage of missing values between 5% to 95%, we re-run the k-means algorithm with a higher number of clusters to generate subsets that contain only populated values.

Table 3.2 contains the missing value indicator matrix corresponding to the example in Table 3.1. Observations are grouped using k-means clustering into four subsets indicated by Cluster no. column. The four non-overlapping subsets can be denoted as follows: $S_1^{\{Humidity,CO\}} = \{1,2,3\}$, $S_2^{\{Pressure,CO_2\}} = \{4,5\}$, $S_3^{\{Pressure\}} = \{6,7,8,9\}$, and $S_4^{\{Pressure,Humidity,CO,CO_2\}} = \{10,11\}$, where the elements are row numbers in the table and $S_i^{c_i}$ is the i^{th} subset containing a set of columns c_i. For ease of representation, we represent these subsets as $S_1(Humidity, CO)$, $S_2(Pressure, CO_2)$, $S_3(Pressure)$ and $S_4(Pressure, Humidity, CO, CO_2)$. Here, the terms inside the parentheses

indicate the feature columns used for the respective non-overlapping subsets of data.

Table 3.2: Indicator matrix for missing values in example data

Obs.	Pressure	Humidity	CO	CO_2	Cluster no.
1	1	0	0	1	1
2	1	0	0	1	1
3	1	0	0	1	1
4	0	1	1	0	2
5	0	1	1	0	2
6	0	1	1	1	3
7	0	1	1	1	3
8	0	1	1	1	3
9	0	1	1	1	3
10	0	0	0	0	4
11	0	0	0	0	4

3.3.1.2. Creating overlapping subsets

The overlapping subsets build upon the non-overlapping subsets to utilize the maximum possible number of observations for training candidate models in the subsequent step of the training phase. For each non-overlapping subset defined previously, we define a partial order of subsets such that the features in lower order are subsets of features in higher order subsets. For the previous example, the non-overlapping subsets have the partial order relationships given by:

$S_1(Humidity, CO) \leq S_4(Humidity, Pressure, CO, CO_2)$

$S_2(Pressure, CO_2) \leq S_4(Humidity, Pressure, CO, CO_2)$

$S_3(Pressure) \leq S_2(Pressure, CO_2) \leq S_4(Humidity, Pressure, CO, CO_2)$

$S_4(Humidity, Pressure, CO, CO_2)$

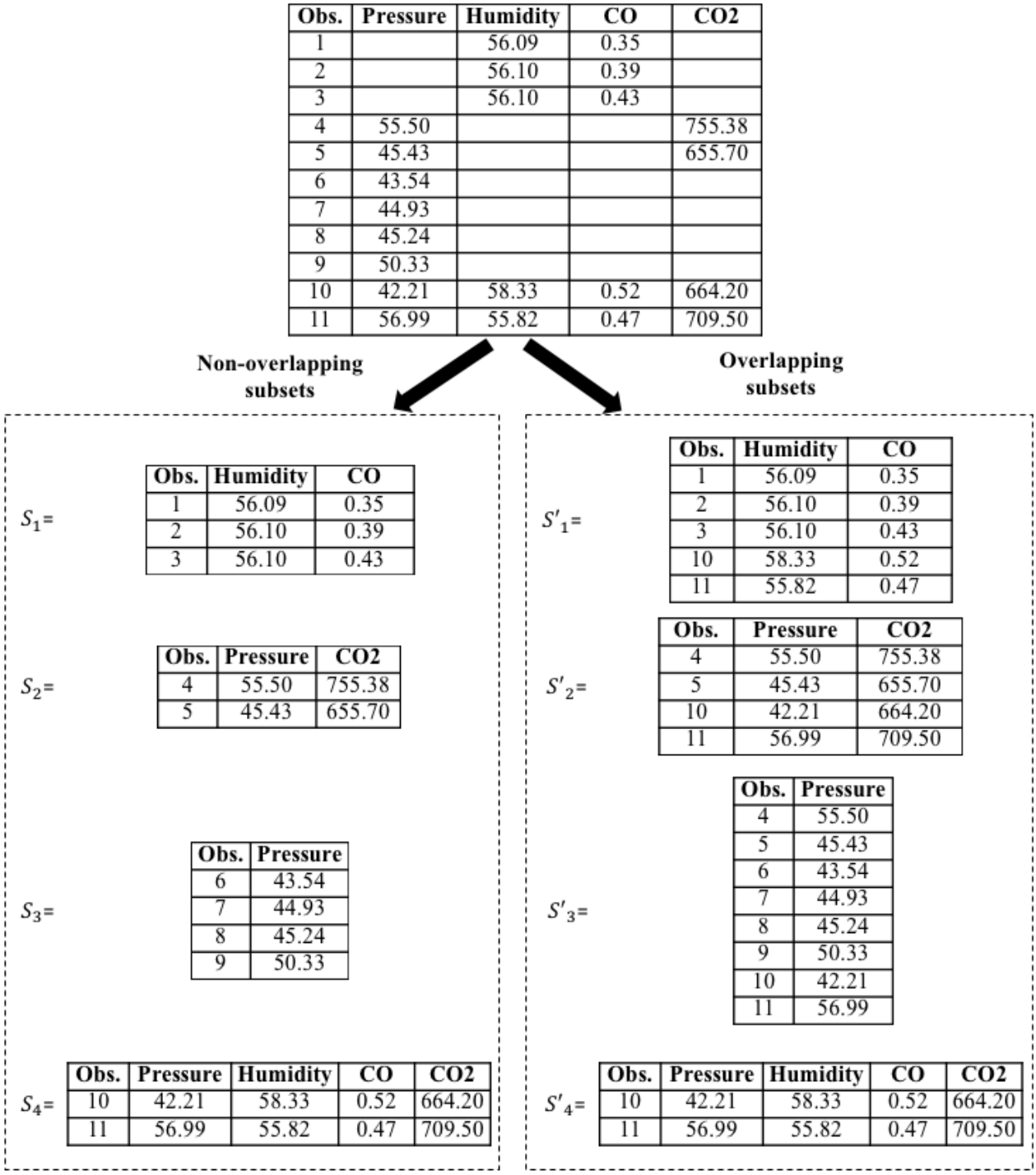


Figure 3.2: Overlapping versus non-overlapping subsets for training candidate reduced models

The data contained in the subsets that are highest in the partial order can be included in the lower ordered subsets, for they contain elements equal to or greater than the set of feature columns

lower in the partial ordering. Using this logic, the dimensions of the overlapping subsets in the given example are as follows:

$S'_1(Humidity, CO) = S_1(Humidity, CO)$[1,2,3,10,11]

$S'_2(Pressure, CO_2) = S_2(Pressure, CO_2)$[4,5,10,11]

$S'_3(Pressure) = S_3(Pressure)$[4,5,6,7,8,9,10,11], and

$S'_4(Pressure, Humidity, CO, CO_2) = S_4(Pressure, Humidity, CO, CO_2)$[10,11]

where $S_i(col_i)[row_i]$ stands for the i^{th} over-lapping subset with col_i as the column vector (features) and row_i is the row vector (observations).

Figure 3.2 shows overlapping, and non-overlapping subsets generated for the example in Table 3.1. If we only consider non-overlapping subsets $\{S_1, S_2, S_3, S_4\}$ for developing candidate models, the number of observations included for training for a given set of features is sub-optimal. However, by creating overlapping subsets $\{S'_1, S'_2, S'_3, S'_4\}$, one can insure that for each set of features, the maximum possible number of observations for training each candidate model are utilized.

3.3.1.3. Training candidate models over overlapping subsets

Using the previous steps, the training data is divided into multiple overlapping subsets that contain only populated values. For each subset, we train independent candidate reduced models such that columns of the subsets are input features of the candidate models.

3.3.1.4. Computing global coefficient scores

We adapt a pooling strategy from literature developed for the multiple imputation method (Rubin 1976). The original strategy averages the values of the parameter estimates across the missing value samples generated in the multiple imputation method to obtain single point estimates (Buuren 2012). Our approach combines results from datasets with different dimensions as shown in the following equation:

$$B_p = \frac{\sum_{i=1}^{k} B_{p,i} \cdot W_i}{\sum_{i=1}^{k} W_i} \quad (3.1)$$

$$W_i = \frac{\sqrt{\dim(i)}}{\varepsilon_i^2} \quad (3.2)$$

where,

B_p = Global co-efficient of p^{th} feature

k = Number of candidate reduced models with significant coefficient for feature p

$B_{p,i}$ = Co-efficient of p^{th} feature in i^{th} model

W_i = Weighting factor for i^{th} model

$\dim(i)$ = (No. of rows) x (No. of columns) is the dimension of the i^{th} training subset

ϵ_i^2 = Mean squared error (MSE) of i^{th} model

The weighting strategy accounts for the size of the dataset (rows), the number of features (columns) as well as in-training predictive performance of candidate reduced models.

Figure 3.3 summarizes the algorithm for the training phase of our method.

Input: Training data D_{train} with at one feature having at least one missing value.

Initialize: Create missing value indicator matrix D'_{train} corresponding to missing value pattern in D_{train} (1 if missing, 0 if value present in each cell)**.**

Algorithm:

i. Create non-overlapping subsets: Cluster tuples of D_{train} into k subsets $\{S_1, S_2, \dots, S_k\}$. Discard columns in each subset with more than 95% missing values.

ii. Create overlapping subsets: Determine the partial order of columns. Assign observations/tuples to overlapping subsets $\{S'_1, S'_2, \dots, S'_k\}$ such that each observation in S'_m is included in S'_n if $S_n \leq S_m$ in the set $\{S_1, S_2, \dots, S_k\}$.

iii. Train candidate models $\{M_1, M_2, \dots, M_k\}$ over each subset in $\{S'_1, S'_2, \dots, S'_k\}$.

iv. Compute the global coefficient score for each feature p as B_p (Refer equation (1.1)).

Output: A set of candidate reduced models and global coefficient scores for each feature in D_{train}.

Figure 3.3: Phase 1 of BRM – Training

3.3.2. Phase 2: Prediction

Big data applications involving integration of data from multiple heterogeneous sources may be expected to have the problem of block-wise missing values in both the test and training samples. Block-wise missing patterns can be considered to be similar across training and test data if the number of sources is assumed to remain the same. We use the Jaccard similarity index (Real and Vargas 1996) to compare the feature vector of a test instance and assign it to the nearest candidate reduced model. That is, the candidate model with the input feature set that has a maximal overlap with the test input feature vector is selected for making a prediction. In case of a tie in Jaccard similarity scores for two or more models, we select the model with higher number of input features. The above steps are repeated for each test instance.

Consider the task of predicting a test instance of the example in Table 3.1. Let us denote the set of candidate reduced models for overlapping subsets $\{S_1', S_2', S_3', S_4'\}$ defined in phase 1 for example, as $\{M_1, M_2, M_3, M_4\}$ in Table 3.1. A test instance with input feature vector $\{Humidity, CO, CO_2\}$ will have a missing value indicator vector given by $[1, 0, 0, 0]$. The Jaccard similarity scores for this test instance with the overlapping subsets are 0.75, 0.25, 0.00, and 0.75 respectively. We choose model M_4 as the reduced model for prediction in this case for two reasons. One, it is one of the models with the highest similarity scores tied with M_1, and two, the number of features is higher in the training set S_4' than the test instance and hence, uses the information more efficiently compared to S_1', whereas the value for CO_2 in test instance is not used for prediction by M_1.

Figure 3.4 summarizes the algorithm for the prediction phase of our method.

Input: Test instance $d \in D_{test}$ and candidate models trained in Phase 1.
Initialize: Create missing value indicator vector d' corresponding to missing value pattern in d (1 if missing, 0 if value present in each cell).
Algorithm:

i. Compute Jaccard similarity index J_i of (not-null) features in d' with feature indicator vectors of each subset $S_i' \in \{S'_1, S'_2, \ldots, S'_k\}$. Select model M_i as:

$$\underset{M_i \in \{M_1, M_2, \ldots, M_k\}}{\operatorname{argmin}} J_i$$

ii. Predict outcome using candidate model M_i

Output: Prediction for test instance d

Figure 3.4: Phase 2 of BRM – Prediction

3.3.3. Computational complexity of BRM

The computational complexity of the training phase of BRM is driven by the clustering algorithm. For instance, the k-means algorithms implemented in standard packages run in $O(nkq)$ time, where n = number of observations, k = number of clusters, and q = maximum iterations until convergence (constant). In the prediction phase of BRM, computing the Jaccard index takes $O(k)$ time per test instance. Existing methods, such as single value imputation (SVI) and multiple imputation (MI), are slower than BRM as shown in Section 3.4, mainly due to the iterative procedures for parameter estimation (Buuren 2012). The existing reduced modeling approach is a special case of BRM. In the existing approach, every combination of the p input features can form a missing pattern, and therefore, $2^p - 1$ models need to be trained. In BRM, the number of candidate models is equal to the number of distinct block-wise missing patterns, which is typically proportional to the number of heterogeneous data sources. While in the worst case, the computational cost of the training stage of our method (a one-time, or non-recurring cost) can increase exponentially with the number of distinct block-wise missing patterns, in practice, the number of heterogeneous data sources is likely to be a reasonably small number (Zheng 2015).

3.4. Analysis and Evaluation

We evaluated our method by applying it to the problem of predicting hourly demand of rental bikes in a bike sharing program. We simulated missing values in the dataset incrementally

to measure the changes in performance. We then checked the scalability and external validity of our method by testing it on a healthcare problem of modeling per-visit patient costs using incomplete EHR data.

3.4.1. Predicting hourly demand of rental bikes in a bike sharing system

A bike-sharing system is a service in which bicycles are made available for shared use to individuals on a short-term basis for a price. A user rents out a bike from a dock and returns it to the same dock or a different one belonging to the bike sharing system. The total count of bikes rented across all the docks per hour indicates the (total) hourly bike demand in the system. Accurate predictions of hourly bike demand are critical for the design, operations, and expansion of bike sharing systems. Prediction performance of models are shown to improve significantly through meaningful inclusion of external sources of data (Ram et al. 2015; Srinivasan et al. 2018; Zheng 2015). Prediction of hourly bike demand can be improved by including information from external sources such as weather conditions, traffic conditions, local bus routes, etc. However, integration of such heterogeneous data sources can cause block-wise missing patterns in the resulting dataset. BRM is therefore a suitable method for solving the problem of predicting hourly bike demand in a bike-sharing system.

We use the historical usage data of Capital Bike Sharing (CBS) system (Fanaee-T and Gama 2013) to evaluate the performance of BRM. The dataset contains hourly counts of bikes rented during 2011-2012 in Washington D.C., USA. We selected the CBS dataset in this study for a number of reasons. The size of the CBS dataset (17389 observations, 10 features) is not too small to be trivial, and not too large so that we could not evaluate the quality of our approach by comparing it against more expensive imputation methods. The CBS dataset has no missing values – a feature we sought since we needed to measure the impact of missing values and imputation. We simulated random missing values for selected features to evaluate our method. To create block-

wise missing patterns similar to what may be seen in a multi-source integration scenario, we sampled 25%, 50%, and 75% of observations from features {hours after midnight, weather situation, holiday} and {Temperature, humidity, wind speed} independently and replaced them with NULL values. We also randomly sampled and deleted 5% of the observations from these features to add a stochastic element to the missing data. Table 3.3 shows the list of features and proportion of missing values for four versions of the dataset (with <5%, 25%, 50%, 75% missing values). We also introduced curvilinear effects (second order for numerical features) to improve overall fit for models.

Table 3.3: List of features and simulated missing values in the bike-sharing dataset

Feature	**Simulation 1**	**Simulation 2**	**Simulation 3**
	Missing ~ 25 %	**Missing ~ 50 %**	**Missing ~75 %**
Date based features (season, month, day)	0	0	0
Hours after midnight	28.76	52.53	76.24
Weather situation	28.84	52.59	76.32
Temperature	28.68	52.45	76.25
Humidity	28.68	52.5	76.21
Wind speed	28.67	52.47	76.25
Time of day	4.99	4.99	4.99

17389 observations for 2011-12. The outcome feature is count/hr. After initial inspection of data, the feature time of day was added as a derived feature to ensure a better fit without feature transformation

Four non-overlapping subsets were detected using k-means clustering algorithm on the missing values indicator matrix of the training data. Four overlapping subsets that contain only populated values were generated using the non-overlapping subsets. Features for the four subsets were S'_1 {Time of day, hours after midnight, Season, Month, Day of week, Holiday, wind speed}, S'_2 {Time of day, hours after midnight, Season, Month, Day of week, Holiday, wind speed, Temperature, Weather situation, Humidity}, S'_3 {Time of day, hours after midnight, Season, Month, Day of week, Temperature, Weather situation, Humidity} and S'_4 {Time of day, hours after

midnight, Season, Month, Day of week}. Candidate reduced models (model 1, model 2, model 3, model 4) were trained on these subsets.

3.4.1.1. Evaluating predictive performance

We evaluated and compared the prediction performance of models trained using BRM with other extant methods. For comparison with BRM, we trained models using following alternative methods:

1. **SVI**: Models developed using a complete dataset prepared through the single value imputation treatment using random forests (Stekhoven and Buhlmann 2012).
2. **CART**: Classification and regression tree as a representative of models that can implicitly handle missing values (Breiman et al. 1984).
3. **ENSEMBLE**: Models developed using the ensemble-based method (Srinivasan, Currim, et al. 2016).
4. **MI**: Models developed using datasets prepared through multiple imputation treatment combined with meta learners (Melville and McQuaid 2012).

The outcome for prediction is hourly bike demand measures in counts/hour in the bike-sharing dataset. An 80:20 split was used for training and test samples for all the predictive models. BRM is equally applicable for explanatory and predictive models. Therefore, we evaluated using linear regression as well as gradient boosting machine (GBM) as the underlying model families for the method comparison (except for CART which is an independent tree-based model). The stepwise method using AIC was used for model selection in linear regression (Neter et al. 1999). The tuning parameters used for GBM were shrinkage = 0.1, interaction = 3, minobsinnode = 10 and number of trees = 500, determined using grid search with cross-validation (Friedman et al. 2001). The stepwise linear regression models are termed as regression models, and GBM models are termed as tree-based models in the rest of the paper for the reader's convenience. Since hourly

bike demand represents count data, we also evaluated the fit of regression models with Poisson errors since count values are known to have a Poisson distribution. Nevertheless, the model fit improvement of Poisson regression over regression models with normal errors was negligible. Therefore, we trained regression models with normal errors for the bike sharing dataset.

We measured the prediction performance of the models using Root mean square error (RMSE) and Mean absolute error (MAE). RMSE and MAE for hourly bike demand y in T test instances can be computed as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}\left(y_t^{actual} - y_t^{predicted}\right)^2}{T}}$$

$$MAE = \sqrt{\frac{\sum_{t=1}^{T}\left|y_t^{actual} - y_t^{predicted}\right|}{T}}$$

The prediction accuracy in terms of RMSE and MAE for our test dataset are shown in Figure 3.5. The figure shows a grouped bar chart comparing the predictive performance using RMSE and MAE (counts/hour) compared for four versions of the Bike sharing data (<5%, 25%, 50%, 75% missing values).

In the absence of any predictive model, the mean value of the outcome in the training dataset is the best guess for the outcome in test instances and is used as a baseline for comparison. Just using the mean value of hourly bike demand in training data as predictions for hourly bike demand in test instances, we get RMSE and MAE as 183.13 counts/hour and 143.26 counts/hour respectively, values much higher than error values of all predictive models shown in Figure 3.5. When fitting regression models, model performance using BRM is equal to or better than MI, ENSEMBLE, and SVI. Model performance using BRM is superior to model performance using all other methods for tree-based models. Single value imputation using Random Forest (SVI) consistently underperforms for regression models as well as tree-based models because of over-

fitting that occurs due to excessive imputation of training data. As the proportion of missing values increases, the predictive performance of MI and ENSEMBLE deteriorates faster for tree-based models as compared to the performance for regression models. BRM consistently outperforms other incomplete data processing methods for regression models as well as tree-based models, trained over datasets with different proportions of missing values.

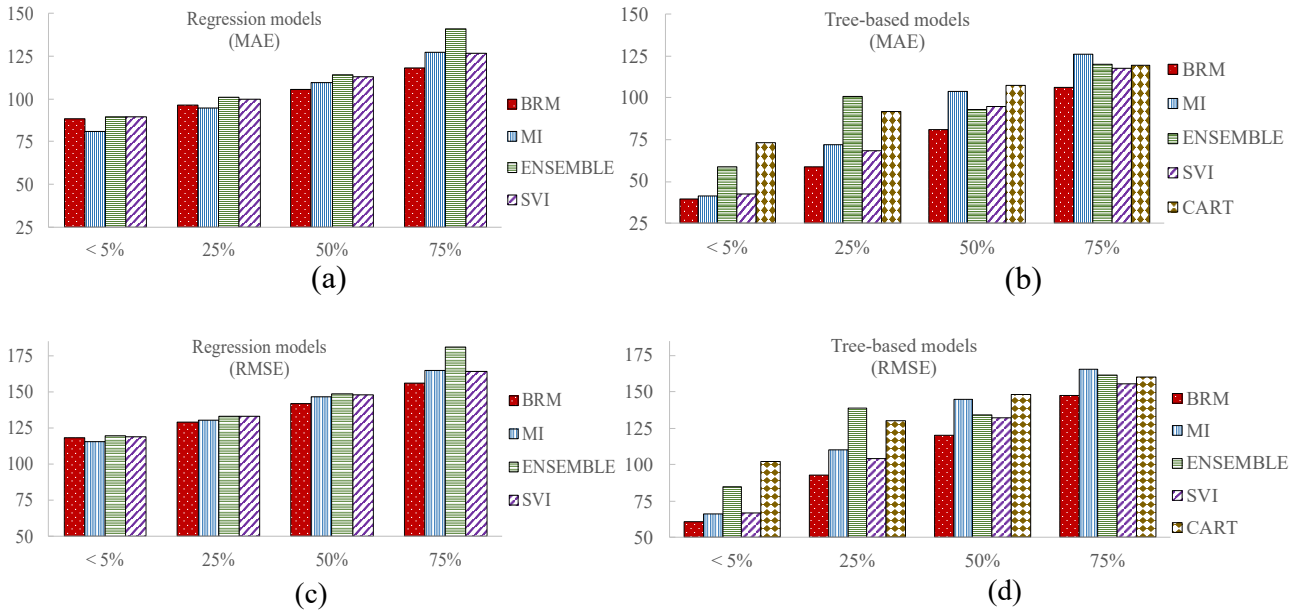


Figure 3.5: Performance comparison using MAE and RMSE metrics for Stepwise and Tree-based models

We measure the improvement in predictive performance of BRM with respect to the other methods using percentage change in RMSE and MAE as $\Delta RMSE_x = 100 \times (RMSE_{Reduced} - RMSE_x)/RMSE_x$ and $\Delta MAE_x = 100 \times (MAE_{Reduced} - MAE_x)/MAE_x$. Table 3.4 and Table 3.5 show the improvement in predictive performance on using BRM over other methods. For example, the third row and first column in Table 3.5 shows that the RMSE of BRM is 17.22 % lower than the RMSE of MI method for dataset with 50% missing values.

Table 3.4 shows that the percentage differences for MI are negative for the simulated datasets with < 5% missing values (also negative at the < 25% level using MAE), indicating that it performs equal to or better than BRM for data with smaller proportions of missing values for

regression modeling. For simulated datasets with 50% or 75% missing values, our method performs better than all other methods for regression modeling by at least 3.5% and up to 16.2%. Further, as we see in Section 3.4.1.3, it is not feasible to use MI for larger datasets. From Table 3.5, we see that our method outperforms other methods including MI for tree-based modeling across levels of missing values.

Table 3.4: Improvement in predictive performance when using BRM as compared to other methods for regression modeling

% Missing values	**MI**		**ENSEMBLE**		**SVI**	
	ΔRMSE	ΔMAE	ΔRMSE	ΔMAE	ΔRMSE	ΔMAE
< 5%	-2.63	-9.72	1.00	0.96	0.63	0.90
25%	1.05	-1.88	2.92	4.51	3.04	3.54
50%	3.56	3.79	4.63	7.44	4.17	6.37
75%	5.36	7.34	13.61	16.24	4.72	6.71

Table 3.5: Improvement in predictive performance when using BRM as compared to other methods for tree-based modeling

% Missing values	**MI**		**ENSEMBLE**		**SVI**		**CART**	
	ΔRMSE	ΔMAE	ΔRMSE	ΔMAE	ΔRMSE	ΔMAE	ΔRMSE	ΔMAE
< 5%	7.71	4.60	28.09	33.05	8.84	7.72	40.05	45.97
25%	15.22	18.31	32.71	41.69	10.67	13.75	28.40	35.87
50%	17.22	22.23	10.39	13.18	9.13	14.85	19.04	24.71
75%	10.69	15.84	8.46	11.52	4.96	9.90	7.98	11.28

Overall, Table 3.4 and Table 3.5 show that the BRM performs much better than existing methods in tackling missing data. In terms of predictive performance, we see gains ranging from 4.6% to 45.9%. As the percentage of missing values increases from 5% to 50%, our method outperforms existing methods by increasing amounts. However, once we reach the level of 75% (or more) missing values, while our method continues to outperform existing methods, the strength of the difference decreases since models in general have poor performance with such a high proportion of missing values.

3.4.1.2. Interpreting regression model coefficients and feature importance

We present the coefficients of regression models trained using BRM to model hourly bike demand in the bike sharing dataset simulated with 50% missing values. The global coefficient score for each input feature is computed using equation 1.1. A regression model trained over the bike sharing dataset with no missing values (i.e., complete data) is used as a benchmark for comparison. Table 3.6 shows a multi-column representation of candidate model coefficients for the bike sharing dataset simulated with 50% missing values, global coefficient scores and model coefficients for complete data.

Table 3.6: Summary of candidate reduced regression models and model with complete data along with global coefficient scores

Input features		**Model for subset 1**	**Model for subset 2**	**Model for subset 3**	**Model for subset 4**	**Global co-efficient scores**	**Model for complete data**
Intercept		247.03	226.76	199.44	280.15	239.90	150.82
Hour of day							-1.94
Time of day	*Early morning*	-176.80	-183.80	-178.96	-183.29	-180.70	-147.89
	Morning	40.05			47.76	44.51	40.57
	Mid-day	56.45	64.76	59.56	45.83	55.63	58.86
	Evening	188.96	164.78	158.12	190.74	176.25	217.39
	Late-evening	*baseline*	*baseline*	*baseline*	*baseline*		*baseline*
Day of the week	*Sunday*	-12.19		-15.99	-13.45	-13.99	-16.62
	Monday						
	Tuesday						
	Wednesday						
	Thursday						
	Friday						
	Saturday	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
Month				-1.46	-1.21	-1.33	
Season	*Spring*	37.48	-43.42	-44.49	33.51	-2.06	-46.10
	Summer	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
	Fall		-25.99	-32.15		-29.71	-27.39
	Winter	-89.45	-65.55	-77.62	-100.68	-85.09	-64.48
Holiday		-38.70	-24.01	*NA*	*NA*	-31.07	-29.29
Weather	*Mist + Few*	*NA*			*NA*		

situation	*clouds*						
	Light Rain, Light snow	*NA*	-49.89	-29.83	*NA*	-37.77	-57.90
	Heavy Rain + Thunderstorm	*NA*			*NA*		
	Clear - Partly cloudy	*NA*	*baseline*	*baseline*	*NA*	*baseline*	*baseline*
Temperature		*NA*	155.41	261.66	*NA*	219.62	268.97
Temperature (curvilinear)		*NA*			*NA*		
Humidity		*NA*	-141.11	-69.96	*NA*	-108.79	-120.82
Humidity (curvilinear)		*NA*		-155.62	*NA*	-155.62	-76.84
Wind speed		99.76	114.23	*NA*	*NA*	106.33	80.06
Wind speed (curvilinear)		-215.54	-254.20	*NA*	*NA*	-233.10	-220.48
Dataset size	No. of observations	8298	4123	8640	17389		17389
	Support (%)	47.72	23.71	49.69	100		-
Model fit	R-squared	0.5286	0.6000	0.5800	0.5235		0.5980

NA = Not applicable

We make the following inferences using coefficients of candidate models and global coefficient scores shown in Table 3.6, while showing the coefficients of the model for complete data for reference. The coefficient for the spring season is positive in the models for subsets 1 and 4 (with weather conditions) but negative in models for subsets 2 and subsets 3 (without weather conditions). This indicates a moderating effect of weather conditions (i.e., weather situation, temperature, humidity) on seasonal bike rental demand. On comparing coefficients of temperature and humidity in the model for subset 2 (with wind speed and holiday) and the model for subset 3 (without wind speed and holiday), we observe that the effects of temperature and humidity on bike demand is moderated by wind speed and holiday. We included respective two-way interactions in the model with complete data and found all interaction coefficients to be significant. Hourly bike demand is low during early mornings (12:00am – 5:00am) and late evenings (7:00pm – 11:59pm), and maximum during evening time (4:00pm – 7:00pm) with demand more than late evenings by

180 counts/hour. All weekdays have similar demand as that of Saturday, but demand drops by approximately 15 counts/hour on Sundays. Bike demand is highest in summer and lowest in winter (80 counts/hour less than summer). Light rain or light snow reduces bike demand by more than 30 counts/hour. Bike demand linearly increases by 220 counts/hour for every 1^{o}C increase in temperature. Humidity and windspeed have curvilinear relationships with hourly bike demand. The regression model with complete data has slightly higher values than the global coefficient scores due to the censoring of 50% values of features in the simulated dataset. In the absence of the BRM method, our inferences would be biased due to exclusion of rows/columns using pruning approaches or distortion of data due to treatment approaches as discussed in Section 3.2. For example, the simple pruning approach commonly adopted by researchers to eliminate missing values, would have led us to conclude that the day of the week had no relationship with bike demand, whereas in reality the demand for bikes is lower on Sundays compared to other days.

Figure 3.6 depicts the feature importance scores normalized to a total of 100 as pie-charts. Features that have less than a 5% importance score have been cumulatively labeled as “others” for ease of interpretation. *Time of day* and *Hours after midnight* are the two most important predictors for predicting bike demand.

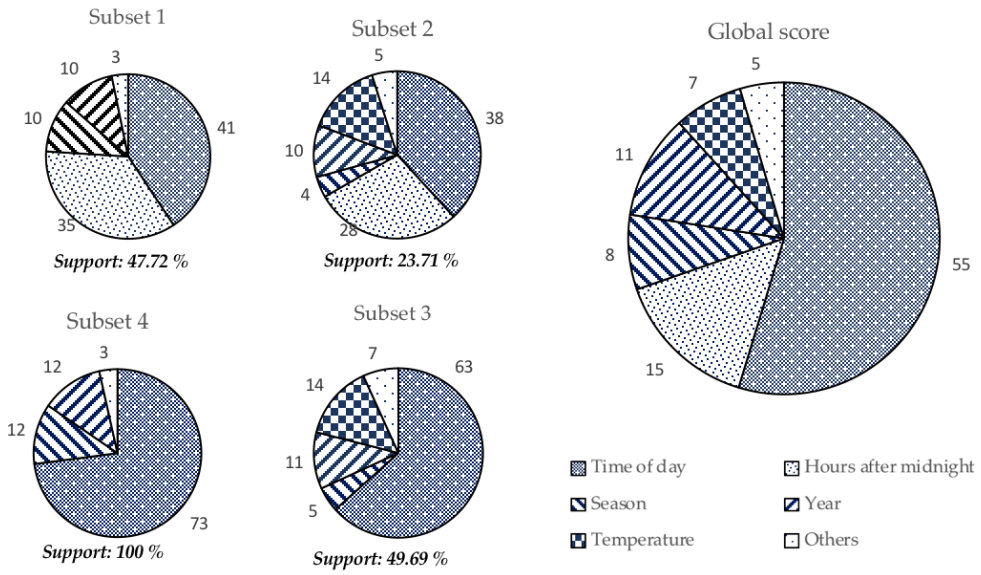


Figure 3.6: Comparison of candidate reduced tree-based models: 50% simulated missing values in dataset

3.4.1.3. Scalability analysis

We evaluate the scalability of BRM in terms of data size as well as number of missing block-patterns, by applying it to versions of the bike sharing dataset with simulated missing values. Running times of alternative methods (i.e., SVI, CART, ENSEMBLE, MI) are also compared, with regression as the underlying model fitting method for BRM, SVI, ENSEMBLE, and MI. All models were run using the R software package on a 16GB RAM, 2.50GHz processor machine. As shown in Figure 3.7, CART models had the lowest running times for all dataset sizes, and it took just 59 seconds to fit a model for 1 million data points. Models with BRM and ENSEMBLE were trained within 300 seconds for a million data points. On the other hand, SVI and MI did not converge beyond 100,000 data points due to memory overflow and intractable iteration times respectively.

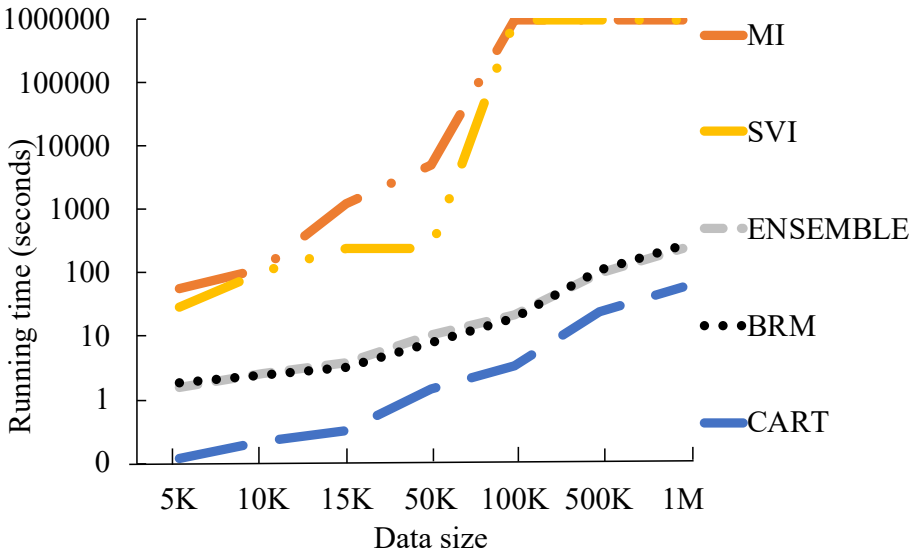


Figure 3.7: Scalability of different block-wise missing value handling methods

Table 3.7 shows the ratio of training time for regression models using existing methods (except CART which is an independent model) to the training time for regression models using BRM method. CART takes about one-fifth of the time as BRM across different sizes of data. ENSEMBLE methods have comparable speeds. However, the running time for SVI is an order of magnitude longer and two orders of magnitude longer for MI than BRM for a dataset with 15,000 points. Based on the results, we argue that BRM is suitable for both smaller datasets and larger datasets with over a million observations.

Table 3.7: Speed of other methods relative to the speed of BRM

No. of rows in dataset	**MI**	**ENSEMBLE**	**SVI**	**CART**
5,000	32.28	0.86	15.85	0.07
10,000	49.29	1.04	44.02	0.09
15,000	389.02	1.15	77.03	0.1
50,000	649.38	1.35	31.53	0.19
100,000	Did not converge	1.1	Did not converge	0.17
500,000	Did not converge	0.91	Did not converge	0.23
1,000,000	Did not converge	0.92	Did not converge	0.22

Figure 3.8 shows running time for regression models using BRM as a function of number

of block-wise missing patterns simulated in the bike sharing dataset. Run time increases from 2.5 seconds for fitting data with 2 block-wise missing patterns to 55.9 seconds for data with 256 block-wise missing patterns. It shows us that BRM can handle a range of block-wise missing patterns within a reasonable amount of time using standard or commodity hardware.

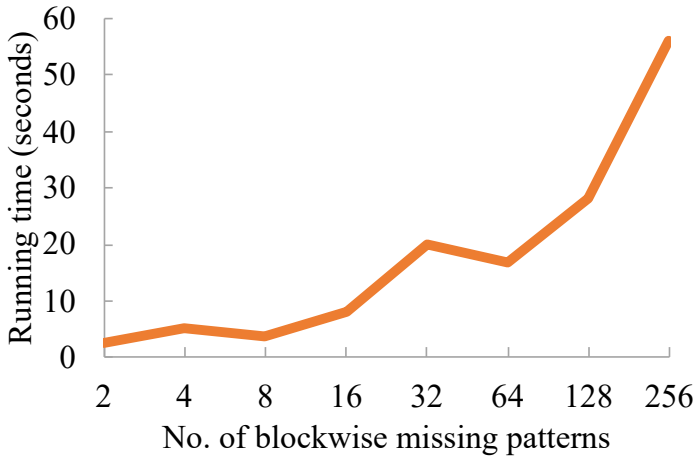


Figure 3.8: Running times of BRM based regression models as a function of number of block-wise missing patterns

3.4.2. Healthcare Application: Determining factors related to patient visit costs

Next, we discuss an application in the healthcare domain to illustrate the scalability and cross-domain suitability of our approach. In healthcare, we often run into the issue of missing values because parts of the data are collected separately, or values are missed during data entry. Furthermore, we often have a problem of a minority class (e.g., patients with specific conditions) or a limited number of observations where discarding data points is not advisable. If we were to prune records or only look at subsets of the data, we may end up with non-representative data or inaccurate results. Understanding factors affecting overall cost of patient visits is an important challenge in healthcare. We model the relationship between visit-related factors and cost per patient visit using linear regression. Modeling visit costs using large-scale electronic health records (EHR) can be useful to understand the factors related to treatment costs across different diagnoses, procedures, and patient populations. Current studies (Bertsimas et al. 2008; Chechulin et al. 2014;

Sushmita et al. 2015) analyze effects of different features on treatment cost and ways to improve predictive performance of models. Very large datasets are available for training the models, and both predictive performance and external validity are known to improve with scale. Our dataset includes 13 million patient visit records from the state of Arizona. This contains unadjusted costs per visit as well as other fields such as age at admission, severity of illness, and payer information. Since patient history has been identified as an important factor influencing treatment costs (Sushmita et al. 2015), we linked the re-admissions table with the patient visits table to get previous costs and time to readmission. The input features used in this study and their corresponding percentage of missing values is given in Table 3.8.

Table 3.8: List of features and missing values in the healthcare cost dataset

Features	**Details**	**% Missing**
Sex, Year, Quarter	Demography and admission year	0
Age at admission	Age in years (categorized to 7 levels)	0.06
Severity of illness	(Score from 0-4)	0
Risk of mortality	(Score from 0-4)	0
Previous charges	Cost of patient during the last visit if recorded (categorized to 4 levels)	87.2
Time to readmission	Days since previous discharge (categorized to 3 levels)	87.17
Payer information	Categorical feature with 5 levels (top 5 payment sources in dataset)	29.33

Total no. of observations is 13,020,716 for the 2011-14 timeframe and the outcome variable is Cost of individual patient visits.

We applied BRM for training regression models to explain the relationship between input factors and patient visit costs. Four non-overlapping subsets were detected using the *k*-means

clustering algorithm on the missing values indicator matrix of the healthcare cost dataset. Using the non-overlapping subsets and their partial order relationships, four overlapping subsets that contain only populated values were created for fitting candidate reduced models. The features for the four subsets were S'_1 {*year*, *quarter*, *sex*, *age*, *severity of illness*, *risk of mortality*, *previous charge*, *days to readmit*, *payer*}, S'_2 {*year*, *quarter*, *sex*, *age*, *severity of illness*, *risk of mortality*, *payer*}, S'_3 {*year*, *quarter*, *sex*, *age*, *severity of illness*, *risk of mortality*, *previous charge*, *days to readmit*} and S'_4 {*year*, *quarter*, *sex*, *age*, *severity of illness*, *risk of mortality*}. We discretized numerical features except *severity of illness* and *risk of mortality* to make it easier to interpret the differences across coefficients. Analysis was performed using Microsoft R version 8.0 (Rickert 2011) on a 16 GB RAM, 2.50 GHz processor machine.

Table 3.9 shows a multi-column representation of candidate reduced regression model coefficients and global coefficient scores for the healthcare cost dataset. We label the corresponding candidate regression models (Model 1, Model 2, Model 3, Model 4) in Table 3.9 as *Payer Re-admission known*, *Payer unknown*, *Re-admission unknown*, and *Payer Re-admission unknown* for the reader's convenience.

Table 3.9: Global coefficients scores and coefficients of the candidate regression models trained using BRM for healthcare cost dataset

Input features		**Model 1** ***(Payer Re-admission known)***	**Model 2 (*Re-admission unknown)***	**Model 3** ***(Payer unknown)***	**Model 4** ***(Payer Re-admission unknown)***	**Global co-efficient score**
Intercept		3631.78	5934.45	3089.92	5224.48	4952.52
Severity of illness		2248.77	1817.95	2240.19	1792.81	1927.41
Risk of mortality		401.30	433.59	351.63	446.78	422.32
Sex	*Female*	163.47	653.02	169.94	652.86	517.51
	Male	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
Year	*2011*	-969.56	-1059.36	-972.70	-1050.91	-1031.66

	2012	-666.44	-779.52	-663.29	-773.55	-745.35
	2013	-400.39	-389.06	-401.23	-376.11	-387.54
	2014	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
Quarter	*One*	-251.14	-210.13	-240.56	-220.71	-223.95
	Two	-279.25	-268.24	-273.90	-266.09	-269.74
	Three	-138.25	-156.51	-123.14	-161.66	-151.15
	Four	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
Age group	*New-Born*	-1036.67	-3817.58	-995.99	-3974.95	-3095.34
	Pre-teen	-574.78	-548.47	-462.34	-590.86	-555.29
	Teenage	143.19	285.63	281.94	289.57	267.43
	Mature-Adult	277.50	2197.22	379.58	2138.35	1655.32
	Senior-below 80	683.24	3019.12	717.32	2854.52	2312.27
	Senior-above 80	-209.22	1053.67	-280.78	798.75	597.01
	Young-Adult	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
Previous charges	*Below 10000 $*	1390.69	*NA*	1420.15	*NA*	1405.95
	Below 50000 $	3454.16	*NA*	3480.28	*NA*	3467.69
	Above 50000 $	4752.61	*NA*	4790.36	*NA*	4772.17
	Below 5000 $	*baseline*	*NA*	*baseline*	*NA*	*baseline*
Days to readmit	*Within one week*	393.80	*NA*	341.40	*NA*	366.66
	Within two weeks		*NA*		*NA*	
	Within a month	*baseline*	*NA*	*baseline*	*NA*	*Baseline*
Payer	*Medicare*	-914.82	-1125.27	*NA*	*NA*	-1067.65
	Medicaid	-759.10	-1185.28	*NA*	*NA*	-1066.27
	Private funding HMO	-640.29	-976.08	*NA*	*NA*	-883.10
	Other	-475.45	-672.32	*NA*	*NA*	-617.81
	Self-pay	*baseline*	*baseline*	*NA*	*NA*	*Baseline*
Dataset size	*No. of observations*	1211761	9195229	1665215	13012378	
	Support (%)	9.31	70.67	12.80	100.00	
Model fit	*R-squared*	0.1239	0.1526	0.1213	0.1442	

NA = Not applicable

Table 3.9 shows that the coefficient of determination for all four models is between 0.12 and 0.16, suggesting that the fit of all the models is equivalent. Model 1 *(Payer Re-admission known)* and Model 3 *(Payer unknown)* indicate that readmission within one-week increases visit costs by 7.5%. Also, readmitted patients with previous visit costs in the range of $[5,000 - 10,000],

$[10,000 - 50,000], $[> 50,000] have higher visit costs by approximately $1,400, $3,500, and $4,800 respectively than visit costs of first-time patients. Model 2 *(Re-admission unknown)* indicates that patients supported by primary payers private funding HMO, Medicaid, Medicare have $900 to $1,100 lower visit costs compared to those recorded as self-pay. However, model 1 *(Payer Re-admission known)* shows that the difference between visit costs of self-paying patients to those supported by HMO, Medicaid, Medicare reduces to $650 to $900 when observations with readmission related information are included. Coefficients of models that include readmission related information (Model 1 *(Payer Re-admission known)* and Model 3 *(Payer unknown)*) are higher than coefficients of models that exclude readmission related information (Model 2 *(Re-admission unknown)* and Model 4 *(Payer Re-admission unknown)*) for patients belonging to age groups: teenagers (13-19 years), mature adults (40-60 years), seniors less than 80 (60-80 years), and seniors older than 80. This indicates that on an average, readmitted patients in these age groups pay more per visit when compared to first time patients belonging to the corresponding age groups. The global coefficient scores for age-groups show that visits for patients belonging to new-born and pre-teen (1-12 years) age group categories cost less than the visits of young adult patients by approximately $3100 and $600, respectively. On the other hand, visits for patients belonging to age groups of teenage (13-19 years), mature adults (40-60 years), seniors less than 80 (60-80 years), and seniors older than 80 years cost more than visits of young patients by approximately $300, $1,700, $2,300 and $600 respectively. The global coefficient scores for years 2011 to 2013 show that each year, patient visit costs increased by approximately 8.5%.

Previous studies (Bertsimas et al. 2008; Chechulin et al. 2014) show that age is positively correlated to patient visit costs. There is also evidence that previous visit costs are positively correlated to revisit costs (Sushmita et al. 2015). Gender, severity of illness, and risk of mortality have also been found to be related to patient visit costs (Gregori et al. 2011). In this study, we

validate the findings from previous studies using a bigger dataset. We extend existing knowledge by including input features that have not been included in a same explanatory model such as payer, previous costs, and severity of illness. Using our method, we are able to train multiple models and compare their coefficients for different combinations of input features. We identified that including patients with readmission in the data moderates the effect of other features in our dataset such as age, payer, gender, severity of illness, and risk of mortality.

Models trained using the alternative methods of MI, SVI and CART do not converge for this type of incomplete data. ENSEMBLE uses a non-overlapping method for creating subsets and hence, will always have less than or equal to the number of observations available for training candidate models using BRM method. Without our method, the naïve approach of pruning rows or columns is most intuitive, leading to modeling of one of the four subsets identified above. Since 87.20% of the 13 million patient records do not have readmission related information, inference about the effects of readmissions can be done using subset 3, which has approximately 1.6 million observations. Similarly, inferences about the effects of payer information in the dataset can be made using subset 2, which has approximately 9.2 million observations. Simultaneous modeling of partial effects of readmissions and payer information can be done using subset 1, which has approximately 1.2 million observations. Using just one of the subsets may induce biased inferences since coefficients of features in a model trained on one subset could be substantially different from the coefficients of corresponding features in the model trained on another subset. For example, a model trained on the subsets of data excluding readmission information (incorrectly) suggests that mature-adults have visit costs $2,100 more than young-adult patients (the baseline), whereas including the readmission information in training dataset reduces the difference to approximately $300, as shown in Table 3.9. It is also problematic to manually select two or more subsets to make collective inferences based on visual inspection of data, since some subsets could be

unintentionally excluded. BRM automatically indicates all the possible subsets of the data with block-wise missing patterns. We no longer need to analyze subsets of the dataset in isolation because the incomplete dataset can be modeled and interpreted simultaneously.

There is a need for accurate modeling of patient visit cost data for population management (Gregori et al. 2011). Identification of potential high-cost patients can support personalized case management programs to address specific patient needs (Bates et al. 2014; Srinivasan et al. 2018). Imputation based methods are computationally intensive and often do not converge for larger datasets. Pruning columns and rows can result in arbitrary choice of training datasets for models. The BRM method is suitable for modeling patient visit data with block-wise missing patterns. It is not only fast and robust, but it also provides a compilation of models corresponding to different missing value patterns to draw inferences. Our method involves minimal synthetic modification to original data through imputation as well as avoids discarding important information via pruning. As a result, our method is a valuable tool to model applications similar to patient visit cost modeling with incomplete data containing block-wise missing patterns.

3.5. Discussion

This study follows the paradigm of design science research by developing an IT artifact that addresses the problems associated with analyzing incomplete data. It provides guidance on how to analyze incomplete data in cases where incomplete data can neither be discarded nor imputed as a pre-processing step before data analysis. There are several examples when we encounter the incomplete data problem, including flat tables obtained using multiple outer joins, participant non-responses in surveys, asynchronous data capture from Internet of Things (IoT) sensors, partial destruction of data, and loss of data during transmission.

Single value imputation and multiple imputation methods have good predictive performance, but they can be difficult to use as the scale of data increases. Ensemble based

methods combine predictions from multiple models and use imputation to process missing values in test instances. However, to optimize the modeling of block-wise missing patterns, it is important to devise a method that minimizes imputation at the training as well as prediction phase. Therefore, we propose BRM to minimize imputation at both the training and prediction phases. The systematic procedure of identification of overlapping subsets that contain only populated values ensures minimal imputation in the training phase. In the prediction phase, our method does not require the test instances to be modified to suit the models. Instead, a suitable model is mapped to each test instance. Consequently, for batches of test instances with missing values for different sets of input features, our method performs seamlessly and efficiently. The global coefficient score assignment is useful for making inferences across all subsets of the incomplete data. With a large number of missing values, drawing inferences only based on coefficients of a single model can be misleading. Therefore, we propose a multiple column representation of results from multiple models and a set of global coefficient scores to draw inferences in this study.

We simulated missing values using the CBS bike-sharing dataset and evaluated the prediction performance of our method. As the proportion of missing values increases, our method significantly outperforms extant methods. We have demonstrated ease of interpretation for the user, using four subsets in this study; however, our method can be used with any number of block-wise missing patterns. We trained regression models using our method and inferred that bike demand is affected by seasonal patterns, weather conditions and wind speed. Even with 50% of feature values missing in a simulated version of the CBS dataset, we were able to determine effects consistent with those found in a model trained over complete data. In addition, by comparing coefficients of the candidate models, we identified a moderating effect of weather conditions on seasonal hourly bike demand as well as a moderating effect of wind speed on the relationships of temperature and humidity with hourly bike demand.

Models trained using BRM for the healthcare patient visit cost application were useful for making multiple inferences. Readmitted patients belonging to the age groups – teenage (13-19 years), mature adults (40-60 years), seniors less than 80 (60-80 years) and seniors older than 80 – pay more per visit on an average when compared to first time patients belonging to the corresponding age groups. Previous visit cost, time to readmission as well as type of payer (e.g., Medicaid, Medicare, self-pay) have an effect on patient visit cost. The results from the application of BRM to the healthcare cost data highlights the value of considering information from multiple sources (i.e., payer, readmissions, etc.) in patient visit cost modeling even if integration of multi-source data results in incomplete data. Global coefficient scores are useful for making collective inference. The difference in model coefficients across subsets indicates heterogeneity in visit characteristics across subsets.

Our current work has some limitations. While we have evaluated our method using algorithms suitable for the problems at hand, the space of known data mining algorithms is naturally very large, and we leave a more comprehensive testing and performance improvement analysis for future work. However, we anticipate that our method will support different algorithms for explanatory and predictive modeling applications (e.g., Bayesian models, deep learning models, semi-supervised models, etc.). In general, we believe that our proposed method will be useful for researchers as well as practitioners in tackling incomplete data. We anticipate that our method will facilitate analysis of incomplete datasets that often have a story to tell but have traditionally been ignored in the past.

3.6. Conclusion

Technology, algorithms, and applications are constantly evolving in the big data era, where new problems require novel solutions. Data integration, outlier detection, influence analysis, real-time prediction, and visualization are some of the sub-domains of data science that require

significantly different treatment in the case of big data. Handling missing values has been a topic of research for several decades and has gained importance with the advent of large volumes and heterogeneity inherent in big data, especially since a poorly chosen technique has the potential to distort the validity of subsequent data analysis.

In this study, we described how to handle incomplete data with block-wise missing patterns. We discussed the limitations of current strategies to deal with block-wise missing values. Thereafter, we introduced a new method to model block-wise missing data based on the idea of reduced modeling. However, the naïve reduced modeling approach suffered from the computational burden of building a very large number of models on training data at run-time based on test instances. Our method, BRM, limits the number of models to train by exploiting block-wise missing patterns in the dataset. We demonstrate the usefulness of our method by evaluating it using datasets from two different domains. Our method improves prediction performance for regression models over existing methods by 4% to 16% and improves prediction performance for tree-based data mining models by at least 5% and up to 50%. Our method can predict as well as provide overall feature scores even in the presence of a relatively large proportion of missing values.

BRM is applicable to explanatory as well as predictive models. It can be used as a framework to analyze incomplete data observed in operational datasets as well as those that require cross-domain integration of multiple datasets (e.g., data from disparate sources such as social media, weather, and the Internet of Things). It can be useful for big data applications that require use of massive yet incomplete datasets and for which imputation is not feasible. Our method is scalable to handle incomplete data in cases where extant single value imputation and multiple imputation methods did not converge. We believe our method is preferable to methods that handle missing values implicitly due to its simplicity and enhanced interpretability. Our method can be

used to model data with moderate to high proportion of block-wise missing values using common methods such as linear regression, classification trees, neural networks, and Bayesian models.

Future work can consider benchmarking and performance evaluation of BRM on datasets of different sizes, across domains, and using other modeling methods. Other improvements can include optimizing the underlying techniques in each phase. For example, future work can investigate the use of alternative clustering techniques for subset identification or different similarity indices for test instance mapping.

The phases in BRM can be executed in a sequential automated manner without requiring manual intervention, from data processing to creation of model results. We are currently developing a package in R for BRM to analyze incomplete data with block-wise missing patterns.

4. ESSAY 3: DETERMINING THE EFFECTS OF SOUND LEVELS ON PHYSIOLOGICAL WELLBEING IN THE WORKPLACE – A FIELD STUDY USING WEARABLE DEVICES

4.1. Introduction

Nearly 50 million workers in the United States spend over one-fifth of their day at their workplace (Bureau of Labor Statistics 2017). On an average, four out of ten US workers report their job and workplace to be stressful and negatively affecting their wellbeing (Harvard School of Public Health 2016). Workplace-related stress and absenteeism cost up to $225 billion, or more than 10% of office workers' contribution to the annual U.S. GDP (CDC Foundation 2018). Past studies have shown that the workplace environment is closely tied to an office worker's wellbeing markers including mental state, productivity, stress, and longevity (Heerwagen and Zagreus 2005; MacNaughton et al. 2016; Thayer et al. 2010). Among various workplace environment factors, workplace sound level[1] has been identified as a significant stressor for white-collared office workers (Frontczak et al. 2012; Seidman and Standring 2010). The psychosocial effects of high sound levels (i.e., loud noise) including its effects on satisfaction, environmental control, social interaction, social support, and perceived insensitivity to social cues in a workplace setting have been studied previously (Rashid and Zimring 2008). However, the underlying mechanism of sound effects on physical health and wellbeing is not yet fully understood (Kraus et al. 2013). A comprehensive understanding of the relationship between workplace sound levels and physiological wellbeing is important for making decisions towards minimizing work-related stress and promoting health and wellness among office-workers.

Mobile and sensor-based content are among the key characteristics of third generation Business Intelligence and Analytics (BI&A) applications (Chen et al. 2012). Wearable health

[1] Section 6.1 of supplementary materials presents a reference for subjective reference of different sound levels in real world

sensor devices or wearables offer the unique opportunity to observe mental and physiological changes in individuals through measurement of activity, heart rate, body temperature, and other health indicators (e.g., blood pressure, sleep quality, breathing rate, etc.). During the early adoption phase of wearable technology, research was primarily directed towards sensor development and architecture (Malhi et al. 2012; Yamada and Lopez 2012). With more and more commercial products being introduced into the market today, research focus is shifting on wearable data analytics and associated design science research applications (Ravi et al. 2017; Sano et al. 2018; J. B. Wang et al. 2015). Wearables-based studies have primarily focused on predictive modeling, with research applications such as ambient assisted living (Rashidi and Mihailidis 2013), human activity recognition (Zhu et al. 2018), reality mining (Pentland et al. 2009), and sports management (Guillén et al. 2011). Recent research applications have also used wearables data for identifying associative/causal patterns using statistical modeling methods (MacNaughton et al. 2016; Thayer et al. 2010; J. B. Wang et al. 2015). Wearable data analytics is a promising area that solicits attention from IS researchers owing to the ubiquitous nature of wearables in today's lifestyle, and the promise of wearables to generate rich, personalized, temporal, and highly-grained information content. In the context of modeling the effects of workplace environment and a worker's wellbeing, wearable data analytics can be used to measure and model real-time physiological responses (e.g., change in heart rate, breathing rate, etc.) occurring due to changes in indoor environments.

We conducted a field study using multiple wearable devices to determine the impact of indoor environment on individual wellbeing. In our study, participants carry out their day to day activities while wearing sensors that continuously record their (short-term) physiological wellbeing state and ambient environmental conditions, including sound level. We observed that the relationship between sound level and two physiological wellbeing measures (i.e., SDNN, normalized-HF) is curvilinear and varies across individuals. We propose new methods to address

challenges in representing curvilinear effects, modeling multiple outcomes simultaneously, and identifying factors contributing to heterogeneity in effects. The first method is a semi-automated method for change point determination in multilevel segmented regression models for representing curvilinear relationships. The second method is a Bayesian latent variable modeling method for simultaneous modeling of multiple outcomes. The third method tackles the problem of modeling individual heterogeneity in the sound-wellbeing relationship using a two-step approach: first, modeling the heterogeneity using a Bayesian varying coefficients models then identifying person-level factors that contribute to the heterogeneity using regularized models. We show that our proposed methods have better predictive performance than existing methods and are elemental in developing critical insights about the sound-wellbeing relationship in the workplace.

Our major findings on the workplace sound-wellbeing relationship are as follows. A sound level threshold value of 50 dBA averaged at 5-minute interval is optimal for physiological wellbeing. A 10 dBA increase in sound levels below 50 dBA is related to a 3.6% increase in physiological wellbeing, whereas a 10 dBA increase in sound levels above 50 dBA is related to decrease in physiological wellbeing by 1.3%. Blood pressure level and computer intensive work are two individual personality factors that moderate the relationship between sound levels and physiological wellbeing. People with higher blood pressure are negatively affected by increase in lower as well as higher sound levels. Computer intensive work is related to an amplification of positive effects of lower sound levels as well as negative effects of higher sound levels. Our findings show that quiet workstations are optimal for the physiological wellbeing of office-workers with high blood pressure, whereas workspaces with moderate sound levels (~50dBA) are suitable for workers with computer intensive work.

In this study, we make three major contributions. One, to our knowledge, our paper is the first to implement a field study using multiple wearable devices to model the environment-

wellbeing phenomenon. Our field study records environmental signals and physiological responses in real-time, while office workers' carry out their daily tasks, which overcomes external validity challenges faced by existing studies. Our field study design can be used as a reference by future design science research applications facing similar challenges. Two, our study proposes new quantitative methods that address three key challenges in modeling digital data generated by wearable devices (i.e., curvilinear modeling, simultaneous modeling, heterogeneity effects modeling). These methods address gaps in modeling methods for multilevel data. Lastly, our study unravels aspects of the sound-wellbeing relationship that were unknown earlier, informing workplace planning policies and practices that affect the health and wellbeing of office workers worldwide. Our conclusions have implications for indoor environment-wellbeing literature as well as healthy workplace design practices.

The rest of this paper is organized as follows. In Section 4.2, we introduce the study background and related literature. In Section 4.3 and Section 4.4, we describe our field study using wearable devices and the need for new modeling methods. In Section 4.5, we expound on the three new methods for sound-wellbeing modeling. Section 4.6 contains the evaluation of our methods and key-learning based on their application on our data, followed by the discussion and conclusions in Section 4.7 and Section 4.8 respectively.

4.2. Background and related literature

4.2.1. Physiological wellbeing

Psychological well-being consists of positive relationships with others, personal mastery, autonomy, a feeling of purpose and meaning in life, and personal growth and development (Ryff 1989). On the other hand, physiological wellbeing is associated with a dynamic, ever-adapting balance in the human physiological system conditioned by momentary demands (Malik et al.

1996). When we are in good health or at a higher physiological wellbeing state, we experience flexibility and resilience in relation to our environment and experiences.

Stress is a major factor that impacts physiological wellbeing (Boron and Boulpaep 2012). The physiological response to the demands put upon the body, also known as physiological stress response, has a direct relationship with the two components of the autonomic nervous system (ANS): sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS). When the body is stressed, the SNS generates the "fight or flight" response where the body shifts all of its energy resources toward fighting off a life threat. In contrast, the PNS indicates "rest and digest" and is involved with restoration, repair, nourishment and detoxification processes in the body. Heart rate variability (HRV) reflects the modulation in the normal rhythm of the heart and is known to assess overall cardiac health and the state of the ANS. HRV is considered as a proxy measure for the physiological wellbeing of a person, i.e., the higher its value, the higher the wellbeing (Xhyheri et al. 2012). The HRV is more widely used when compared to other physiological stress response measures such as salivary cortisol and skin conductance as it can be measured at short intervals using commercially available heart rate monitors, which are relatively less intrusive (Acharya et al. 2006; Xhyheri et al. 2012). While many measures of HRV exist, each serves as a slightly different lens in terms of viewing the body's physiological stress response (Shaffer and Ginsberg 2017). The mean of standard deviation for all successive R-R intervals (SDNN) is a global index of HRV and reflects longer term circulation differences or the overall activity in the ANS. The normalized high frequency component (normalized-HF) is the ratio between the absolute value of the High Frequency and the difference between Total Power and Very Low Frequency bands in the frequency domain power spectrum of heart rate that emphasizes changes in parasympathetic regulation. High values of SDNN and normalized-HF have consistently been

found to indicate better health and wellbeing (Soares-Miranda et al. 2014). Our study considers both SDNN and normalized-HF.

4.2.2. Workplace and wellbeing

White-collar workers in U.S. spend a majority of their active hours in a day in enclosed workplaces (Bureau of Labor Statistics 2017). The workplace environment not only affects people at work (Heerwagen and Zagreus 2005) but is also known to have a carry-over effect on their personal lives outside the office (Lindberg et al. 2018). Workplace characteristics consist of elements such as workstation design (i.e., workstation type, workstation area, furniture, nature view, etc.), indoor environment quality (i.e., ambient sound levels, temperature, humidity, air quality, etc.), social influence (i.e., interaction with colleagues in close vicinity, proximity with team members, tele-working facilities, etc.) and amenities (i.e., proximate breakout areas, availability of quiet spaces, control over thermostat and window-blinds, etc.). Indoor workplace environment quality is closely tied to mood, productivity, activity and longevity of office workers (Backé et al. 2012; Kivimäki et al. 2012). Air-quality factors such as carbon dioxide are shown to impair cognitive performance and degrade physiological wellbeing in the workplace (MacNaughton et al. 2016). Workstation type (e.g., open bench seating, cubicles, private space, etc.) and structure (e.g., length of passage, furniture, workstation area, etc.) also affect worker stress and physiological wellbeing (Lindberg et al. 2018; Thayer et al. 2010).

Our study is part of a multi-disciplinary research program titled Wellbuilt-for-Wellbeing (WB2) with a focus on understanding the relationship between workplace of white-collar office-workers, described in detail in section 4.3. As part of the program, we conducted research investigations on multiple issues. These include, proposing a new sensor-based sleep quality index (Lee et al. 2018), exploring the inter-relationship between mobile survey prompt inputs and wearables (Ghahramani et al. 2018), devising new data integration mechanisms (Srinivasan, Ram,

et al. 2016), and identifying lagged effects of independent variables on outcomes, all of which are measured via smart-wearable systems (Srinivasan et al. 2017). Impact of different aspects of indoor environment on wellbeing reported or currently being investigated include, sound level, workstation type (Lindberg et al. 2018), temperature, and relative humidity (Razjouyan et al. 2019). In terms of office worker activity and self-induced stress, Lindberg et al. (2018) report that activity level at work is related to work-related stress and also tied to average post-work stress levels of white-collar office- workers. Among all workplace-related factors, high ambient sound levels or noise in the work environment is reported to be one of the highest stressors for white-collared office workers in the US (Frontczak et al. 2012; Heerwagen and Zagreus 2005; Seidman and Standring 2010). Therefore, this study focuses on workplace sound level and its effects on worker's physiological wellbeing.

4.2.3. Workplace sound levels and wellbeing

Sources of sounds in offices include other people's conversations, telephone-calls, and mechanical equipment. Physical workstation design is tightly coupled with ambient sound level exposure; consequently, sound level is an important environmental factor to be considered in intelligent space planning and design (Kjellberg et al. 1996). There is substantial literature on the effects of sound in office settings on social, psychological, and performance-based wellbeing (Kjellberg et al. 1996; Rashid and Zimring 2008). Sound level amplitudes have been shown to not only affect our mood and productivity, but also affect our physiological state of wellbeing (Jahncke et al. 2011; Lee et al. 2010; Lusk et al. 2002). Understanding the effect of workplace sound levels on an individual's physiological wellbeing is important for making decisions towards minimizing fatigue in office workers and promoting their health and wellness. Table 4.1 lists the literature on effects of workplace sound levels on an individual's physiological wellbeing in chronological order.

Table 4.1: Literature on effect of workplace sound levels on physiological wellbeing

Study	Input	Outcome(s)	Study design	Findings
(Lusk et al. 2002)	Areas with sound levels averaged across a 5 years interval	Blood pressure and heart rate	N=374; Correlating person-level noise exposure with physiological wellbeing; Method: Linear regression	Areas with high sound levels are predictive of increase in blood pressure
(Lee et al. 2010)	Discrete sound levels	HRV (LF, LF/HF), Mean blood pressure, Mean heart rate	N=16; Treatment = Sound level exposure of No noise, 50 dBA, 60 dBA, 70 dBA and 80 dBA for 5 minutes with 2 minutes interval; Method: Repeated measures ANOVA; Spearman's Rho	HRV decreases with higher sound level exposures, but no change in blood pressure and mean heart rate
(Jahncke et al. 2011)	Noisy background, river sounds, nature movie	Cortisol, Catecholamines, self-rating of tiredness, mood	N=47; Treatment = Completed tasks for 2 hours each in a low and high noise conditions; Repeated measures ANOVA	Though noisy background and river sounds have an effect on psychological outcomes, they had not significant effect on physiological outcomes
(Kraus et al. 2013)	Sound levels	HRV (LF/HF, SDNN, RMSSD)	N=110; Prospective panel study with participants spending up to 7.5 hours in a room; Method = Additive mixed models	Sound levels have a positive effect below 65 dBA on SDNN, but is not significantly related to any of the other outcomes
(Sim et al. 2015)	Sound types, sound levels	HRV (SDNN, HF, LF/HF)	N=40; Treatment: 45 dBA exposure for 5 minutes; Method: Linear regression	Increase in sound level negatively affects physiological wellbeing. Sound types do not have a significant effect on physiological outcomes

(Walker et al. 2016)	Noise exposure at 75 dBA at low frequency and high-frequency	HRV (SDNN, LF, RMSSD), blood pressure, salivary cortisol, amylase	N=10; Treatment = 40 minutes noise exposure; Method=Multivariate multilevel regression	High sound levels at low-frequencies and high-frequencies have significant negative effect on HRV
(Park and Lee 2017)	Floor impact noises ranging from 31.5 dBA to 63 dBA	Noticeability, Annoyance, Heart rate, electrodermal activity, respiration rate	N=21; Treatment = 5 sessions of 15 minutes of different floor impact noises; Method=Repeated measures ANOVA	Annoyance, noticeability, electrodermal activity and respiration rate increases with sound level, but no significant change in heart rate. Physiological responses are not affected by noise source.
(Cvijanović et al. 2017)	Sound levels	Mental effort, HRV (LF, LF/HF) and skin conductance	N=40; Treatment = 6 dBA background noise added while participants completed collaborative tasks; Method=Multilevel regression	Though mental effort required increases with sound levels, effect on physiological wellbeing was not significant
(Srinivasan et al. 2017)	Sound levels, Temperature, CO_2, Humidity, Atmospheric pressure	HRV (SDNN, RMSSD, normalized HF, LF/HF)	N=231; Mixed lasso for identify length of cumulative lagged effect of inputs on outcomes	Sound level has an instantaneous effect on HRV whereas other environment factors have a lagged effect of one hour

From Table 4.1, we see that the majority of past studies employed controlled experiments to determine the causal effect of different sound level exposures on an individual's physiological wellbeing. While some studies revealed a negative relationship between high sound levels (i.e., noise) and physiological wellbeing measures (Lee et al. 2010; Lusk et al. 2002; Walker et al. 2016),

other studies reported inconclusive results (Cvijanović et al. 2017; Jahncke et al. 2011). Noise source and noise types were observed to not have a significant effect on physiological wellbeing (Park and Lee 2017; Sun Sim et al. 2015). Also, the effect of sound level on physiological outcomes, if present, were observed to be consistent for low as well as high sound frequencies (Walker et al. 2016). Recent studies identified that the effects of sound levels on physiological wellbeing are non-monotonic (Kraus et al. 2013), as well as instantaneous (Srinivasan et al. 2017).

Existing studies suffer from four major limitations. First, the majority of the studies in the past employed experiments with a limited set of treatments, few subjects, and limited control variables such that the results cannot be easily generalizable to the real office workplaces. Second, studies report results from multiple models corresponding to different measures of physiological wellbeing (Cvijanović et al. 2017; Park and Lee 2017; Sim et al. 2015) making it difficult to derive insights and take actions. Third, even though it has been established that the sound-wellbeing relationship is not linear (Kraus et al. 2013), the nature of its relationship is still not clear. That is, how physiological wellbeing varies as a function of workplace sound levels is not yet modeled. Learning the optimal sound levels for individual physiological wellbeing can be useful for reducing the overall fatigue of office workers through intelligent workplace design. Lastly, the heterogeneity in sound effects on physiological wellbeing (i.e., how the effects of sound level varies across individuals) is important for personalized space planning but has not been studied previously. To summarize, previous studies employed controlled experiments with few subjects and standard modeling methods (e.g., analysis of variance, linear regression, etc.); as a result, they suffer from low external validity and are unable to capture different aspects of the sound-wellbeing relationship. Therefore, we introduce a novel observational study design using wearable devices and propose new methods to overcome the limitations of existing workplace sound-wellbeing studies.

4.3. Field study using wearable devices

We conducted a multi-phase field study between May 2015 and August 2016 as part of the U.S. General Services Administration's Wellbuilt-for-Wellbeing (WB2) research program to understand the impact of indoor environment on the wellbeing of white collar office-workers (Sternberg et al. 2016). In the study, self-described healthy adult workers involved in a variety of office-based roles for the U.S. government were recruited across four federal office buildings in the Mid-Atlantic and Southern regions of the country. Buildings were selected for their representation of common office workstation types across the U.S. General Services Administration's portfolio of over 370 million square feet of office space that houses over 1 million employees. Staff in sections of each office building from organizations with leadership approval were offered the opportunity to participate. After giving written informed consent, participants completed an intake survey consisting of demographic questions. Participants wore two sensors for three days while carrying out their day-to-day activities, a heart and physical activity monitor, and a personal environment quality sensor-based device. The study also included experience sampling mobile surveys to collect individuals' perceived/psychological responses at periodic intervals of 1 to 2 hours, as well as using, stationary environmental sensors mounted on multiple walls in the study areas. Figure 4.1 shows a visual representation of the data collection mechanism used in the study using wearable devices, mobile-based surveys and wall-mounted sensors. The heart and physical activity monitor is a chest-worn wearable device named *EcgMove* 3 developed by *movisens* (Verkuil et al. 2016). The personal-environment quality sensor-based device is a multi-modal sensing neckwear device developed by *Aclima, Inc.* that measures ambient environment conditions including sound levels. Sound level exposure was measured as A-weighted continuous sound pressure levels reported in units of A-weighted decibels (dBA), a

measurement of the relative loudness of sounds relative to absolute silence as perceived by the human ear.

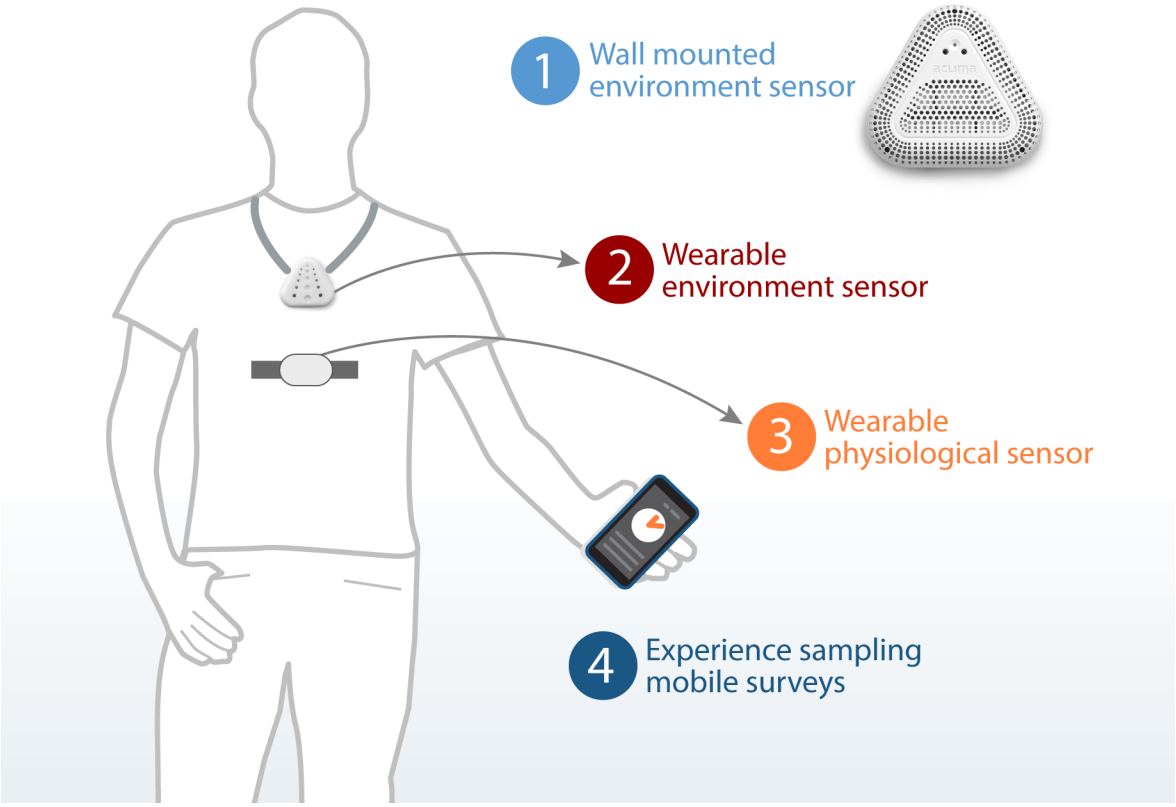


Figure 4.1: Study data collection mechanism consisting of two wearable sensors, a mobile survey application and wall mounted sensors

Section 6.4 of the supplementary materials provides description of all the information collected during the field study.

4.4. Preliminary analysis

We trained a two-level classical multilevel regression model[2] for our data for different physiological outcomes as shown below:

[2] Section 6.2 of the supplementary materials summarizes multilevel regression modeling using Classical and Bayesian approaches

$$y_{ij} = \beta_0 + \gamma_{0j} + \sum_{k=1}^{K} \beta_k x_{kij} + \sum_{m=1}^{M} \gamma_{mj} z_{mij} + \epsilon_{ij} \quad (4.1)$$

In Equation (4.1), y_{ij} is the physiological wellbeing measure (i.e., SDNN, normalized-HF) for the i^{th} observation and j^{th}individual, β_0 is the fixed intercept, $\mathrm{B} = \{\beta_1, \dots, \beta_K\}$ are coefficients for K fixed effects $\{x_1, x_2, \dots, x_K\}$, $\Gamma_0 = \{\gamma_{01}, \gamma_{02}, \dots, \gamma_{0j}, \dots, \gamma_{0J}\}$ are J random intercepts for each individual, $\Gamma = \{\gamma_{11}, \dots, \gamma_{1j}, \dots, \gamma_{MJ}\}$ are coefficients for $M\ X\ J$ random effects $\{z_1, z_2, \dots, z_M\}$, and ϵ_{ij} is the residual error. We assume a variance component structure for the covariance matrix of the random effects coefficients since it makes the least assumptions (Raudenbush and Bryk 2002). Sound level was included as a fixed effect as well as random effect in the model. We considered including higher-levels in the multilevel model (i.e., organization type, buildings, participant cohorts, work type, etc.), but the model fit did not improve significantly; hence, we restrict the model to vary at two levels, i.e., for variables varying at within-individual level (level-1) and for variables varying at between-individual level (level-2) in the data. We tested all covariates from section 6.4 of the supplementary materials and only retained significant variables in the model.

Consistent with previous studies (Kraus et al. 2013; Lee et al. 2010), we observed a curvilinear relationship between sound levels and physiological wellbeing measures (i.e., SDNN, normalized-HF) in two dimensional scatter plots. We found that the fixed-effect of the sound level variable, first order as well as second order, in the two univariate multilevel regression models for SDNN and normalized-HF as outcomes were significant, i.e., $\beta_{Sound,SDNN} = 0.1038$ ($p = 0.0000$), $\beta_{Sound^2,SDNN} = -0.0075$ ($p = 0.0000$), $\beta_{Sound,normalized-HF} = -0.0979$ ($p = 0.0000$), and $\beta_{Sound^2,normalized-HF} = 0.0013$ ($p = 0.015$). Thus, we can infer that sound level has a curvilinear effect on SDNN and normalized-HF. Secondly, we found that including sound

level as a random-effect improves the quality of fit of the multilevel model. This implied that the effects of sound level on the two physiological wellbeing measures varies across individuals. To study the sound-wellbeing relationship further, we required methods to (i) model the curvilinear relationship of sound level on two physiological wellbeing measures (i.e., SDNN and normalized-HF), (ii) simultaneously model the effects of sound level on two different measures of physiological wellbeing (i.e., SDNN and normalized-HF), and (iii) model the heterogeneity in the effects of sound on physiological wellbeing. Corresponding to each modeling problem, we formulate the following research questions:

Research question 1 (RQ1): How can we effectively model curvilinear relationships between sound level and the two wellbeing measures (i.e., SDNN and normalized-HF)?

Research question 2 (RQ2): How can we simultaneously model the relationship between sound level and two wellbeing measures (i.e., SDNN and normalized-HF)?

Research question 3 (RQ3): What are the factors contributing to the individual heterogeneity in effects of sound level on wellbeing?

In the next section, we consider each modeling problem, review corresponding modeling methods literature, identify the research gaps, and propose new methods to address the corresponding research questions.

4.5. Modeling methods

4.5.1. Modeling curvilinear effects

RQ1: How can we effectively model curvilinear relationships between sound level and the two wellbeing measures (i.e., SDNN and normalized-HF)?

4.5.1.1. Existing methods

A curvilinear relationship between an input and an outcome is commonly observed in IS (Liu and Goodhue 2012; Pant and Srinivasan 2010; Xue et al. 2011) and other disciplines (Geng et al. 2017). Considering techniques in extant literature, we observe that polynomial regression models account for higher order relationships but they are not directly interpretable (Durban et al. 2005). Segmented regression is an optimal approach for modeling curvilinear relationships, for it is robust, has fewer underlying assumptions and is easier to interpret (Jirschitzka et al. 2016). The primary challenge in using a segmented regression approach is the determination of change points linking the input segments (Shuai et al. 2003). Common procedures to determine change points in segmented regression models for simple data (Shuai et al. 2003) cannot be used for modeling the sound-wellbeing relationship since determination of the likelihood function is not straightforward for the multilevel data structure in our study. A recent method was proposed based on maximum-likelihood estimation of a continuous functional approximation of the piece-wise linear function (Muggeo et al. 2014) as an alternative to subjective assignment based on visualization of pair-wise plots (Kraus et al. 2013). However, this method estimates multiple change points automatically with no scope for user inputs into the estimation process (e.g., including or dropping change points if they are at extremities of the input distribution, etc.). Therefore, we conclude existing procedures for determining change points in studies employing segmented models for sound-wellbeing are either ad-hoc or analytically complex, leading to problems such as low external validity and overfitting respectively. To summarize, there is a need for a validated method to determine the change points in segmented multilevel models that is robust, efficient, and transparent. With such a method, we can accurately determine change points that create piece-wise linear functions of the sound-wellbeing relationship and facilitate direct interpretation from the linear model.

4.5.1.2. Semi-automated change point determination method

Consider the multilevel model described in Section 4.4. Sound level as an input variable varying at level-1 (i.e., fixed-effect) and having a curvilinear relationship with the outcome can be expressed as the sum of segmented variables as follows:

$$x_r = x_{1r} \cdot I(x_r < \eta_1) + x_{2r} \cdot I(\eta_1 \leq x_r < \eta_2) + \cdots + x_{kr} \cdot I(\eta_k \leq x_r) \quad (4.2)$$

In Equation (4.2), $\mathrm{H} = \{\eta_1, \eta_2, \ldots, \eta_k\}$ is a set of k change points defined for the input variable x_r. $I(\varphi)$ is an indicator function equal to 1 if condition φ is true; otherwise it is 0. As can be seen, the problem here is to estimate each change point η_i as well as to determine the total number of change points k. We propose a three step semi-automated method to estimate the change points and determine k as follows:

Step I: Fit a Generalized additive mixed model and visualize component smooth functions

Fit input x_r as a non-parametric spline in a Generalized Additive Mixed Model (GAMM) and visualize its component smooth function (Faraway 2006). Identify the order of the curve by inspecting the number of extrema (i.e., minima and maxima), and set the value of k. Note the value of the maxima and minima to be used as starting points in a linear search algorithm in the next step. This step is also used to determine whether or not to opt for a segmented model over a linear model, by inspecting the curvilinear nature of component smooth function.

Step II: Perform a linear search for change points using optimization with box constraints.

Consider a model fit metric such as Akaike information criteria (AIC), Bayesian information criteria (BIC), Deviation or Mean-squared error as an optimizing function. Select a suitable range around each starting point selected in step I, and run the optimization algorithm with the given range as a box constraint (Brent 2013). Select set of change points H that maximize model fit.

Step III: Fit the segmented multilevel regression model as shown below:

$$y_{ij} = \beta_0 + \gamma_{0j} + \sum_{s \in S} \beta_{rs} x_{rij} I(x_{rij} \in s) + \sum_{k=1,\ k \neq r}^{K} \beta_k x_{kij} + \sum_{m=1}^{M} \gamma_{mj} z_{mij} + \epsilon_{ij} \qquad (4.3)$$

In Equation (4.3), S is a set of segments constructed using change points H identified in Step II for input x_r. Significance of the effect of input variable, x_r, at each segment, s, can be determined by inspecting the corresponding fixed effects coefficient, β_{rs}, under regular conditions.

4.5.2. Simultaneous modeling of multiple outcomes

RQ2: How can we simultaneously model the relationship between sound level and two wellbeing measures (i.e., SDNN and normalized-HF)?

4.5.2.1. Existing methods

Existing studies analyzing the effects of sound level on multiple physiological wellbeing measures fit a different model for each outcome and report coefficients for each of the models separately (Cvijanović et al. 2017; Kraus et al. 2013; Park and Lee 2017; Sim et al. 2015). Interpretation and communication of results from multiple models for decision-making can be challenging. A statistical model with a single set of coefficients for multiple outcomes, known as simultaneous modeling, is suitable for this purpose (Baldwin et al. 2014; Das et al. 2004; Pituch and Stevens 2016). Simultaneous modeling differs from multivariate modeling, where coefficients are estimated for each outcome along with cross-correlation parameters (Lin et al. 2017; Ritz et al. 2017). For example, for three outcomes and three inputs, a simultaneous multiple regression model will contain three coefficients (excluding the intercept), whereas a multivariate regression modeling procedure will estimate nine coefficients (excluding the intercepts for outcomes) and corresponding covariance between the coefficients. Simultaneous modeling can be done by carrying out a univariate transformation of the outcomes after accounting for heterogeneity in error

variances (Baldwin et al. 2014; Faraway 2016; Pituch and Stevens 2016). In the univariate transformation method, even though different outcomes have different error variances in the model, the effects of input variables are assumed to be uniform across outcomes. For example, for a model measuring stress using breathing rate and heart rate as two health indicators, one can expect that the effects of inputs on each of the outcomes are different in scale. Latent variable modeling is an alternative approach for simultaneous modeling of multiple outcomes (Muthén 2002). However, classical latent variable modeling approaches (e.g. structural equation modeling) traditionally require individual items of a latent construct to be theoretically related and have construct validity (Kline 2012). Secondly, for multilevel modeling, diagnostic checking for structural equation modeling is more challenging to satisfy all the assumptions required in the classical approach (Hox 2013). The estimation procedure becomes more complex with a large number of random effects. Hence, there is a need for a new method that can overcome challenges of existing methods. Such a method can be used for making inferences over the effects of sound level on the two physiological wellbeing measures, SDNN and normalized-HF, examined in our study.

4.5.2.2. Bayesian latent variable modeling method

Consider a Bayesian latent variable model (Merkle and Wang 2016) for outcomes $Y = \{y_1, y_2, \dots, y_h, \dots, y_H\}$ as follows:

$$y_{ih} | \, \theta_i, \gamma_h, \lambda_h, \sigma_{\text{ih}} \sim N\left(\mu_{ih}, \sigma_{ih}^2\right) \tag{4.4}$$

$$\mu_{ih} = \gamma_h + \sum_{k=1}^{m} \lambda_{hk} \theta_{ik} \tag{4.5}$$

$$\theta_{ik} \sim N_m(0, \Phi) \tag{4.6}$$

For simultaneous modeling, we set $m = 1$ in the previous equation and express the latent variable as an outcome of a multilevel regression model as shown below:

$$\theta_{ij} = \beta_0 + \gamma_{0j} + \sum_{k=1}^{K} \beta_k x_{kij} + \sum_{m=1}^{M} \gamma_{mj} z_{mij} + \xi_{ij} \tag{4.7}$$

$$\gamma_{0j} \sim N\left(0, \sigma_{\gamma_0}^2\right), \gamma_{mj} \sim N\left(0, \sigma_{\gamma_m}^2\right), \xi_{ij} \sim N(0, \sigma_\theta^2) \tag{4.8}$$

Upon centering the outcomes and dropping the outcome intercept parameter γ_h, we can combine the within-individual level error variances (i.e., σ_{ih}^2 and σ_θ^2). The resultant Bayesian latent variable model is represented as follows:

$$y_{hij} - \overline{y_{hij}} = \left(\beta_0 + \gamma_{0j} + \sum_{k=1}^{K} \beta_k x_{kij} + \sum_{m=1}^{M} \gamma_{mj} z_{mij}\right) \cdot \lambda_h + \epsilon_{ij}^{(h)} \tag{4.9}$$

$$\gamma_{0j} \sim N\left(0, \sigma_{\gamma_0}^2\right), \gamma_{mj} \sim N\left(0, \sigma_{\gamma_m}^2\right), \epsilon_{ij}^{(h)} \sim N(0, \sigma_h^2) \tag{4.10}$$

This Bayesian latent variable model (i.e., Equations (4.9) and (4.10)) can be used for simultaneously modeling the effects of sound level on the two physiological wellbeing measures, SDNN and normalized-HF. The factor loadings, λ_h, automatically assign different weights to each outcome (i.e., λ_1 and λ_2), overcoming the limitation in the existing univariate transformation-based modeling methods. To ensure identifiability, we set $\lambda_1$as 1, and set $\lambda_h$relative to λ_1, following Merkle and Wang (2016).

The univariate transformation-based modeling method is a special case of latent variable modeling method, where we set the factor loadings of all outcomes, λ_h, to 1 (Refer to section 6.3 of the supplementary materials for model representation in a univariate transformation-based modeling method). A corresponding latent variable model can be developed using a classical approach, where the outcomes can be considered as reflective measures for a latent construct, modeled as

function of input variables using a two-level structural equation model (SEM). Software such as Mplus, LISREL, EQS, lavaan, OpenMx can fit two-level SEM with random intercepts (Skrondal and Rabe-Hesketh 2004). In the multilevel SEM model, each outcome y_{ijh} is split into a within and a between component as follows:

$$y_{ij} = \left(y_{ij} - \bar{y_j}\right) + \bar{y_j} = y_W + y_B \tag{4.11}$$

In Equation (4.11), both the within and between covariance components are treated as orthogonal and additive latent variables (Heck and Thomas 2015). The maximum likelihood estimate for parameters is derived by minimizing the overall loglikelihood which is the sum of likelihood of data from J individuals. The latent variable model using the classical approach offers less flexibility than its Bayesian counterpart, for it solicits more data-related assumptions and does not account for random effects of sound level (Heck and Thomas 2015; Kline 2012).

4.5.3. Modeling heterogeneity in effects

RQ3: What are the factors contributing to the individual heterogeneity in the effects of sound level on wellbeing?

4.5.3.1. Existing methods

While it is of interest to understand the overall effect of an input on an outcome in a population, insights regarding how and why effects differ across individuals can be valuable (Abrams and Hens 2015; Dingemanse and Dochtermann 2013; Gimenez et al. 2018). The random-effects indicate the presence of individual heterogeneity in effects of an input on the outcome in a multilevel model (Raudenbush and Bryk 2002). A naïve approach to identify factors contributing to individual heterogeneity is to introduce each factor in an interaction term with the input variable and test its significance, an approach known as slopes-as-outcomes modeling (Becker et al. 2013; Raudenbush and Bryk 2002). However, this hypothesis testing-based approach is sensitive to noise

in data, increases the chances of a Type II error (i.e., even if person-level factor contributes to individual heterogeneity, it is insignificant as a moderator in the model), and becomes cumbersome as the number of potential factors increases. There are no existing validated methods for the identification of factors contributing to individual heterogeneity in effects measured by random effects in multilevel models. Therefore, there is a need for a new method to identify person-level factors associated with individual heterogeneity in the effects of sound level on wellbeing.

4.5.3.2. Varying-coefficients modeling method

We propose the varying-coefficients modeling method as a two-step procedure: step 1, quantifying heterogeneity, and step 2, identifying factors that contribute to heterogeneity. The method can be used to identify person-level variables that explain the heterogeneity in the effects of sound level on physiological wellbeing across individuals.

In the first step, we fit a Bayesian hierarchical linear model with all input variables with varying coefficients having normal priors with non-zero means. Person-level variables (e.g., age, BMI, gender, etc.) are not included in the model since their value is constant for each individual (i.e., varying coefficients for person-level variables have distribution with zero variance). The hierarchical Bayesian linear model for step 1 is given in Equations (4.12) and (4.13).

$$y_{hij} = \left(\gamma_{0j} + \sum_{m=1}^{M} \gamma_{mj} z_{mij}\right) \cdot \lambda_h + \epsilon_{ij}^{(h)} \tag{4.12}$$

$$\gamma_{0j} \sim N\left(\mu_{\gamma_0}, \sigma_{\gamma_0}^2\right), \gamma_{mj} \sim N\left(\mu_{\gamma_m}, \sigma_{\gamma_m}^2\right), m \in \mathbb{Z}_M\,, \epsilon_{ij}^{(h)} \sim N(0, \sigma_h^2) \tag{4.13}$$

Note that the mean values μ_{γ_0} and μ_{γ_m} in Equation (4.13) are analogous to the model intercept and the corresponding fixed effect coefficients of the m^{th} variable in the Bayesian latent variable model (i.e., Equations (4.9) and (4.10)).

In the second step, we formulate the varying coefficients of sound level as an outcome of a linear model with person-level variables as the input variables as given in Equation (4.14).

$$\gamma_{rj} = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pj} + \epsilon_j,\ \epsilon_j \sim N(0, \sigma_r^2) \tag{4.14}$$

In Equation (4.14), $\Gamma_r = \{\gamma_{r1}, \gamma_{r2}, \dots, \gamma_{rJ}\}$ are the varying coefficients for sound level in the Bayesian hierarchical linear model from step 1, $\{x_{1\cdot}, x_{2\cdot}, \dots, x_{P\cdot}\}$ are P person-level variables, and ϵ_j is a normally distributed residual error varying across J individuals.

The problem of identifying person-level factors contributing to individual heterogeneity effects is presented as a variable selection problem in our linear model. Traditional stepwise feature selection methods for regression models are ridden with challenges such as sensitivity to changes in data and low external validity (Hastie et al. 2009). These challenges are particularly relevant in our problem, where there are multiple person-level variables that could be factors contributing to heterogeneity in the sound effects on wellbeing across individuals. Therefore, we choose three regularization based methods, *lasso* (Tibshirani 1996), *elasticnet* (Zou and Hastie 2005), and *adaptive lasso* (Zou 2006) to determine significant inputs in the linear model. The *elasticnet* and *adaptive lasso* methods are improvements over the *lasso* feature selection method and account for correlated features and possess oracle properties (Hastie et al. 2009). For each of these methods, the problem of feature selection is represented in the regularization modeling framework as an optimization problem, as $\underset{\beta}{\operatorname{argmin}}\, f_{Loss}(\beta) + f_{Penalty}(\beta)$. The loss function for linear regression is the sum of squared errors given by $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{p=1}^{P} x_{ip}\,\beta_p\right)^2$ and the penalty function for regularized models is given in Equations (4.15), (4.16), and (4.17).

$$f_{Lasso-Penalty}(\beta) = \lambda \sum_{p=1}^{P} |\beta_p| \quad (4.15)$$

$$f_{Elasticnet-Penalty}(\beta) = \lambda[(1-\alpha)\sum_{p=1}^{P}|\beta_p|^2/2 + \alpha \sum_{p=1}^{P}|\beta_p|] \quad (4.16)$$

$$f_{Adaptive\ lasso-Penalty}(\beta) = \lambda \sum_{p=1}^{P} w_p|\beta_P| \quad (4.17)$$

The hyperparameters λ and α are determined using grid-search procedure (Hastie et al. 2009). The initial adaptive weights are set as $\frac{1}{|\beta_p^{OLS}|}$, or inversely proportional to the absolute values of naïve regression coefficients of inputs as proposed by Zou (2006). We choose the person-level variables that have non-zero coefficients in all three regularized models as the final set of factors contributing to individual heterogeneity effects.

Table 4.2 summarizes existing and proposed methods for addressing the three challenges in sound-wellbeing modeling.

Table 4.2: Methods addressing challenges in sound-wellbeing modeling

Existing/Proposed methods	**Modeling challenges**		
	Modeling curvilinear effects (using segmented models)	**Simultaneous modeling of multiple outcomes**	**Modeling heterogeneity in effects**
Existing methods	• Heuristic approach (Kraus et al. 2013) • Maximum likelihood based method (Muggeo et al. 2014)	• Classical approach – Univariate method (Baldwin et al. 2014)	• Slopes-as-outcomes modeling method (Raudenbush and Bryk 2002)
Proposed methods	• Semi-automated change point determination method	• Bayesian latent variable modeling method	• Varying-coefficients modeling method

4.6. Analysis

4.6.1. Data pre-processing

A total of 248 office workers expressed interest in participating in our study (described in section 4.3), representing approximately 12% of the worker force located in areas of the office buildings where recruitment took place. Pregnant women and those wearing pacemakers or insulin pumps were excluded. Participants taking medication known to affect cardiac activity were noted but not excluded. Due to scheduling problems, sickness and exclusionary criteria, 17 office workers did not participate, resulting in a total enrolment of 231 participants. Due to unexpected changes in work schedules, eight of the 231 participants were only observed for two, rather than the full three days.

The heart rate variability measures SDNN and normalized-HF were calculated using cardiac activity measured by EcgMove3 according to the guidelines of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology (Malik et al. 1996). Physical activity levels were assessed in G (i.e. 1 unit of gravitational force) from the EcgMove3's triaxial accelerometer sensor (Razjouyan et al. 2018). Sound levels were aggregated at 5-minute intervals to be integrated with physiological wellbeing measures SDNN and normalized-HF, assuming no lagged effects (Srinivasan et al. 2017). Only observations with both outcome values present were considered in the analysis. Observations with outcome values above the 99.5^{th} percentile were discarded. Age and BMI were discretized to five and four levels, respectively, for ease of interpretation. Data of participants with less than one hour of recorded data were excluded from analysis. Missing values in input variables were imputed using mean values. Our final dataset contained 31,557 observations aggregated at five-minute intervals and processed approximately 200,000 minutes of wearable data streams from the 231 participants. Apart from sound level as

the input variable and SDNN and normalized-HF as the outcomes, person-level variables (e.g., age, gender, etc.), temporal indicators (time of day, day of the week), and physical activity levels were included as covariates in the statistical models. Observations from day 1 and day 2 of participation of all participants were considered as the training dataset, and day 3 observations were used as the holdout sample (i.e., test dataset) for evaluating the predictive performance of models. Summary statistics of relevant intrapersonal variables (i.e., wearable device based repeated measures and temporal information) and interpersonal variables (i.e., person-level information) in this study are shown in Table 4.3.

Table 4.3: Summary statistics of our data

Variable	**Summary**			
INTRAPERSONAL				
Numerical	**Mean**	**SD**	**Units**	**% missing**
SDNN	53.08	23.33	ms	-
Normalized-HF	19.81	12.70	%	-
Sound level	51.85	8.79	dBA	4.29
Physical activity level	0.1738	0.3164	G	0.07
Categorical	**Category**	**Hours:Mins**	**Proportion**	**% missing**
Time of day				0
	Morning	1224:10	45.76	
	Afternoon	1039:30	38.85	
	Evening	411:15	15.37	
Day of week				0
	Monday	449:25	16.80	
	Tuesday	860:50	32.18	
	Wednesday	916:55	34.28	
	Thursday	431:50	16.14	

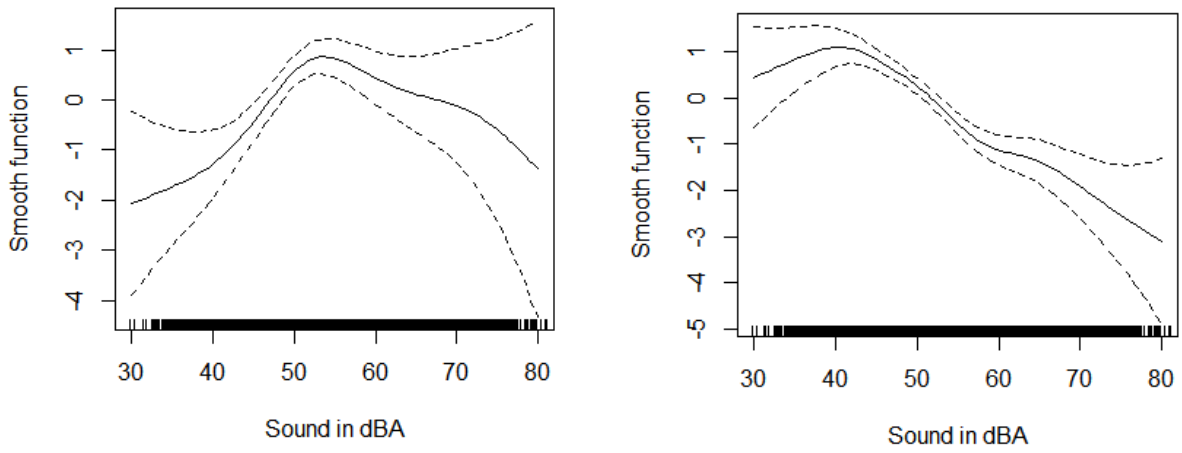
	Friday	15:45	0.59	
INTERPERSONAL				
Numerical	**Mean**	**SD**	**Units**	**% missing**
Neuroticism	3.21	0.97	Scale 1-7	10.38
Noise sensitivity	4.05	1.17	Scale 1-7	9.52
Average sound exposure	51.99	4.89	dBA	4.33
Categorical	**Category**	**No. of participants**	**Proportion**	**% missing**
Age				9.95
	Less than 30 years	30	12.98	
	30 - 39 years	62	26.83	
	40 - 49 years	43	18.61	
	50 - 59 years	56	24.24	
	60 years or above	17	7.36	
Gender				12.12
	Male	88	38.09	
	Female	115	49.78	
BMI				10.39
	18.5 - 25	76	32.9	
	25.1 - 30	81	35.06	
	30.1 - 35	30	12.98	
	Above 35.1	20	8.66	
Computer-dominant work				8.66
	Yes	93	40.26	
	No	118	51.08	
Management work				8.66
	Yes	69	29.87	
	No	142	61.47	

Technical work				8.66
	Yes	90	38.96	
	No	121	52.38	
Meeting heavy work				8.66
	Yes	42	18.18	
	No	169	73.16	
Sleep problems				9.09
	Yes	42	18.18	
	No	168	72.73	
High blood pressure				9.09
	Yes	42	18.18	
	No	168	72.73	
Anxiety				9.09
	Yes	38	16.45	
	No	172	74.46	

4.6.2. Modeling curvilinear effects of sound levels on physiological wellbeing

We applied and evaluated our semi-automated method to determine the change points for fitting segmented multilevel models, associating sound level with SDNN and normalized-HF as physiological wellbeing outcomes.

The first step of the semi-automated method is to fit a Generalized Additive Mixed Model (GAMM) and visualize its sound level component smooth function. The component smooth functions of sound level on the outcomes, SDNN and normalized-HF, are shown in Figure 4.2(a) and Figure 4.2(b), respectively. Smooth functions in both models were observed to be curvilinear with a single maximum across the range of sound level in the dataset.



(a) (b)

Figure 4.2: Component smooth function of sound level in GAMM for (a) SDNN as outcome and (b) Normalized-HF as outcome

The second step of the semi-automated method is to perform a linear search for change points using optimization with box constraints. For models corresponding to outcomes SDNN and normalized-HF, we chose starting values of 55 dBA and 45 dBA across search ranges [40, 60] and [30, 50], respectively, for running a linear search of change points. Brent's optimization algorithm with box constraints (Brent 2013) was used as the linear search method. We identified 51 dBA and 39 dBA as change points for two models with SDNN and normalized-HF as outcomes, respectively. Finally, we fitted segmented multilevel regression models using these change points over the training data. Fixed-effects coefficients of sound level segments in models with outcomes SDNN and normalized-HF are as shown in Table 4.4.

Table 4.4: Fixed-effects coefficients of sound level in segmented multilevel models

Outcome	**Segment**	**Coefficient (SE)**
SDNN	Sound level < 51 dBA	0.1425 (0.06)**
	Sound level ≥ 51 dBA	-0.0682 (0.03) **
Normalized-HF	Sound level < 39 dBA	Insignificant
	Sound level ≥ 39 dBA	-0.0998 (0.02)***

** = $p < .05$, *** = $p < .01$

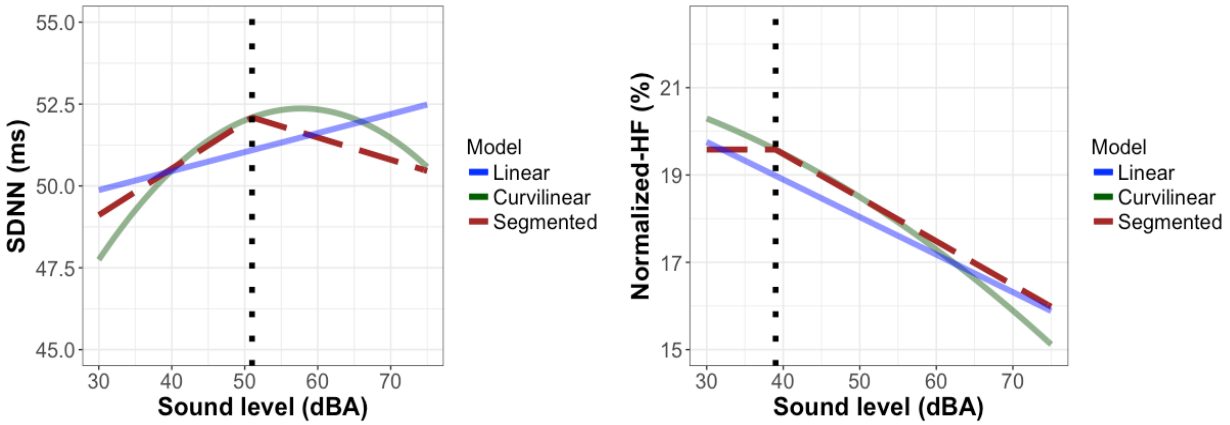
We tested the robustness of the change point estimates by varying search ranges and starting points around the maxima identified in previous step and got estimates within +/- 1 dBA tolerance of the previous estimates. Further, we found that other sophisticated optimization algorithms such as BFGS and L-BFGS (Fletcher 2013) gave similar estimates for change points. We compared the performance of the segmented multilevel model with change points determined using our proposed method with the performance of multilevel models with: (i) sound level as linear input, (ii) sound level as curvilinear input (i.e., first order and second order effects) (iii) sound level segmented using the maximum likelihood approach (Muggeo et al. 2014), and (iv) sound level segmented using ad-hoc approach (Kraus et al. 2013). Using the maximum likelihood method (Muggeo et al. 2014), the change points determined for sound level were 53 dBA and 76 dBA for models with SDNN and normalized HF as outcomes respectively. For the ad-hoc method, we inspect the component smooth curves of the GAMM models and set the change points as 55 dBA and 45 dBA for models with SDNN and normalized HF as outcomes, respectively. The fixed effects model was used as a baseline, representing the case when only fixed effects of sound level are considered in the multilevel model. Sound level was included in the fixed as well as random effects components of other models. Model fit was checked using pseudo R-Squared (Nakagawa and Schielzeth 2013). Predictive performance was compared using Root Mean Squared error (RMSE) and Mean Absolute Prediction Error (MAPE) on the test dataset. The model fit and prediction accuracy comparisons across models are shown in Table 4.5. The model fit and error estimates for best performing models are highlighted for reader convenience. Better model fit and predictive performance corresponds to a higher value of R-squared and lower error values. The models with segmented inputs perform better than models with linear inputs, but they are equivalent to models with curvilinear inputs in terms of fit and predictive performance. Table 4.5 shows that segmented models with change points determined using our method are better than

segmented models with change points using inspection alone (ad-hoc method) or using the maximum likelihood method.

Table 4.5: Model fit and predictive performance comparison of segmented multilevel models

Model	SDNN			Normalized-HF		
	R-sq.	**RMSE (ms)**	**MAPE (%)**	**R-sq.**	**RMSE (ms)**	**MAPE (%)**
Fixed effects only (baseline)	0.5098	17.84	26.19	0.5026	9.17	46.12
Linear inputs	0.5555	17.44	25.18	0.5202	9.08	44.71
Curvilinear inputs	0.5815	17.21	24.73	**0.5329**	8.97	**44.17**
Segmented inputs using ad-hoc method (Kraus et al. 2013)	0.5832	17.21	24.71	0.5316	8.97	44.19
Segmented inputs using Maximum Likelihood method (Muggeo et al. 2014)	0.5837	**17.20**	24.71	0.5319	8.98	44.18
Segmented inputs using our semi-automated method	**0.5838**	**17.20**	**24.69**	0.5323	**8.96**	44.18

The fixed effects coefficient of sound level in the linear, curvilinear, and segmented model are visually represented in Figure 4.3. Figure 4.3 shows that segmented models represent the curvilinear relationship better than a linear model and are easier to interpret than the curvilinear model in terms of unit change in outcome as a function of unit change in the sound level.



(a) (b)

Figure 4.3: Trajectory of linear, curvilinear and segmented fixed-effects coefficients for (a) SDNN as outcome, and (b) Normalized-HF as outcome

4.6.3. Simultaneous modeling of sound level effects on wellbeing measures

We evaluated and applied the Bayesian latent variable modeling method for simultaneously modeling effects of sound level on SDNN and normalized-HF. In the model, fixed-effects were introduced for variables sound level, physical activity level, time of day, day of week, age group, BMI group and gender, and random-effects were introduced for variables sound level and physical activity level. We repeated the change point determination procedure for segmented multilevel models to get a single optimal sound level for outcomes SDNN and normalized-HF. The component smooth function for the GAMM model for sound level is shown in Figure 4.4. We selected starting values of 45 dBA and search range [40, 60] for running linear search of change point and identified 50 dBA as an optimal change point for the segmented regression dual-outcome model. We used the classical univariate transformation method for simultaneous modeling of outcomes in the change point determination procedure since the linear search procedure with Bayesian variable modeling method took more than a day to converge.

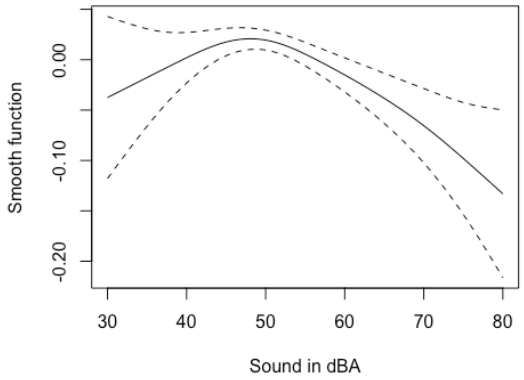


Figure 4.4: Component smooth function of sound level in GAMM for physiological wellbeing as a bivariate function of SDNN and Normalized-HF

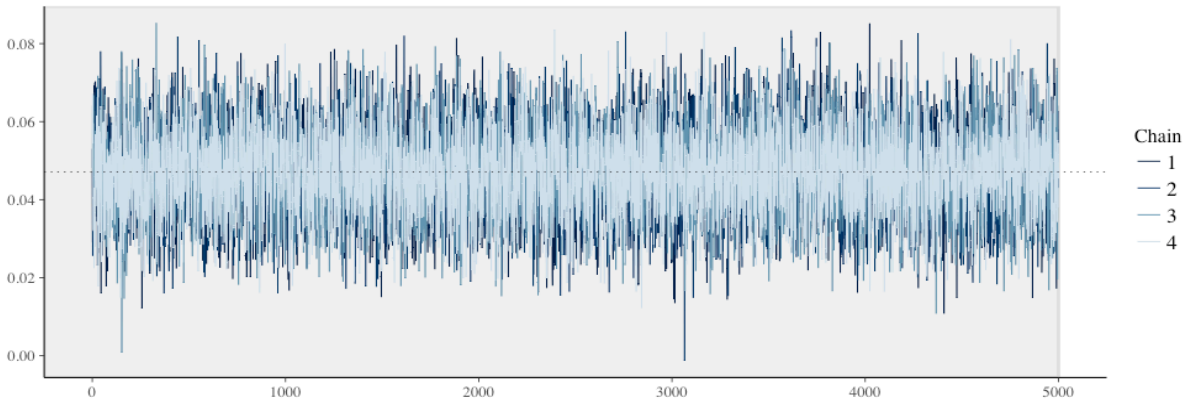
We standardized the input (sound level) as well as the outcomes (SDNN and normalized-HF) to remove sensitivity and challenges in posterior estimation convergence due to scale differences in the units. For models fit using the Bayesian approach, parameters were assigned a diffused Normal prior, and the error variances were assigned a diffused half-Cauchy prior (Gelman and Hill 2007). The Hamiltonian Monte Carlo algorithm was used for sampling four parallel chains (Carpenter et al. 2017). The R-hat statistic cutoff < 1.1 and zero divergence check were used as validation tests for posterior estimates of parameters and assessing quality of fit (Carpenter et al. 2017).

The mean posterior distribution estimates and the 90% credible intervals (values between 5^{th} and 95^{th} percentile of the posterior distribution) of the fixed effects coefficients for models fitted using the Bayesian latent variable modeling method is given in Table 4.6. The posterior estimates of the fixed-effects of sound level, time of day, day of week, physical activity level, age, and BMI indicate that they are interpersonal and intrapersonal factors related to an individual's physiological wellbeing at workplace. The trace plots of four chains of MCMC draws for the coefficient of sound level for the two conditions, sound level < 50 dBA and sound level >= 50 dBA shown in Figure 4.5, indicating good convergence. Posterior estimates of all the parameters

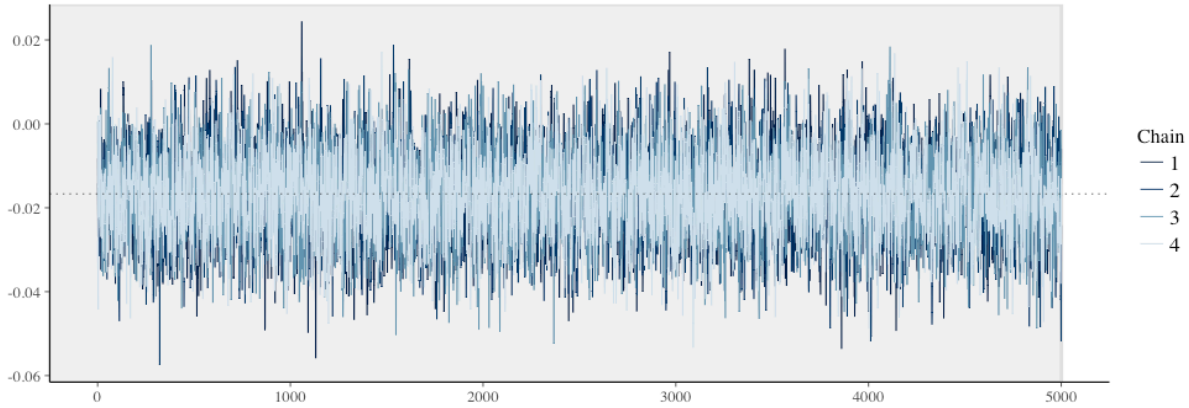
in the model (fixed-effects as well as varying effects) had R-hat values less than 1.1, and the model convergence report indicated zero divergence check, indicating an acceptable model fit.

Table 4.6: Fixed effects of models using different simultaneous outcomes modeling methods

Coefficients	**Posterior estimate (mean)**	**90% Credible interval**
Sound $level_{Normalized}$ (< 50 dBA)	0.0471	(0.0199 – 0.0648)
Sound $level_{Normalized}$ (>= 50 dBA)	-0.0167	(-0.0337 – -0.0042)
Physical activity $level_{Normalized}$	0.2756	(0.2316 – 0.2932)
Time of day – Morning	*Baseline*	
Time of day – Afternoon	-0.1479	(-0.1675 – -0.1277)
Time of day – Evening	-0.0939	(-0.1206 – -0.0690)
Day of week – Monday	*Baseline*	
Day of week – Tuesday	-0.1301	(-0.2670 – 0.0092)
Day of week – Wednesday	-0.0571	(-0.0888 – -0.0108)
Day of week – Thursday	-0.0588	(-0.0886 – -0.0287)
Day of week –Friday	-0.0430	(-0.0836 – 0.0277)
Age group – Below 30	*Baseline*	
Age group – 30-40	0.1361	(-0.1439 – 0.4115)
Age group – 40-50	-0.1468	(-0.4495 – 0.1641)
Age group – 50-60	-0.3119	(-0.6235 – -0.0038)
Age group – Above 60	-0.4413	(-0.7475 – -0.0132)
BMI group – Below 25	*Baseline*	
BMI group – 25-30	-0.2278	(-0.4281 – -0.0165)
BMI group – 30-35	-0.3619	(-0.6751 – -0.0896)
BMI group – Above 35	-0.6169	(-0.9768 – -0.2363)
Gender – Male	*Baseline*	
Gender – Female	-0.0439	(-0.2278 – 0.1435)



(a)



(b)

Figure 4.5: MCMC trace plots for coefficient of sound levels in following Bayesian latent variable model for: (a) sound levels < 50 dBA, and (b) sound levels >= 50 dBA

The fixed effect of sound level in the Bayesian latent variable model represents the global effects of sound on individual wellbeing after accounting for individual heterogeneity effects through the varying effects coefficients. The coefficient for sound level in Table 4.6 indicates change in physiological wellbeing by a standard deviation (SD) related to a unit standard deviation (SD) change in sound level as both input and outcomes are standardized. Knowing that 68.3% of variability is explained by 1 SD of a normal distribution and the SD of sound level in the dataset is 8.79 dBA (refer to Table 4.3), we can make the following inferences. For sound amplitudes lower than 50 dBA, a 10 dBA increase in sound level is related to a 3.6% increase in physiological

wellbeing. For sound amplitudes higher than 50 dBA, a 10 dBA increase in sound level is related to decrease in physiological wellbeing by 1.3%.

We compared the predictive performance of the Bayesian latent variable modeling method with the following three alternative methods for simultaneous modeling of multiple outcomes: (i) the classical univariate transformation method (Baldwin et al. 2014), (ii) the univariate transformation method trained using a Bayesian approach, and (iii) the classical multilevel structural equation modeling method (Kline 2011). Models using the classical approach are trained using the R packages lavaan (Rosseel 2012) and nlme (Pinheiro et al. 2007) in a 16 GB RAM, 2.7 GHz processor PC, whereas models using the Bayesian approach are written and executed using Stan program through the RStan interface (Carpenter et al. 2017), in a high performance computer cluster with 28 nodes (192 GB RAM per node, Intel Haswell V3 28 core processors). The RMSE and MAPE of the models trained using the four methods are given in Table 4.7.

Table 4.7: Comparing predictive performance of different simultaneous modeling methods

Model		**SDNN**		**Normalized HF**	
		RMSE	**MAPE**	**RMSE**	**MAPE**
Classical	*Univariate*	20.12	34.13	10.98	54.36
	Latent	23.71	44.78	11.22	57.10
Bayesian	*Univariate*	21.50	37.39	10.04	52.64
	Latent	**17.06**	**26.56**	**8.90**	**44.36**

Table 4.7 shows that the model trained using the Bayesian latent variable modeling method has the lowest prediction errors RMSE and MAPE, indicating that our method is superior to other methods for simultaneous modeling of multiple outcomes.

4.6.4. Individual heterogeneity in sound level effects on physiological wellbeing

The heterogeneity in the effect of sound level on physiological wellbeing across individuals is accounted for by the varying coefficients of sound level input in the Bayesian latent variable

model. Figure 4.6 shows a caterpillar plot visualization of posterior estimates of varying coefficients of sound level and their 60% credible interval (values between 20th percentile and 80th percentile of the posterior distribution) in the Bayesian latent variable model. The vertical lines show the corresponding fixed effects coefficients of sound level. The spread of mean values of posterior estimates of the varying coefficients indicate substantial individual heterogeneity effects. We applied the varying-coefficients method to identify person-level variables contributing to individual heterogeneity in sound level effects on physiological wellbeing.

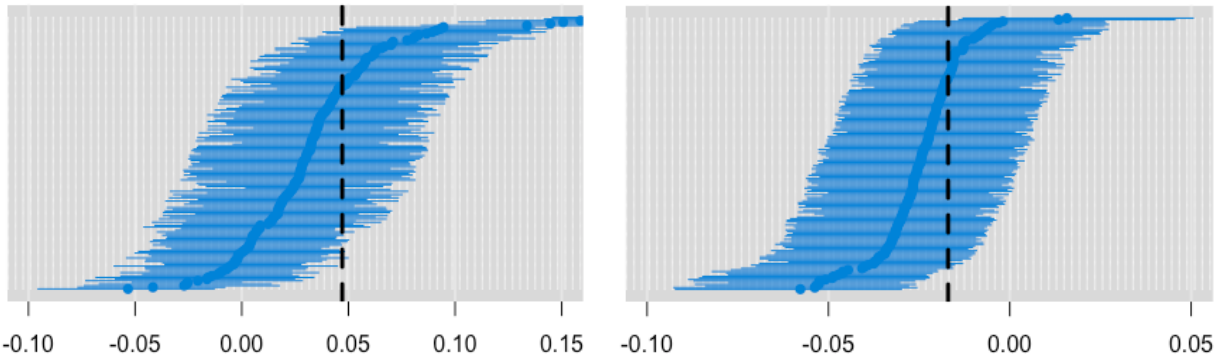


Figure 4.6: Caterpillar plots of posterior estimates of varying coefficients of sound level and their 60% credible interval in the Bayesian latent variable model for (a) sound level < 50 dBA, and (b) sound level >= 50 dBA.

We considered two subsets of the data, one with sound levels less than 50 dBA and the other with sound levels greater than or equal to 50 dBA, to fit two independent models. By fitting two independent models for instances with high sound levels (>= 50dbA) and instances with low sound levels (< 50 dBA), we were able to make independent inferences about individual heterogeneity effects for each scenario.

The varying-coefficients modeling method uses a two-step procedure for modeling heterogeneity effects and for identifying factors contributing to the heterogeneity effects. We fitted a Bayesian hierarchical model with input variables sound level, physical activity level, time of day, and day of week as variables with varying coefficients with normal prior having non-zero

means. In step 2, we used lasso, elasticnet, and adaptive-lasso regularized models to identify person-level variables that have a significant relation with the varying coefficients of sound level. The coefficients for the regularized feature selection models are shown in Table 4.8.

Table 4.8: Coefficients of person-level input variables in regularized models in varying coefficients modeling method

Predictors	**Below 50 dBA**			**Above 50 dBA**		
	Lasso	**Elastic-net**	**Adaptive lasso**	**Lasso**	**Elastic-net**	**Adaptive lasso**
Neuroticism						
Noise sensitivity						
Age group - Below 30	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
Age group - 30-40				-0.0011	-0.0076	-0.0002
Age group - 40-50						
Age group - 50-60	-0.0026	-0.0141	-0.0003			
Age group - Above 60	-0.0047	-0.0224	-0.0007	0.0084	0.0160	0.0010
BMI group - Below 25	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*	*baseline*
BMI group - 25-30	-0.0009	-0.0044	-0.0001		0.0004	
BMI group - 30-35				-0.0001	-0.0096	
BMI group - Above 35				0.0076	0.0123	0.0011
HighBP - Yes	-0.0133	-0.0764	-0.0021	-0.0207	-0.0203	-0.0042
Anxiety - Yes	-0.0015	-0.0013	-0.0002	0.0060	0.0148	0.0007
Sleep problems - Yes						
Computer use intensive work - Yes	0.0187	0.0881	0.0036			
Managerial work - Yes						
Meeting intensive work – Yes						
Technical work - Yes						
Average sound exposure						

Table 4.8 shows that Age groups, BMI groups, High BP (blood pressure), Anxiety, and

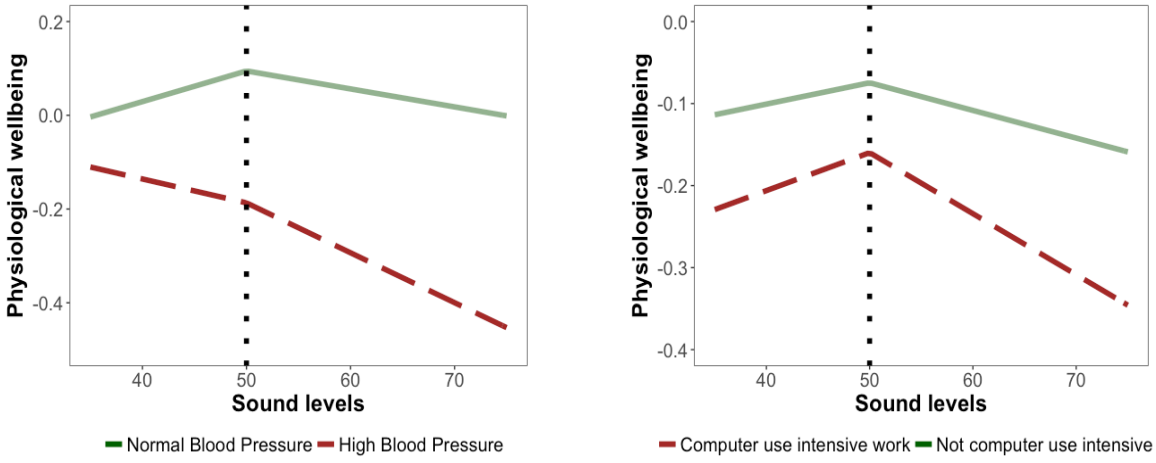
Computer use intensive worktype are the person level factors related to the variability in coefficients of sound level in the physiological wellbeing models. The blank cells show that coefficients of corresponding variables have been shrunk to zero in the corresponding feature selection method (i.e., lasso, adaptive lasso, elasticnet). To evaluate the performance of the varying coefficient modeling method, we compared the predictive performance of multilevel models with three set of input variables: (i) inputs including no person-level variables as moderators, (ii) inputs including all person-level variables as moderators, and (iii) inputs including person-level variables identified by varying-coefficients modeling method as moderators. Moderators were included as two-way interactions with fixed effects of sound level in the multilevel models. Table 4.9 shows the prediction errors of all three models with respect to outcomes SDNN and normalized HF. (Univariate outcomes were reconstructed using latent factors estimated with Bayesian latent variable modeling.) Table 4.9 shows that the model including the person-level variables identified using the varying-coefficients modeling method since moderators have the smallest RMSE and MAPE values as compared to other models.

Table 4.9: Performance comparison of multilevel models with different set of moderators

Moderators of sound level in multilevel model	**SDNN**		**Normalized HF**	
	RMSE	**MAPE**	**RMSE**	**MAPE**
No moderators	17.06	26.56	8.90	44.36
All person-level variables	19.66	31.38	11.13	47.73
Person-level variables identified by varying-coefficients modeling method	**16.65**	**24.97**	**8.41**	**43.23**

High BP and Computer use intensive worktype are person level factors that contribute most to the between-individual heterogeneity in sound level effects on physiological wellbeing (see Table 4.8). Figure 4.7(a) and Figure 4.7(b) are plots showing the change in outcome due to introducing interaction effects of High BP and Computer use intensive worktype variables with sound level fixed effects in multilevel model respectively. The fixed-effect coefficient of Sound level in multilevel models with stratified datasets participants belonging to categories Normal BP,

High BP, Computer use intensive worktype, and Not computer use intensive worktype are given in Table 4.10.



(a) (b)

Figure 4.7: Interaction plots of the top two person-level variables moderating the sound-wellbeing relationship

Figure 4.7(a) and Table 4.10 show that office-workers with high blood pressure are more negatively affected than participants with normal blood pressure. Figure 4.7(b) and Table 4.10 show that office-workers involved in computer intensive work have higher positive effects of sound levels on physiological wellbeing at amplitudes less than 50 dBA, but they have higher negative effects of sound levels on physiological wellbeing at amplitudes more than 50 dBA compared to other office-workers.

Table 4.10: Coefficients of sound level in stratified datasets

Stratification		**Standardized coefficients of sound level**	
		Less than 50 dBA	**Greater than 50 dBA**
Blood pressure	Normal BP	0.0232	-0.0239
	High BP	-0.0181	-0.0665
Type of work	Not computer use intensive	0.0092	-0.0211
	Computer use intensive	0.0165	-0.0461

4.6.5. Post-analysis group comparisons

To validate the presence of optimal sound level for physiological wellbeing at 50 dBA and the influence of blood pressure and work involving intensive computer use in moderating the sound-wellbeing relationship, we conducted post-hoc comparison of wellbeing across different stratified populations for three sound level conditions: sound level less than 45 dBA, sound level between 45 dBA and 55 dBA, and sound level greater than 55 dBA. Table 4.11 shows the post-hoc comparisons of mean wellbeing score adjusted for random effects for the three sound level ranges for different sub-populations in our data. In support of our finding that 50 dBA is an optimal sound level at workplace, we find that sound level range 45-55 dBA has the highest mean adjusted wellbeing score across the complete population, when compared to low and high sound level ranges. However, for individuals with high blood pressure, the lowest sound level range (i.e., sound level <= 45 dBA) is optimal, unlike individuals with normal blood pressure. Finally, individuals with computer use intensive work have a lower mean adjusted wellbeing score for low as well as high sound level ranges (i.e., sound level <=45 dBA and sound level > 55 dBA), when compared to individuals with regular computer use at work. These post-analysis group comparison findings validate the findings based on our proposed methods.

Table 4.11: Post-hoc group comparisons across sound level ranges

Sub-population	**Mean adjusted wellbeing score**††		
	sound <= 45 dBA	**45 dBA < sound <= 55 dBA**	**sound > 55 dBA**
Complete dataset	0.0054	0.0174	-0.0203
High BP†	-0.1395	-0.1600	-0.1884
Normal BP	0.0120	0.0302	-0.0081
Intensive computer use	-0.0407	-0.0186	-0.0985
Regular computer use	0.0229	0.0333	0.0144

† Repeated measures MANOVA shows significant differences across sound level ranges for each sub-population except high BP

†† Mean value of wellbeing is adjusted for random effects using estimated marginal means procedure (Searle 1980)

4.7. Discussion

Understanding the effects of ambient sound levels on an office-worker's physical health is critical for informing workplace design practices promoting health and wellbeing, which in turn facilitates positive organizational impacts. Traditional study methodologies such as controlled experiments and survey-based methods, are not suitable for sound-wellbeing modeling as real office environment ecosystems are more complex than simulated environments with a limited set of treatments and controls. We overcame the challenge of low external validity by using wearable sensors to collect continuous measurements of environmental and physiological conditions of multiple persons simultaneously. A preliminary analysis of the collected data showed that sound level has a curvilinear effect on two physiological wellbeing measures (i.e., SDNN and Normalized-HF) and the effect varies across individuals. We proposed three new methods for addressing the gap in statistical modeling methods for representing curvilinear effects, simultaneous modeling of multiple outcomes, and identifying factors contributing to individual heterogeneity in order to model the sound-wellbeing relationship.

4.7.1. Relevance to IS

IT artifacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems) (Hevner et al. 2004). Our major contributions include introduction of a unique study design using wearable devices for analyzing complex relationships, new quantitative methods for modeling digital data generated by wearables and informing workplace planning policies and practices that affect the health and wellbeing of office workers worldwide. Using wearable devices for analyzing workplace wellbeing allows us to get a fine-

grained measurement of the impact of sound on individual health. Our study allows researchers and practitioners to not only reconcile some of the differences in past work on the effect of sound on wellbeing, but to also separate out factors that should be controlled in future work (i.e., blood pressure and nature of work). One of the benefits and promises of information systems is enabling personalization. Individuals are affected by low and high sound levels in different ways. As wearable technology becomes widely available, personalized measurement is feasible and allows understanding the impact of our surroundings at an individual-level. This can improve workplace design and personal choices to maximize wellbeing, and in turn, improve our ability to function at our best in the workplace.

In analytics, predictive modeling and statistical modeling go side-by-side as one predicts the future using existing data, focusing on informing us on "What will be", while the other illuminates hidden patterns and tells us about "What is" with respect to a phenomenon. Both of them are important for creating value out of data generated from digital sources such as wearable devices. As the number of wearable technology-based applications increases in the future, the quantum of available data to analyze will exponentially increase and warrant more advancements in statistical modeling for meaningful pattern interpretations. Our method contributions in statistical modeling of wearables generated data are timely in IS research, as the discipline is widening its scope in design science using novel data sources including wearable devices (Abbasi et al. 2016; Chen et al. 2012; Rai et al. 2017).

4.7.2. Future research directions

Our study is an archetypal sensor-based field study employing multiple wearable sensors that collected data from participants for multiple days. It opens up a broader avenue for inter-disciplinary work with scientists in information systems, architecture, organizational behavior, environmental science, and physiology. In the future, novel design science applications such as

physician fatigue management, social interaction impact analysis, and sports performance optimization can be based on our study design, where it is possible for study participants to wear one or more sensors while conducting their day-to-day activities. Even though curvilinear effects, multiple outcomes and heterogeneity effects have been presented in the context of sound-wellbeing modeling, they are modeling challenges encountered in a wider set of applications employing multilevel models. Our proposed methods can be used for applications such as patient monitoring systems, military fitness management programs, smart diet applications, etc., which have multilevel streaming data. In this paper, we consider a two-level structure with variables varying at a within-individual (level-1) and at an individual level (level-2). Nonetheless, the proposed methods can be applied to data with more than two levels of grouping structures without loss of generality. We propose the following guidelines for applying the statistical modeling methods to future applications:

(a) **Modeling curvilinear relationships:** Curvilinear relationships can be explored using scatter plot visualizations and validated using second-order coefficients in models. In cases where curvilinear relationships of more than one input are required to be modeled while determining change points for each corresponding input, other inputs should be retained as smooth functions in the model.

(b) **Simultaneous modeling of multiple outcomes:** We have conducted simultaneous modeling of two outcomes using a single latent variable (i.e., physiological wellbeing) in our study. However, there can be theoretical justification to model outcomes using two or more latent variable constructs. Future studies can extend our method using parameter expansion (Merkle and Wang 2016) or other suitable approaches for including multiple latent variables in the hierarchical Bayesian latent variable model.

(c) **Modeling heterogeneity in effects:** The varying-coefficient modeling method identifies person-level factors contributing to individual heterogeneity in effects, therefore, sufficient person-level information should be collected before the analysis. The voting method using three regularized models can be replaced by other feature selection methods based on specific business needs.

4.7.3. Managerial implications

Understanding and managing the workplace environment have organization-wide cost and benefit implications. For example, companies trying to increase communication and interaction among employees, with additional benefits such as cost savings and higher activity levels (Lindberg et al. 2018), opt for an open-office design. Nevertheless, these efforts have been shown to have negative effects in terms of loss of privacy and focus due to higher sound levels (Kim and de Dear 2013; Mak and Lui 2012). We posit that considering workplace sound level exposure in design decisions is important, for it is related to the physiological wellbeing of individuals. We determined that short term sound exposure around 50 dBA is optimal for the health and wellbeing of white-collared office workers. That is, workplaces that encourage soft conversations are better for employees than extremely quiet or noisy workplaces. For sound amplitudes lower than 50 dBA, a 10 dBA increase in sound level is related to a 3.6% increase in physiological wellbeing. For sound amplitudes higher than 50 dBA, a 10 dBA increase in sound level is related to decrease in physiological wellbeing by 1.3%. We also show that the effects of sound vary across worker population. In particular, workers with high blood pressure and involved in computer dominant work are more sensitive to ambient sound levels than others. Since a significant part of our lives is spent in the workplace (Bureau of Labor Statistics 2017), relatively modest changes to our short-term physiological wellbeing due to workplace environment factors such as the ambient sound can impact our long-term health and wellbeing due to our prolonged exposure to them. Organizations

should consider the effects of sound as part of a broader design plan to improve employee wellbeing through workplace design. This includes making better workplace design decisions regarding the following: number of workstations in a given area, types of workstations (e.g., private office, cubicle, open bench seating), seating area per workstation, distance from nearest window, partition height between workstations, nature of work (e.g., customer care, analysis, creative, etc.), etc. Workers are encouraged to follow workplace etiquette aimed at moderating ambient sound levels (e.g., taking calls from personal workstations, maintaining quieter peer conversations, etc.). This can help the organization by promoting overall satisfaction levels as well as the health and wellness of workers. Workers who report having high blood pressure should have provisioning to opt for quiet workplaces. Workers who spend a majority of their time working on their computers should be informed about health benefits of choosing workplaces that are not too quiet or too noisy.

4.7.4. Study limitations

In this study, we have focused on modeling the effects of workplace sound levels on the physiological wellbeing of office workers, but we have not collected information about the sound types (e.g., conversation, mechanical background noise, etc.) and frequencies (e.g., low, speech, high-tone, etc.) due to individual privacy concerns and sensor technology limitations. Since office sound type and frequency do not moderate the effects of sound level on physiological wellbeing outcomes (Sun Sim et al. 2015; Walker et al. 2016), we believe our findings will still hold when controlling for the type and frequency of ambient sounds. Future studies can focus on the effects of sound type and sound frequency and use our study design and modeling methods. Secondly, we have aggregated the sound level and other level-1 variables at 5-minute intervals to match the grain of short-term physiological wellbeing measures SDNN and normalized-HF; since the latter cannot be determined meaningfully for grains finer than 5 minutes (Kleiger et al. 2005; Pereira et al. 2017;

Xhyheri et al. 2012). Therefore, the lasting effects of spikes in sound level due to sudden events (e.g., shouting, breaking glass, etc.) have not been investigated. Nevertheless, the effects of events repeated multiple times as well as background noises consistent across the five-minute interval are accounted for in our models. The Bayesian approach used in this paper for simultaneous modeling of outcomes and modeling heterogeneity effects leads to better performance when compared to the classical (i.e., frequentist) approaches at the cost of computing power and time. Posterior estimation of parameters in Bayesian models take much more time than estimation of coefficients in classical models. This limitation can be partially overcome by using parallel processing and high-performance computing clusters to estimate parameters of the Bayesian models.

4.8. Conclusion

The majority of the working population spend a significant part of their active hours in office workspaces. Research shows that the workplace environment has an impact on an individual's work-related stress and wellbeing. Among the workplace environment factors, the ambient sound level is reported to be one of the highest stressors. We conducted a novel field study using wearable devices, where participants carried out their day-to-day activities wearing sensors that continuously recorded their physiological wellbeing state and ambient sound level. We fitted multilevel regression models to the resultant data and observed evidence of a relationship between sound level and two physiological wellbeing measures (i.e., SDNN, normalized-HF). To better understand the mechanism of the sound-wellbeing relationship, we proposed three new statistical modeling methods for representing curvilinear effects, simultaneous modeling of multiple outcomes, and identifying factors contributing to individual heterogeneity effects. Our methods have better predictive performance than existing methods for each of the three modeling problems. Using our methods, we infer that on average an individual's physiological wellbeing is optimal when sound level in the workplace is 50 dBA. Age, body-mass-index, high blood pressure, anxiety,

and computer use intensive work are person level factors that contribute to heterogeneity in sound level effects on physiological wellbeing across our study population. Our modeling method shows that workers with higher blood pressure are more negatively affected by an increase in sound levels, whereas workers with computer intensive work are more negatively affected by sound level extremities (i.e., quietude and loud noise). Our study informs policies and practices for designing workplaces with optimal sound levels in general and customized for certain subsets of the office worker population. It proposes new quantitative modeling methods to facilitate the advancement of IS in BI&A 3.0 through meaningful use of sensor-based content.

Moving forward, we plan to extend our analysis to determine the effects of other indoor environment quality factors including temperature, CO_2, relative humidity, and light intensity on the physiological wellbeing of office workers. We also plan to validate our findings using additional (causal) experiments for specific target groups and discrete values of sound levels identified in this study. Finally, we plan to investigate other factors contributing to the heterogeneity in sound level effects on physiological wellbeing such as extent of social interaction interactions between office-workers and average sound level exposure during leisure time.

5. CONCLUSIONS

5.1. Dissertation summary

Analytics is showing more and more promise in tackling problems in healthcare with advancements in big data and digital technology. Health analytics is one of the most fastest growing disciplines today with research opportunities in a wide range of topics. Innovative methods are required for processing and analyzing the complex data generated in health analytics applications. My dissertation consists of three essays that introduces a collection of novel quantitative methods to addressing specific challenges in analytics when applied to the health care application domains of digital health and preventive care (DHPC). Essay 1 introduces a new method of feature engineering using disease co-occurrence networks to predict high-cost patients at point of admission with limited input features available. The high-cost patient encounter prediction at the point-of-admission (HPEPP) model that is proposed in this essay has the potential to improve targeted care management and reduce health care expenditures. Essay 2 proposes a new method to analyze incomplete data with block-wise missing patterns by automatically identifying patterns in the dataset that can be utilized to train multiple reduced models to help minimize imputation and loss of information. It is simple, efficient, and scalable for large incomplete datasets from a variety of domains that contain block-wise missing values. It is effective in explanatory as well as predictive modeling and can be used in design science applications with data exhibiting block-wise missing patterns. Essay 3 presents a digital health analytics problem using a multi-sensor field study. It introduces methods to represent curvilinear relationships, simultaneously model multiple outcomes, and model heterogeneity in effects of sound level on wellbeing in the workplace. These methods not only contribute to digital health

analytics, but also facilitate the advancement of IS in BI&A 3.0 through meaningful use of sensor-based content. This study informs policies and practices for designing workplaces with optimal sound levels as well as identifies certain subsets of the worker population with higher vulnerability to office sound levels.

To summarize, this dissertation focuses on developing novel predictive and explanatory modeling methods to address problems in DHPC analytics. These methods tackle the following challenges – making early prediction using limited predictors, analyzing incomplete data with block-wise missing patterns, modeling curvilinear relationships, modeling multiple outcomes simultaneously, and identifying factors related to individual heterogeneity in effects. These methods address important modeling challenges in DHPC analytics and make broad contributions to business analytics, health IS, and design science research.

5.2. Relevance and future research

Data science, business intelligence, machine learning, artificial intelligence, big data analytics, and other related data analysis paradigms of the 21st century are distinguished by the underlying treatment to the data (i.e., querying, mining patterns, quantitative modeling, etc.). Analytics (or data science) is defined as the process of extracting knowledge from data and can be considered as the superset of all the above listed data analysis paradigms. Analytics methods can be grouped in different ways – regression-classification-optimization, structured-unstructured, supervised-unsupervised, shallow-deep, static-dynamic, descriptive-associative-causal, querying-mining-modeling, or explanatory-predictive – according to the underlying problems they address. Health analytics has expanded in depth and reach due to the availability of massive amounts of public data and advancement in big data technologies. Smart wearable sensors and cellphone-based tracking technology can be used to monitor, analyze and improve individual health and wellbeing. Digital footprints on social media can be used to gather insights on mental and physical

health of populations. Newer preventive care analytics applications include high-risk patient prediction, disease prediction, disease progression modeling, drug interaction analysis, drug adherence analysis, cumulative dosing analysis, lifestyle-health analysis. It is also worth exploring the possibility of combining digital health and preventive care technologies to enhance human wellbeing. Electronic health records contain information about disease onset and health problems while digital health data contain lifestyle and wellbeing related information. Fusing information from both sources can help to better understand the inter-relationship between lifestyle and diseases. Such a rich fusion of data can be useful for developing advanced pre-emptive care and diagnostics. In this way, DHPC offers unlimited potential for innovations in analytics methods. Robust and efficient methods for DHPC analytics can facilitate the optimal use of digital health information generated every day to support preventive care practices and improve the overall health and wellbeing of the society.

In my dissertation, I have focused on challenges in explanatory and predictive modeling for DHPC applications. Future research can focus on addressing other challenges in DHPC analytics including modeling lagged effects, modeling with multi-collinear features, fusing heterogeneous information, interpreting complex black-box models (e.g., deep learning), causal modeling for big data, providing local explanations for complex models, etc.

6. APPENDIX

6.1. Noise sources and their sound levels (Essay 3)

Table 6.1: Noise sources and sound levels[3]

Noise Source	Decibel Level (dBA)	Decibel Effect
Jet take-off (at 25 meters)	150	Eardrum rupture
Aircraft carrier deck	140	
Military jet aircraft take-off from aircraft carrier with afterburner at 50 ft (130 dBA).	130	
Thunderclap, chain saw. Oxygen torch (121 dBA).	120	Painful. 32 times as loud as 70 dBA.
Steel mill, auto horn at 1 meter. Turbo-fan aircraft at takeoff power at 200 ft (118 dBA). Riveting machine (110 dBA); live rock music (108 - 114 dBA).	110	Average human pain threshold. 16 times as loud as 70 dBA.
Jet take-off (at 305 meters), use of outboard motor, power lawn mower, motorcycle, farm tractor, jackhammer, garbage truck. Boeing 707 or DC-8 aircraft at one nautical mile (6080 ft) before landing (106 dBA); jet flyover at 1000 feet (103 dBA); Bell J-2A helicopter at 100 ft (100 dBA).	100	8 times as loud as 70 dBA. Serious damage possible in 8-hour exposure.
Boeing 737 or DC-9 aircraft at one nautical mile (6080 ft) before landing (97 dBA); power mower (96 dBA); motorcycle at 25 ft (90 dBA). Newspaper press (97 dBA).	90	4 times as loud as 70 dBA. Likely damage in 8-hour exposure.
Garbage disposal, dishwasher, average factory, freight train (at 15 meters). Car wash at 20 ft (89 dBA); propeller plane flyover at 1000 ft (88 dBA); diesel truck 40 mph at 50 ft (84 dBA); diesel train at 45 mph at 100 ft (83 dBA). Food blender (88 dBA); milling machine (85 dBA); garbage disposal (80 dBA).	80	2 times as loud as 70 dBA. Possible damage in 8-hour exposure.
Passenger car at 65 mph at 25 ft (77 dBA); freeway at 50 ft from pavement edge 10 a.m. (76 dBA). Living room music (76 dBA); radio or TV-	70	Arbitrary base of comparison. Upper 70s are annoyingly loud to

[3] Source: IAC Acoustic website: http://www.industrialnoisecontrol.com/comparative-noise-examples.htm

audio, vacuum cleaner (70 dBA).		some people.
Conversation in restaurant, office, background music, Air conditioning unit at 100 feet.	60	Half as loud as 70 dBA. Fairly quiet.
Quiet suburb, conversation at home. Large electrical transformers at 100 feet.	50	One-fourth as loud as 70 dBA.
Library, bird calls (44 dBA); lowest limit of urban ambient sound	40	One-eighth as loud as 70 dBA.
Quiet rural area.	30	One-sixteenth as loud as 70 dBA. Very Quiet.
Whisper, rustling leaves	20	
Breathing	10	Barely audible

6.2. Multilevel model inference using Classical and Bayesian approaches (Essay 3)

Multilevel or hierarchical levels of grouped data are a commonly occurring phenomenon (Raudenbush and Bryk 2002). For example, in organizational studies, information about firms as well as workers are available such that there exists a hierarchical structured data of individual workers nested within multiple firms. Multilevel models, also called as hierarchical linear models, random coefficients models, mixed-effects models, are statistical models with parameters that capture variability across multiple levels of data.

In the classical or frequentist approach, multilevel models can be considered as an extension of an ordinary least squares (OLS) regression model used to analyze variance in the outcome variables when the predictor variables are at varying hierarchical levels. A two-level hierarchical linear model can be mathematically expressed as follows:

$$\text{Level 1:} \qquad Y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} V_{kij} + r_{ij} \tag{6.1}$$

$$\text{Level 2:} \qquad \beta_{kj} = \gamma_{k0} + \sum_{m=1}^{M} \gamma_{km} W_{mj} + u_{kj} \tag{6.2}$$

where Y_{ij} is the outcome, $\beta_{kj}$are the level-1 coefficients, V_{kij} are level-1 input variables, r_{ij} are level-1 residuals, γ_{km} are level-2 coefficients, W_{mj} are level-2 input variables and u_{kj} are level-2 variables for i^{th}observation of j^{th}individual for $k \in \mathbb{Z}_K$ and $m \in \mathbb{Z}_M$. The assumptions for the model are as follows:

$$E(r_{ij}) = 0; var(r_{ij}) = \sigma^2;\ E(u_{kj}) = 0; cov(u_{kj}, r_{ij}) = 0\ \forall i, j, k;\ \begin{bmatrix} u_{11} & \cdots \\ \cdots & u_{kj} \end{bmatrix} = T \tag{6.3}$$

where T is the level-2 variance covariance component that model the inter-relationship between level-2 errors. Combining equations (1) and (2), we can represent hierarchical linear models as follows:

$$y_{ij} = \beta_0 + \gamma_{0j} + \sum_{k=1}^{K} \beta_k x_{kij} + \sum_{m=1}^{M} \gamma_{mj} z_{mij} + \epsilon_{ij} \tag{6.4}$$

where $\beta = \{\beta_0, \beta_1, \ldots, \beta_K\}$ are fixed effects coefficients, $\gamma = \{\gamma_{0j}, \gamma_{1j}, \ldots, \gamma_{Mj}\}$ are random-effects coefficients for J groups $j \in \mathbb{Z}_J$, and ϵ_{ij} is the sum of fixed-effects error and random-effects error components. In matrix notation, the above equation is represented as follows:

$$Y = \alpha + X\beta + Z\gamma + \epsilon \tag{6.5}$$

where X is a matrix of fixed effects, and Z is a matrix of random effects. Conditional to the above assumptions, the parameters in the model can be estimation by maximizing the likelihood function y as shown below:

$$y \sim N(\alpha + X\beta, \sigma^2 I + Z'TZ) \tag{6.6}$$

The significance of the fixed effects and random effects are tested using Wald test, Likelihood Ratio Test, F-test, parametric bootstrap or MCMC methods (Raudenbush and Bryk 2002). Model fit can be compared using AIC, deviance and R-squared approximations (Nakagawa and Schielzeth 2013).

Bayesians, on the other in-hand, describe their beliefs about the unknows in a hierarchical linear model before observing data with prior distributions and the following likelihood function:

$$y \sim N(\alpha + X\beta + Zb, \sigma^2 I) \tag{6.7}$$

A single level regression disregards between-group heterogeneity is called model with complete pooling and can yield parameter estimates that are wrong if there is between-group heterogeneity. On the other hand, regression models for each group of the level-2 data independently are called modeling with no pooling and result imprecise parameter estimates, for they ignore common

variance across groups. Hierarchical linear models are considered as a subset of Hierarchical Bayesian models that are models with partial pooling (Gelman and Hill 2007). Parameters are allowed to vary by group at lower levels of the hierarchy while estimating common parameters at higher levels. Note that the level-2 and higher effects are not part of the error variance as in the classical/frequentist approach but modeled as parameters themselves (also called varying coefficients). The varying parameters have hyper-parameters that are estimated based on level-2 and higher order grouping in the data. The estimated posterior distribution of parameters for a hierarchical linear model with normally distributed error and identity link function has the following form:

$$p(\alpha, \beta, \gamma, \sigma_Y, \sigma_\gamma | Y, X, Z, U) \propto \tag{6.8}$$

$$\prod_{j=1}^{J} \prod \prod_{i=1}^{n_j} N(Y| \ \beta_0 + \gamma_{0j} + \beta X + \gamma_j Z, \sigma_Y^2) \prod \prod_{j=1}^{J} N(\gamma_{0j}, \gamma_j | \alpha_0 + \alpha U, \ \sigma_\gamma^2)$$

MCMC estimation approaches such as Metropolis Hastings, Gibbs Sampling, and Hamiltonian Monte Carlo families of methods are used to estimate the posterior probability given the prior distribution of all parameters and likelihood of given data (Gelman et al. 2014). Comparison of implementations and general purpose software packages for classical and Bayesian multilevel modeling is done in West and Galecki (2011), Mai and Zhang (2018) respectively.

6.3. Univariate transformation-based modeling method (Essay 3)

In the univariate transformation-based modeling method, we first convert the data into a long format as shown in Figure 6.1.

Row no.	Person ID	Outcome 1	Outcome 2
1	1	36.48	10.05
2	1	37.12	7.33
3	1	60.83	13.68
4	2	38.99	7.11
5	2	73.80	10.75
6	3	48.02	12.22
…	…	…	…

Row no.	Person ID	Outcome label	Value
1	1	Outcome 1	36.48
2	1	Outcome 1	37.12
3	1	Outcome 1	60.83
4	2	Outcome 1	38.99
5	2	Outcome 1	73.80
6	3	Outcome 1	48.02
7	1	Outcome 2	10.05
8	1	Outcome 2	7.33
9	1	Outcome 2	13.68
10	2	Outcome 2	7.11
11	2	Outcome 2	10.75
12	3	Outcome 2	12.22
…	…	…	…

Figure 6.1: Converting data in wide format to long format for univariate representation of multivariate outcomes

Equation resembles the equation for multilevel model for univariate outcome, except that each variable is subscripted with h which indicates the value for the h^{th} outcome.

$$y_{hij} = \beta_0 + \gamma_{0j} + \sum_{k=1}^{K} \beta_k x_{hkij} + \sum_{m=1}^{M} \gamma_{mj} z_{hmij} + \epsilon_{ij}^{(h)} \tag{6.9}$$

Here, residual errors, $\epsilon_{ij}^{(h)}$, are defined as $N(0, \sigma_h^2)$ with estimated independent error variances for each outcome h. Marginal contributions of any outcome on input variables can be derived by including an indicator function in an interaction effect. For example, to derive marginal effects of level-1 variables, model is shown as follows:

$$y_{hij} = \beta_0 + \gamma_{0j} + \sum_{k=1}^{K} \beta_k x_{hkij} \cdot I(h) + \sum_{m=1}^{M} \gamma_{mj} z_{hmij} + \epsilon_{ij}^{(h)} \tag{6.10}$$

where $I(\varphi)$ is an indicator function equal to 1 if condition φ is true, otherwise it is 0.

6.4. Information collected in wellbuilt for wellbeing study (Essay 3)

Table 6.2: Description of all the information collected during the wearable sensors-based field study

Variable name	Data source	Description
P_ID	Participant log	Identification number unique to each participant
cohort	Participant log	Which study cohort the participant belongs to
Timestamp	Movisens HRV monitor	Observation timestamp aggregated to a 5-minute grain
norm. LF	Movisens HRV monitor	Normalized low frequency component of heart rate
norm. HF	Movisens HRV monitor	Normalized high frequency component of heart rate
LF/HF	Movisens HRV monitor	Ratio of low and high frequency components of heart rate
SDNN	Movisens HRV monitor	Standard deviation between N-N beat intervals of heart rate
RMSSD	Movisens HRV monitor	Root mean squared standard deviation of N-N beat interval of heart rate
Activity	Movisens HRV monitor	Activity state of participants measured using an accelerometer
Wearnode_Temperature	Wearnodes IEQ sensor	Temperature measured in Celsius using the wearable sensor
Wearnode_Sound	Wearnodes IEQ sensor	Sound leve measured in dBA using the wearable sensor
Wearnode_Relative_humidity	Wearnodes IEQ sensor	Relative humidity measured in % of volume in air using the wearable sensor
Wearnode_CO2	Wearnodes IEQ sensor	CO2 measured in ppm using the wearable sensor
Wearnode_Pressure	Wearnodes IEQ sensor	Pressure measured in bars using the wearable sensor
Wearnode_Light	Wearnodes IEQ sensor	Light intensity measured in lux using the wearable sensor
Wearnode_Absolute_humidity	Wearnodes IEQ sensor	Absolute humidity measured in grams of water vapor per cubic meter volume of air using the wearable sensor
Wall_co	Wall mounted IEQ sensors	Carbon monoxide measured in ppm using wall mounted sensor
Wall_co2	Wall mounted IEQ sensors	CO2 measured in ppm using wall mounted sensor

Wall_humidity	Wall mounted IEQ sensors	Carbon monoxide measured in ppm using wall mounted sensor
Wall_light	Wall mounted IEQ sensors	Light intensity measured in lux using wall mounted sensor
Wall_pressure	Wall mounted IEQ sensors	Pressure measured in bars using wall mounted sensor
Wall_temperature	Wall mounted IEQ sensors	Temperature measured in celcius using wall mounted sensor
Wall_sound	Wall mounted IEQ sensors	Sound level measured in dBA using wall mounted sensor
Wall_pm	Wall mounted IEQ sensors	Particulate matter (2.5 Micron) measured in ppm using wall mounted sensor
Wall_tvoc	Wall mounted IEQ sensors	Total volatile organic compounds measured using wall mounted sensor
space	Experience sampling surveys	Which location the participant is in while answering the experience sampling survey
what_been_doing	Experience sampling surveys	Which task the participant is doing while answering the experience sampling survey
who_interact_with	Experience sampling surveys	With whom is the participant interacting with (colleague, supervisor, client, alone) while answering the experience sampling survey
tense	Experience sampling surveys	How tense participant is feeling on a scale between 1-7 while answering the experience sampling survey
content	Experience sampling surveys	How content participant is feeling on a scale between 1-7 while answering the experience sampling survey
sad	Experience sampling surveys	How sad participant is feeling on a scale between 1-7 while answering the experience sampling survey
alert	Experience sampling surveys	How alert participant is feeling on a scale between 1-7 while answering the experience sampling survey
tired	Experience sampling surveys	How tired participant is feeling on a scale between 1-7 while answering the experience sampling survey
happy	Experience sampling surveys	How happy participant is feeling on a scale between 1-7 while answering the experience sampling survey
upset	Experience sampling surveys	How upset participant is feeling on a scale between 1-7 while answering the experience sampling survey
calm	Experience sampling surveys	How calm participant is feeling on a scale between 1-7 while answering the experience sampling survey
focused	Experience sampling surveys	How focused participant is feeling on a scale between 1-7 while answering the experience sampling survey
productive	Experience sampling surveys	How productive participant is feeling on a scale between 1-7 while answering the experience sampling

		survey
engaged	Experience sampling surveys	How engaged participant is feeling on a scale between 1-7 while answering the experience sampling survey
pleasant	Experience sampling surveys	How pleasant participant is feeling on a scale between 1-7 while answering the experience sampling survey
demanding	Experience sampling surveys	How demanding participant is feeling on a scale between 1-7 while answering the experience sampling survey
noise	Experience sampling surveys	How the participant is feeling about the environment's noise level on a scale between 1-7 while answering the experience sampling survey
light	Experience sampling surveys	How the participant is feeling about the environment's light intensity on a scale between 1-7 while answering the experience sampling survey
temperature	Experience sampling surveys	How the participant is feeling about the environment's temperature level on a scale between 1-7 while answering the experience sampling survey
air	Experience sampling surveys	How the participant is feeling about the environment's air quality level on a scale between 1-7 while answering the experience sampling survey
caffeine	Experience sampling surveys	How many cups of coffee did the participant consume during the last one hour
cigarette	Experience sampling surveys	How many cigarettes did the participant smoke during the last one hour
ToD	Derived	Time of day discretized to six levels
DoW	Derived	Day of week
month	Derived	Month of the year
Gender	Intake Survey	Gender of the participant
AntiDepression	Intake Survey	Is the participant taking anti-depression medication
SleepingAid	Intake Survey	Is the participant taking sleeping aid medication
CardioVascularMedicine	Intake Survey	Is the participant taking cardiovascular medication
Ethnicity	Intake Survey	Ethnicity of participant
Age	Intake Survey	Age of participant
Experience	Intake Survey	Experience of the participant within the organization
Qualification	Intake Survey	Qualification of the participant
BMI	Intake Survey	Body mass index of the participant
HighBP	Intake Survey	Does the participant suffer from high blood pressure
Heart.disease	Intake Survey	Does the participant suffer from heart disease
Musculoskeletal.	Intake Survey	Does the participant suffer from musculoskeletal problems

Depression	Intake Survey	Does the participant suffer from depression
Anxiety	Intake Survey	Does the participant suffer from anxiety
Sleep.problems	Intake Survey	Does the participant suffer from sleep problems
Pain_neck_knee	Intake Survey	Does the participant suffer from pain in the neck or knee
Average_alcohol_intake	Intake Survey	Amount of alcohol intake each week
Smoker	Intake Survey	Is the participant a smoker
Extraversion	Intake Survey	Big-5 personality trait of participant on a scale between 1 and 7
Agreeableness	Intake Survey	Big-5 personality trait of participant on a scale between 1 and 7
Conscientiousness	Intake Survey	Big-5 personality trait of participant on a scale between 1 and 7
Neuroticism	Intake Survey	Big-5 personality trait of participant on a scale between 1 and 7
Openness	Intake Survey	Big-5 personality trait of participant on a scale between 1 and 7
Noise_sensitivity	Intake Survey	Noise sensitivity of participant on a scale between 1 and 7
No.night	Sleep data	No of nights data available for the participant
start_time_in_Bed	Sleep data	Start time in bed
stop_time_in_Bed	Sleep data	Time when got out of bed
Day	Sleep data	Day 1, 2 or 3 of participant
Sleephours	Sleep data	Number of hours of sleep averaged over three days
IndividualC_windowShades.x	Spatial characteristics of workstation	If participant has access over Window shades controls
IndividualC_windowAC.x	Spatial characteristics of workstation	If participant has access over A/C controls
group	Spatial characteristics of workstation	Which task group participant belongs to
Base_WS	Spatial characteristics of workstation	The workstation assigned to participant
Floor	Spatial characteristics of workstation	Floor of participants workstation
Wing	Spatial characteristics of workstation	Wing of participants workstation
Building	Spatial characteristics of workstation	Building of participants workstation
StudyWeek	Spatial characteristics of workstation	Study week (not participant specific)

Office_type	Spatial characteristics of workstation	Type of office (open office, cubicle, private office)
Perimeter_core	Spatial characteristics of workstation	Core perimeter of workstation of participant
Distance_window	Spatial characteristics of workstation	Distance from window for participant's workstation
Window_view_sit	Spatial characteristics of workstation	Does the participant's workstation have a window view when sitting
Window_view_stand	Spatial characteristics of workstation	Does the participant's workstation have a window view when standing
Nature_view	Spatial characteristics of workstation	Does the participant's workstation have a naturel view
ThermalzoneSize	Spatial characteristics of workstation	What is the thermal zone size of the workstation area
PerimeterSystem	Spatial characteristics of workstation	What is the perimeter system of the workstation area
CoreHVACControls	Spatial characteristics of workstation	What are the core HVAC controls of the workstation area
Diffuser.density	Spatial characteristics of workstation	What is the diffuser density of the workstation area
SeasonalSwitchover	Spatial characteristics of workstation	Is there a seasonal switchover system available for workstation area
Heating_CoolingSeason	Spatial characteristics of workstation	Is it a heating season (winter) or cooling season (summer)
ReturnAirDensity	Spatial characteristics of workstation	What is the return air density of the workstation area
DedicatedExhausts	Spatial characteristics of workstation	Are there dedicated exhausts in the workstation area
Ceiling.fixture.type	Spatial characteristics of workstation	What is the ceiling fixture type of the workstation area
height	Spatial characteristics of workstation	What is the ceiling height of the workstation area
Ceiling.light.lens.type	Spatial characteristics of workstation	What is the ceiling light lens type of the workstation area
Level.of.ceiling.light.control	Spatial characteristics of workstation	What is the level of ceiling lights of the workstation area
Ceiling_lightControl	Spatial characteristics of workstation	Is there a ceiling light control for participant's workstation area
X..seated.views	Spatial characteristics of workstation	Percentage of seated views in workstation area
Per_seated_coding	Spatial characteristics	Number of seats in workstation area

	of workstation	
Distribution.of.open.closed.spaces	Spatial characteristics of workstation	Distribution of open/closed workspaces
ceiling.quality	Spatial characteristics of workstation	Type of ceiling
Floor.quality...Open.Office.	Spatial characteristics of workstation	type of floor (carpet, tiles, etc.)
Floor.quality...New...Corridor.	Spatial characteristics of workstation	Type of floor in nearby corridor
Size.density.of..workstations..net.	Spatial characteristics of workstation	Size of workstation (net)
Size.density.of.open.workstations..gross.	Spatial characteristics of workstation	Size of workstation (gross)
Partition.height	Spatial characteristics of workstation	Height of partition between adjacent workstations
number.of.sides	Spatial characteristics of workstation	Number of sides of partitions
HVAC.noise	Spatial characteristics of workstation	HVAC noise level
Masking.sound	Spatial characteristics of workstation	Is there sound masking available
Speaker.Phones.Y.N	Spatial characteristics of workstation	Are there speaker phones in workstation
Main.Corridor.Location...Within.Workstation.Area.Y.N	Spatial characteristics of workstation	Main corridor location within workstation
Available.Quiet.Spaces.Y.N	Spatial characteristics of workstation	Are quiet spaces available in workstation area
Presence.of.Informal.Meeting.Areas.Y.N	Spatial characteristics of workstation	Are informal meeting areas available within workstation area
steps	Sitting bout and derived activity data	Number of steps taken during a time period (e.g., in an hour)
Activity and bouts information	Sitting bout and derived activity data	Is the participant sitting, standing, walking or laying at a given time
PSQI	Sitting bout and derived activity data	Pittsburg sleep quality index - Derived based on questionnaire prompts
SDNN_baseLine	Sitting bout and derived activity data	Base stress level of participant at the beginning of the day
Manage/Supervise	Work description data	Does the participant have a manage/supervise job role
Technical/Professional	Work description data	Does the participant have a technical/professional job role

Administrative/Support	Work description data	Does the participant have an administrative/support job role
Other	Work description data	Job role not among the ones listed
Contracting	Work description data	Does the participant have a contracting job role
Collaborate/Meeting.Heavy	Work description data	Does the participant have a collaborate/meeting heavy job role
Analysis	Work description data	Does the participant have an analysis job role
Computer/Office.Dominant	Work description data	Does the participant have a computer/office dominant job role
Workcode	Work description data	What is the work code of the participant
short.description.of.work	Work description data	Description of work as reported by participant

7. REFERENCES

Abbasi, A., Sarker, S., and Chiang, R. 2016. "Big Data Research in Information Systems: Toward an Inclusive Research Agenda," *Journal of the Association for Information Systems*.

Abrams, S., and Hens, N. 2015. "Modeling Individual Heterogeneity in the Acquisition of Recurrent Infections: An Application to Parvovirus B19," *Biostatistics*.

Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., and Suri, J. S. 2006. "Heart Rate Variability: A Review," *Medical and Biological Engineering and Computing*, pp. 1031–1051.

Agarwal, R., and Dhar, V. 2014. "Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research," *Information Systems Research* (25:3), pp. 443–448.

Agarwal, R., Gao, G., DesRoches, C., and Jha, A. K. 2010. "Research Commentary-The Digital Transformation of Healthcare: Current Status and the Road Ahead," *Information Systems Research* (21:4), pp. 796–809. (http://pubsonline.informs.org809.https//doi.org/10.1287/isre.1100.0327http://www.informs.org).

Ahsen, M. E., Ayvaci, M. U. S., and Raghunathan, S. 2018. "When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis," *Information Systems Research*, INFORMS, pp. 1–20.

Ash, A. S., Zhao, Y., Ellis, R. P., and Schlein Kramer, M. 2001. "Finding Future High-Cost Cases: Comparing Prior Cost versus Diagnosis-Based Methods," *Health Services Research* (36), pp. 194–206.

Backé, E. M., Seidler, A., Latza, U., Rossnagel, K., and Schumann, B. 2012. "The Role of Psychosocial Stress at Work for the Development of Cardiovascular Diseases: A Systematic

Review," *International Archives of Occupational and Environmental Health*.

Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., and Atkins, D. C. 2014. "Analyzing Multiple Outcomes in Clinical Research Using Multivariate Multilevel Models," *Journal of Consulting and Clinical Psychology* (82:5), pp. 920–930.

Barabasi, A., and Frangos, J. 2014. *Linked: The New Science Of Networks Science*, Perseus Books Group.

Bardhan, I., Oh, J., Zheng, Z., and Kirksey, K. 2015. "Predictive Analytics for Readmission of Patients with Congestive Heart Failure," *Information Systems Research* (26:1), pp. 19–39.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A., and Escobar, G. 2014. "Big Data in Health Care: Using Analytics to Identify and Manage High-Risk and High-Cost Patients," *Health Affairs* (33:7), pp. 1123–1131.

Batista, G., and Monard, M. C. 2003. "An Analysis of Four Missing Data Treatment Methods for Supervised Learning," *Applied Artificial Intelligence* (17:5–6), Taylor & Francis Group, pp. 519–533.

Becker, J.-M., Rai, A., Ringle, C. M., and Völckner, F. 2013. "Discovering Unobserved Heterogeneity in Structural Equation Models to Avert Validity Threats," *MIS Quarterly* (37:3), Society for Information Management and The Management Information Systems Research Center, pp. 665–694.

Bertsimas, D., Bjarnadóttir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., and Wang, G. 2008. "Algorithmic Prediction of Health-Care Costs," *Operations Research* (56:6), INFORMS, pp. 1382–1392.

Blondel, V. D., Guillaume, J. L., Lambiotte, R., and Lefebvre, E. 2008. "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment* (10), p. 6.

Boron, W. F., and Boulpaep, E. L. 2012. "Medical Physiology," *Physiology*.

Breiman, L. 1999. "Random Forest," *Machine Learning* (45:5), pp. 1–35.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. 1984. "Classification and Regression Trees," *The Wadsworth Statistics/Probability Series* (Vol. 19).

Brent, R. P. 2013. *Algorithms for Minimization without Derivatives*, Courier Corporation.

Bureau of Labor Statistics. 2017. "Average Weekly Hours and Overtime of All Employees on Private Nonfarm Payrolls by Industry Sector, Seasonally Adjusted." (https://www.bls.gov/news.release/pdf/atus.pdf).

Bureau of Public Health Statistics. 2017. *Arizona Hospital Discharge Public Use Files*, Arizona Department of Health Services. (http://www.azdhs.gov/preparedness/public-health-statistics/hospital-discharge-data/index.php).

Buuren, S. Van. 2012. *Flexible Imputation of Missing Data*, CRC press.

Van Buuren, S., and Groothuis-Oudshoorn, K. 2011. "Multivariate Imputation by Chained Equations," *Journal Of Statistical Software* (45:3), pp. 1–67.

Califf, R. M., and Pencina, M. J. 2013. "Predictive Models in Heart Failure: Who Cares?," *Circulation. Heart Failure* (6:5), pp. 877–878.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. 2017. "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*.

CDC Foundation. 2018. "Business Pulse: Healthy Workforce | CDC Foundation." (https://www.cdcfoundation.org/businesspulse/healthy-workforce, accessed September 11, 2018).

census.gov. 2016. "Population Census Quickfacts." (http://www.census.gov/quickfacts/table/PST045215/00, accessed March 8, 2019).

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. 1987. "A New Method of Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation," *Journal of Chronic Diseases* (40:5), pp. 373–383.

Chechulin, Y., Nazerian, A., Rais, S., and Malikov, K. 2014. "Predicting Patients with High Risk of Becoming High-Cost Healthcare Users in Ontario (Canada)," *Healthcare Policy* (9:3), pp. 68–79.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data To Big Impact," *MIS Quarterly* (36:4).

cms.gov. 2016. "2015-2025 Projections of National Health Expenditures." (https://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2016-Press-releases-items/2016-07-13.html, accessed March 12, 2019).

Csárdi, G., and Nepusz, T. 2006. "The Igraph Software Package for Complex Network Research," *InterJournal Complex Systems* (1695), p. 1695.

Cvijanović, N., Kechichian, P., Janse, K., and Kohlrausch, A. 2017. "Effects of Noise on Arousal in a Speech Communication Setting," *Speech Communication* (88), pp. 127–136.

Das, A., Poole, W. K., and Bada, H. S. 2004. "A Repeated Measures Approach for Simultaneous Modeling of Multiple Neurobehavioral Outcomes in Newborns Exposed to Cocaine in Utero," *American Journal of Epidemiology*.

Dingemanse, N. J., and Dochtermann, N. A. 2013. "Quantifying Individual Variation in Behaviour: Mixed-Effect Modelling Approaches," *Journal of Animal Ecology*.

Divo, M. J., Casanova, C., Marin, J. M., Pinto-Plata, V. M., De-Torres, J. P., Zulueta, J. J., Cabrera, C., Zagaceta, J., Sanchez-Salcedo, P., Berto, J., Davila, R. B., Alcaide, A. B., Cote, C., and Celli, B. R. 2015. "COPD Comorbidities Network," *European Respiratory Journal* (46:3), pp. 640–650.

Durban, M., Harezlak, J., Wand, M. P., and Carroll, R. J. 2005. "Simple Fitting of Subject-Specific Curves for Longitudinal Data," *Statistics in Medicine* (24:8), pp. 1153–1167.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*, Cambridge University Press.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. 1998. "Comorbidity Measures for Use with Administrative Data," *Medical Care* (36:1), pp. 8–27.

Fanaee-T, H., and Gama, J. 2013. "Event Labeling Combining Ensemble Detectors and Background Knowledge," *Progress in Artificial Intelligence* (2), pp. 113–127.

Faraway, J. J. 2006. "Extending the Linear Model With R: Generalized Linear, Mixed Effects and Nonparametric Regression Models," *Journal of the American Statistical Association*, pp. 1–28.

Faraway, J. J. 2016. "Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models," *CRC Press*.

Fletcher, R. 2013. "Practical Methods of Optimization," *John Wiley & Sons* (53), p. 456.

Friedman, J. H., Kohavi, R., and Yun, Y. 1996. "Lazy Decision Tree," in *AAAI/IAAI, Vol. 1*, pp. 717–724.

Friedman, J., Hastie, T., and Tibshirani, R. 2001. "The Elements of Statistical Learning," *Springer Series in Statistics* (Vol. 1).

Friedman, N. 1997. "Learning Belief Networks in the Presence of Missing Values and Hidden Variables," *International Conference on Machine Learning*, pp. 125–133.

Frontczak, M., Schiavon, S., Goins, J., Arens, E., Zhang, H., and Wargocki, P. 2012. "Quantitative Relationships between Occupant Satisfaction and Satisfaction Aspects of Indoor Environmental Quality and Building Design," *Indoor Air* (22:2), pp. 119–31.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. 2014.

"Bayesian Data Analysis," *Bayesian Data Analysis*, CRC press.

Gelman, A., and Hill, J. 2007. "Data Analysis Using Regression and Multilevel/Hierarchical Models," *Cambridge*.

Geng, Y., Ji, W., Lin, B., and Zhu, Y. 2017. "The Impact of Thermal Environment on Occupant IEQ Perception and Productivity," *Building and Environment*.

Ghahramani, A., Pantelic, J., Lindberg, C., Mehl, M., Srinivasan, K., Gilligan, B., and Arens, E. 2018. "Learning Occupants' Workplace Interactions from Wearable and Stationary Ambient Sensing Systems," *Applied Energy* (230), Elsevier, pp. 42–51.

Gimenez, O., Cam, E., and Gaillard, J. M. 2018. "Individual Heterogeneity and Capture–Recapture Models: What, Why and How?," *Oikos* (127), pp. 664–686.

Graham, J. W. 2012. *Missing Data Analysis and Design*, Springer Science & Business Media.

Gregori, D., Petrinco, M., Bo, S., Desideri, A., Merletti, F., and 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), pp. 331–341.

Guillén, S., Arredondo, M. T., and Castellano, E. 2011. "A Survey of Commercial Wearable Systems for Sport Application," in *Wearable Monitoring Systems*.

Harvard School of Public Health. 2016. "The Workplace and Health." (https://news.harvard.edu/wp-content/uploads/2016/07/npr-rwjf-harvard-workplace-and-health-poll-report.pdf).

Hassan, M., Coulet, A., Toussaint, Y., Cnrs, L., and Nge, I. 2014. "Learning Subgraph Patterns from Text for Extracting Disease – Symptom Relationships," in *1st International Workshop on Interactions between Data Mining and Natural Language Processing*.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction," *Springer Series in Statistics*, The Mathematical

Intelligencer.

Hayes, S. L., Salzberg, C. A., McCarthy, D., Radley, D. C., Abrams, M. K., Shah, T., and Anderson, G. F. 2016. "High-Need, High-Cost Patients: Who Are They and How Do They Use Health Care?," *The Commonwealth Fund. 3rd Ed.* (Vol. 26), New York.

HCUP. 2013. "Nationwide Inpatient Sample," *Healthcare Cost and Utilization Project (HCUP)*. (http://www.hcup-us.ahrq.gov/, accessed January 16, 2017).

Heck, R. H., and Thomas, S. L. 2015. "An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches Using Mplus," *Hodder Arnold*, Routledge.

Heerwagen, J., and Zagreus, L. 2005. "The Human Factors of Sustainable Building Design: Post Occupancy Evaluation of the Philip Merrill Environmental Center," *Indoor Environmental Quality*.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75–105.

Hidalgo, C. A., Blumm, N., Barabási, A. L., and Christakis, N. A. 2009. "A Dynamic Network Approach for the Study of Human Phenotypes," *PLoS Computational Biology* (5:4).

Hox, J. J. 2013. "Multilevel Regression and Multilevel Structural Equation Modeling," *The Oxford Handbook of Quantitative Methods* (2), pp. 281–294.

Jahncke, H., Hygge, S., Halin, N., Green, A. M., and Dimberg, K. 2011. "Open-Plan Office Noise: Cognitive Performance and Restoration," *Journal of Environmental Psychology* (31:4), pp. 373–382.

Jiang, Z., Sarkar, S., De, P., and Dey, D. 2007. "A Framework for Reconciling Attribute Values from Multiple Data Sources," *Management Science* (53:12), pp. 1946–1963.

Jirschitzka, J., Kimmerle, J., and Cress, U. 2016. "A New Method for Re-Analyzing Evaluation Bias: Piecewise Growth Curve Modeling Reveals an Asymmetry in the Evaluation of pro

and Con Arguments," *PLoS ONE* (11:2).

Kim, J., and de Dear, R. 2013. "Workspace Satisfaction: The Privacy-Communication Trade-off in Open-Plan Offices," *Journal of Environmental Psychology* (36), Academic Press, pp. 18–26.

Kivimäki, M., Nyberg, S. T., Batty, G. D., Fransson, E. I., Heikkilä, K., Alfredsson, L., Bjorner, J. B., Borritz, M., Burr, H., Casini, A., Clays, E., De Bacquer, D., Dragano, N., Ferrie, J. E., Geuskens, G. A., Goldberg, M., Hamer, M., Hooftman, W. E., Houtman, I. L., Joensuu, M., Jokela, M., Kittel, F., Knutsson, A., Koskenvuo, M., Koskinen, A., Kouvonen, A., Kumari, M., Madsen, I. E. H., Marmot, M. G., Nielsen, M. L., Nordin, M., Oksanen, T., Pentti, J., Rugulies, R., Salo, P., Siegrist, J., Singh-Manoux, A., Suominen, S. B., Väänänen, A., Vahtera, J., Virtanen, M., Westerholm, P. J. M., Westerlund, H., Zins, M., Steptoe, A., and Theorell, T. 2012. "Job Strain as a Risk Factor for Coronary Heart Disease: A Collaborative Meta-Analysis of Individual Participant Data," *The Lancet*.

Kjellberg, A., Landström, U., Tesarz, M., Söderberg, L., and Åkerlund, E. 1996. "The Effects of Nonphysical Noise Characteristics, Ongoing Task and Noise Sensitivity on Annoyance and Distraction Due to Noise at Work," *Journal of Environmental Psychology* (16:2), pp. 123–136.

Kleiger, R. E., Stein, P. K., and Bigger, J. T. 2005. "Heart Rate Variability: Measurement and Clinical Utility," *Annals of Noninvasive Electrocardiology : The Official Journal of the International Society for Holter and Noninvasive Electrocardiology, Inc* (10:1), pp. 88–101.

Klimek, P., Kautzky-Willer, A., Chmiel, A., Schiller-Frühwirth, I., and Thurner, S. 2015. "Quantification of Diabetes Comorbidity Risks across Life Using Nation-Wide Big Claims Data," *PLoS Computational Biology* (11:4), pp. 1–16.

Kline, R. B. 2011. "Principles and Practice of Structural Equation Modeling," *Guilford*

Publication (3rd editio., Vol. 156).

Kline, R. B. 2012. "Assumptions in Structural Equation Modeling," *Handbook of Structural Equation Modeling*.

Kohli, R., and Tan, S. S.-L. 2016. "Electronic Health Records: How Can IS Researchers Contribute to Transforming Healthcare?," *MIS Quarterly* (40:3), Society for Information Management and The Management Information Systems Research Center, pp. 553–573.

Kraus, U., Schneider, A., Breitner, S., Hampel, R., Rückerl, R., Pitz, M., Geruschkat, U., Belcredi, P., Radon, K., and Peters, A. 2013. "Individual Daytime Noise Exposure during Routine Activities and Heart Rate Variability in Adults: A Repeated Measures Study," *Environmental Health Perspectives* (121:5), pp. 607–612.

Kuhn, M. 2015. "A Short Introduction to the Caret Package," *R Foundation for Statistical Computing*, pp. 1–10.

Lee, G.-S., Chen, M.-L., and Wang, G.-Y. 2010. "Evoked Response of Heart Rate Variability Using Short-Duration White Noise," *Autonomic Neuroscience : Basic & Clinical*, pp. 94–97.

Lee, H., Razjouyan, J., Nguyen, H., Lindburg, C., Srinivasan, K., Gilligan, B., Canada, K., Sharafkhaneh, A., Mehl, M., Currim, F., Ram, S., Lunden, M., Heerwagen, J., Kampschroer, K., Sternberg, E., and Najafi, B. 2018. "Sensor-Based Sleep Quality Index (SB-SQI): A New Metric to Examine the Association of Office Workstation Type on Stress and Sleep," *Sensors* (pre-print), Preprints.

Lin, Y.-K., Chen, H., Brown, R. A., Li, S.-H., and Yang, H.-J. 2017. "Healthcare Predictive Analytics for Risk Profiling in Chronic Care: A Bayesian Multitask Learning Approach," *MIS Quarterly* (41:2), pp. 473–495.

Lindberg, C. M., Srinivasan, K., Gilligan, B., Razjouyan, J., Lee, H., Najafi, B., Canada, K. J.,

Mehl, M. R., Currim, F., Ram, S., Lunden, M. M., Heerwagen, J. H., Kampschroer, K., and Sternberg, E. M. 2018. "Effects of Office Workstation Type on Physical Activity and Stress," *Occupational and Environmental Medicine* (10:75), pp. 689–695.

Little, R. J. A., and Rubin, D. B. 2002. *Statistical Analysis with Missing Data*, Hoboken, NJ, USA: John Wiley & Sons, Inc.

Liu, B. Q., and Goodhue, D. L. 2012. "Two Worlds of Trust for Potential E-Commerce Users: Humans as Cognitive Misers," *Information Systems Research* (23:4), pp. 1246–1262.

Liu, J., Ma, J., Wang, J., Zeng, D. D., Song, H., Wang, L., and Cao, Z. 2016. "Comorbidity Analysis According to Sex and Age in Hypertension Patients in China," *International Journal of Medical Sciences* (13:2), pp. 99–107.

Loh, W.-Y., and Shih, Y.-S. 1997. "Split Selection Methods for Classification Trees," *Statistica Sinica* (7:4), pp. 815–840.

Lusk, S. L., Hagerty, B. M., Gillespie, B., and Caruso, C. C. 2002. "Chronic Effects of Workplace Noise on Blood Pressure and Heart Rate," *Archives of Environmental Health* (57:4), pp. 273–281.

MacNaughton, P., Spengler, J., Vallarino, J., Santanam, S., Satish, U., and Allen, J. 2016. "Environmental Perceptions and Health before and after Relocation to a Green Building," *Building and Environment* (104), pp. 138–144.

Mai, Y., and Zhang, Z. 2018. "Software Packages for Bayesian Multilevel Modeling," *Structural Equation Modeling: A Multidisciplinary Journal* (25:4), Routledge, pp. 650–658.

Mak, C., and Lui, Y. 2012. "The Effect of Sound on Office Productivity," *Building Services Engineering Research and Technology* (33:3), SAGE PublicationsSage UK: London, England, pp. 339–345.

Malhi, K., Mukhopadhyay, S. C., Schnepper, J., Haefke, M., and Ewald, H. 2012. "A Zigbee-

Based Wearable Physiological Parameters Monitoring System," *IEEE Sensors Journal*.

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., and Schwartz, P. J. 1996. "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use," *European Heart Journal*.

Melville, N., and McQuaid, M. 2012. "Generating Shareable Statistical Databases for Business Value: Multiple Imputation with Multimodal Perturbation," *Information Systems Research* (23:2), pp. 559–574.

Merkle, E. C., and Wang, T. 2016. "Bayesian Latent Variable Models for the Analysis of Experimental Psychology Data," *Psychonomic Bulletin & Review*.

Moturu, S. T., Johnson, W. G., and Liu, H. 2007. "Predicting Future High-Cost Patients: A Real-World Risk Modeling Application," in *IEEE International Conference on Bioinformatics and Biomedicine*, , November, pp. 202–208.

Moturu, S. T., Liu, H., and Johnson, W. G. 2008. "Understanding the Effects of Sampling on Healthcare Risk Modeling for the Prediction of Future High-Cost Patients," *Biomedical Engineering Systems and Technologies* (25), pp. 493–506.

Muggeo, V. M., Atkins, D. C., Gallop, R. J., and Dimidjian, S. 2014. "Segmented Mixed Models with Random Changepoints: A Maximum Likelihood Approach with Application to Treatment for Depression Study," *Statistical Modelling* (14), pp. 293–313.

Muthén, B. O. 2002. "Beyond SEM : General Latent Variable Modelling," *Behaviormetrika*.

Nakagawa, S., and Schielzeth, H. 2013. "A General and Simple Method for Obtaining R2 from Generalized Linear Mixed-Effects Models," *Methods in Ecology and Evolution* (4:2), pp. 133–142.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. 1999. "Applied Linear Statistical Models," *Irwin Series in Statistics* (Vol. 1).

Newman, M. E. J., and Girvan, M. 2003. "Finding and Evaluating Community Structure in Networks," *Physical Review E* (69:2).

Pant, G., and Srinivasan, P. 2010. "Predicting Web Page Status," *Information Systems Research* (21:2), pp. 345–364.

Park, S. H., and Lee, P. J. 2017. "Effects of Floor Impact Noise on Psychophysiological Responses," *Building and Environment* (116), Pergamon, pp. 173–181.

Parssian, A., Sarkar, S., and Jacob, V. S. 2004. "Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product," *Management Science* (50:7), pp. 967–982.

Pentland, A., Lazer, D., Brewer, D., and Heibeck, T. 2009. "Using Reality Mining to Improve Public Health and Medicine," in *Studies in Health Technology and Informatics*.

Pereira, T., Almeida, P. R., Cunha, J. P. S., and Aguiar, A. 2017. "Heart Rate Variability Metrics for Fine-Grained Stress Level Assessment," *Computer Methods and Programs in Biomedicine*.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D. 2007. "Nlme: Linear and Nonlinear Mixed Effects Models," *R Package Version 3*.

Pituch, K. A., and Stevens, J. P. 2016. "Applied Multivariate Statistics for the Social Sciences," *Routledge*.

Rai, A., Burton-Jones, A., Chen, H., Gupta, A., Hevner, A. R., Ketter, W., Parsons, J., Rao, H. R., Sarkar, S., and Yoo, Y. 2017. "Editor's Comments: Diversity of Design Science Research," *Management Information Systems Quarterly* (41).

Ram, S., Wang, Y., Currim, F., and Currim, S. 2015. "Using Big Data for Predicting Freshmen Retention," in *Proceedings of the 35th International Conference on Information Systems (ICIS)*, pp. 1–16.

Rashid, M., and Zimring, C. 2008. "A Review of the Empirical Literature on the Relationships Between Indoor Environment and Stress in Health Care and Office Settings: Problems and Prospects of Sharing Evidence," *Environment and Behavior* (40:2), pp. 151–190.

Rashidi, P., and Mihailidis, A. 2013. "A Survey on Ambient-Assisted Living Tools for Older Adults," *IEEE Journal of Biomedical and Health Informatics*.

Raudenbush, S. W., and Bryk, A. S. 2002. "Hierarchical Linear Models: Applications and Data Analysis Methods," *Advanced Quantitative Techniques in the Social Sciences 1* (Vol. 2nd).

Ravi, D., Wong, C., Lo, B., and Yang, G.-Z. 2017. "A Deep Learning Approach to On-Node Sensor Data Analytics for Mobile or Wearable Devices," *IEEE Journal of Biomedical and Health Informatics*.

Razjouyan, J., Lee, H., Gilligan, B., Lindberg, C., Nguyen, H., Canada, K., Burton, A., Sharafkhaneh, A., Srinivasan, K., Currim, F., Ram, S., Mehl, M. R., Goebel, N., Lunden, M., Bhangar, S., Heerwagen, J., Kampschroer, K., Sternberg, E. M., and Najafi, B. 2019. *Wellbuilt for Wellbeing: Controlling Relative Humidity in the Workplace Matters for Our Health*, (Working paper).

Razjouyan, J., Naik, A. D., Horstman, M. J., Kunik, M. E., Amirmazaheri, M., Zhou, H., Sharafkhaneh, A., and Najafi, B. 2018. "Wearable Sensors and the Assessment of Frailty among Vulnerable Older Adults: An Observational Cohort Study," *Sensors*.

Real, R., and Vargas, J. M. 1996. "The Probabilistic Basis of Jaccard's Index of Similarity," *Systematic Biology*, pp. 380–385.

Rickert, J. 2011. "Big Data Analysis with Revolution R Enterprise," *Revolution White Paper*.

Ritz, C., Pilmann Laursen, R., and Trab Damsgaard, C. 2017. "Simultaneous Inference for Multilevel Linear Mixed Models—with an Application to a Large-Scale School Meal Study," *Journal of the Royal Statistical Society. Series C: Applied Statistics* (66:2), pp. 295–

311.

Roque, F. S., Jensen, P. B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., Jensen, L. J., and Brunak, S. 2011. "Using Electronic Patient Records to Discover Disease Correlations and Stratify Patient Cohorts," *PLoS Computational Biology* (7:8), p. e1002141.

Rosseel, Y. 2012. "Lavaan: An R Package for Structural Equation Modeling," *Journal of Statistical Software* (48:2), pp. 1–36.

Rubin, D. B. 1976. "Inference and Missing Data," *Biometrika* (63(3):3), pp. 581–592.

Ryff, C. D. 1989. "Happiness Is Everything, or Is It? Explorations on the Meaning of Psychological Well-Being.," *Journal of Personality and Social Psychology*.

Saar-Tsechansky, M., and Provost, F. 2007. "Handling Missing Values When Applying Classification Models," *Journal of Machine Learning Research* (8), pp. 1625–1657.

Sano, A., Taylor, S., Mchill, A. W., Andrew, ;, Phillips, J. K., Barger, L. K., Klerman, E., and Picard, R. 2018. "Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study," *Journal of Medical Internet Research* (20:6), pp. 210–216.

Schafer, J. L., and Graham, J. W. 2002. "Missing Data: Our View of the State of the Art," *Psychological Methods* (7:2), pp. 142–177.

Schuurmans, D., and Greiner, R. 1997. "Learning to Classify Incomplete Examples," *Computational Learning Theory and Natural Learning Systems, Making Learning Systems Practical* (4), pp. 87–105.

Seidman, M. D., and Standring, R. T. 2010. "Noise and Quality of Life," *International Journal of Environmental Research and Public Health* (7:7), pp. 3730–3738.

Shaffer, F., and Ginsberg, J. P. 2017. "An Overview of Heart Rate Variability Metrics and

Norms," *Frontiers in Public Health* (5).

Shi, C., Kong, X., Yu, P. S., Xie, S., and Wu, B. 2012. "Relevance Search in Heterogeneous Networks," in *Proceedings of the 15th International Conference on Extending Database Technology*, pp. 180–191.

Shuai, X., Zhou, Z., and Yost, R. S. 2003. "Using Segmented Regression Models to Fit Soil Nutrient and Soybean Grain Yield Changes Due to Liming," *Journal of Agricultural, Biological, and Environmental Statistics* (8:2), pp. 240–252.

Sim, C. S., Sung, J. H., Cheon, S. H., Lee, J. M., Lee, J. W., and Lee, J. 2015. "The Effects of Different Noise Types on Heart Rate Variability in Men," *Yonsei Medical Journal* (56:1), pp. 235–243.

Skrondal, A., and Rabe-Hesketh, S. 2004. "Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models," *CRC Press*.

Soares-Miranda, L., Sattelmair, J., Chaves, P., Duncan, G. E., Siscovick, D. S., Stein, P. K., and Mozaffarian, D. 2014. "Physical Activity and Heart Rate Variability in Older Adults: The Cardiovascular Health Study," *Circulation*.

Srinivasan, K., Currim, F., and Ram, S. 2018. "Predicting High Cost Patients at Point of Admission Using Network Science," *IEEE Journal of Biomedical and Health Informatics* (22:6), pp. 1970–1977.

Srinivasan, K., Currim, F., Ram, S., Lindberg, C., Sternberg, E., Skeath, P., Najafi, B., Razjouyan, J., Lee, H., Foe-parker, C., Goebel, N., Herzl, R., Mehl, M. R., Gilligan, B., Heerwagen, J., Kampschroer, K., and Canada, K. 2016. "Feature Importance and Prediction Modeling for Multi-Source Healthcare Data with Missing Values," in *Proceedings of the 6th International Conference on Digital Health 2016*, Montreal, Canada: ACM Press, pp. 47–54.

Srinivasan, K., Currim, F., Ram, S., Mehl, M. R., Lindberg, C., Sternberg, E., Skeath, P., Najafi, B., Razjouyan, J., Lee, H.-K., Lunden, M., Goebel, N., Andrews, S., Herzl, D., Herzl, R., Gilligan, B., Heerwagen, J., Kampschroer, K., and Canada, K. 2017. "A Regularization Approach for Identifying Cumulative Lagged Effects in Smart Health Applications," in *Proceedings of the 7th International Conference on Digital Health*, London, United Kingdom: ACM Press, pp. 99–103.

Srinivasan, K., Ram, S., Lindberg, C., Lee, H., Foe-parker, C., Mehl, M. R., Gilligan, B., Canada, K., Currim, F., Ram, S., Lindberg, C., Sternberg, E., Skeath, P., Najafi, B., Razjouyan, J., Lee, H., Foe-parker, C., Goebel, N., Herzl, R., Mehl, M. R., Gilligan, B., Heerwagen, J., Kampschroer, K., Canada, K., Goebel, N., Herzl, R., Mehl, M. R., Gilligan, B., Heerwagen, J., Kampschroer, K., Canada, K., Currim, F., Ram, S., Lindberg, C., Sternberg, E., Skeath, P., Najafi, B., Razjouyan, J., Lee, H., Foe-parker, C., Mehl, M. R., Gilligan, B., Canada, K., Goebel, N., Herzl, R., Mehl, M. R., Gilligan, B., Heerwagen, J., Kampschroer, K., Canada, K., Goebel, N., Herzl, R., Mehl, M. R., Gilligan, B., Heerwagen, J., Kampschroer, K., Canada, K., Currim, F., Ram, S., Lindberg, C., Sternberg, E., Skeath, P., Najafi, B., Razjouyan, J., Lee, H., Foe-parker, C., Mehl, M. R., Gilligan, B., and Canada, K. 2016. "Feature Importance and Prediction Modeling for Multi-Source Healthcare Data with Missing Values," in *Proceedings of the 6th International Conference on Digital Health 2016*, Montreal, Canada: ACM Press, pp. 47–54.

Staats, B. R., Dai, H., Hofmann, D., and Milkman, K. L. 2017. "Motivating Process Compliance Through Individual Electronic Monitoring: An Empirical Examination of Hand Hygiene in Healthcare," *Management Science* (63:5), pp. 1563–1585.

Steinhaeuser, K., and Chawla, N. V. 2009. "A Network-Based Approach to Understanding and Predicting Diseases," in *Social Computing and Behavioral Modeling*, pp. 1–8.

Stekhoven, D. J., and Buhlmann, P. 2012. "Missforest-Non-Parametric Missing Value Imputation for Mixed-Type Data," *Bioinformatics* (28:1), pp. 112–118.

Sternberg, E., Gilligan, B., and Lindberg, C. 2016. "Health and Wellbeing in GSA Office Buildings and Beyond," Philadelphia.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., and Carpenter, J. R. 2009. "Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls," *BMJ* (338), British Medical Journal Publishing Group.

Sun Sim, C., Hyun Sung, J., Hyeon Cheon, S., Myung Lee, J., Won Lee, J., and Lee, J. 2015. "The Effects of Different Noise Types on Heart Rate Variability in Men," *Yonsei Medical Journal* (56:1).

Sushmita, S., Newman, S., Marquardt, J., Ram, P., De Cock, M., Teredesai, A., Prasad, V., Cock, M. De, and Teredesai, A. 2015. "Population Cost Prediction on Public Healthcare Datasets," in *Proceedings of the 5th International Conference on Digital Health 2015*, ACM Press, pp. 87–94.

Thayer, J. F., Verkuil, B., Brosschot, J. F., Kampschroer, K., West, A., Sterling, C., Christie, I. C., Abernethy, D. R., Sollers, J. J., Cizza, G., Marques, A. H., and Sternberg, E. M. 2010. "Effects of the Physical Work Environment on Physiological Measures of Stress," *European Journal of Cardiovascular Prevention and Rehabilitation* (17:4), pp. 431–9.

The Commonwealth Fund. 2014. "US Health System Ranks Last Among Eleven Countries on Measures of Access, Equity, Quality, Efficiency, and Healthy Lives." (https://www.commonwealthfund.org/press-release/2014/us-health-system-ranks-last-among-eleven-countries-measures-access-equity?redirect_source=/publications/press-releases/2014/jun/us-health-system-ranks-last).

Tibshirani, R. 1996. “Regression Selection and Shrinkage via the Lasso,” *Journal of the Royal Statistical Society B*, pp. 267–288.

Verkuil, B., Brosschot, J. F., Tollenaar, M. S., Lane, R. D., and Thayer, J. F. 2016. “Prolonged Non-Metabolic Heart Rate Variability Reduction as a Physiological Marker of Psychological Stress in Daily Life,” *Annals of Behavioral Medicine*.

Walker, E. D., Brammer, A., Cherniack, M. G., Laden, F., and Cavallari, J. M. 2016. “Cardiovascular and Stress Responses to Short-Term Noise Exposures—A Panel Study in Healthy Males,” *Environmental Research* (150:October), pp. 391–397.

Wang, J. B., Cadmus-Bertram, L. A., Natarajan, L., White, M. M., Madanat, H., Nichols, J. F., Ayala, G. X., and Pierce, J. P. 2015. “Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial,” *Telemedicine and E-Health* (21:10), Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA, pp. 782–792.

Wang, Y., Currim, F., Currim, S., and Ram, S. 2015. “Class Imbalance Learning and Novel Feature Extraction Methods for Predicting Freshman Retention,” in *Proceedings of the Fifteenth Annual Workshop on Information Technology*, Dallas, Texas, pp. 1–15.

West, B. T., and Galecki, A. T. 2011. “An Overview of Current Software Procedures for Fitting Linear Mixed Models,” *The American Statistician* (65:4), pp. 274–282.

Xhyheri, B., Manfrini, O., Mazzolini, M., Pizzi, C., and Bugiardini, R. 2012. “Heart Rate Variability Today,” *Progress in Cardiovascular Diseases* (55:3), pp. 321–331.

Xiang, S., Yuan, L., Fan, W., Wang, Y., Thompson, P. M., and Ye, J. 2013. “Multi-Source Learning with Block-Wise Missing Data for Alzheimer’s Disease Prediction,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 185.

Xue, L., Ray, G., and Gu, B. 2011. "Environmental Uncertainty and IT Infrastructure Governance: A Curvilinear Relationship," *Information Systems Research* (22:2), pp. 389–399.

Yamada, I., and Lopez, G. 2012. "Wearable Sensing Systems for Healthcare Monitoring," in *2012 Symposium on VLSI Technology (VLSIT)*, IEEE, June, pp. 5–10.

Yin, D., Bond, S. D., and Zhang, H. 2014. "Anxious or Angry? Effects of Discrete Emotions on the Perceived Helpfulness of Online Reviews," *MIS Quarterly* (38:2), pp. 539–560.

Zheng, Y. 2015. "Methodologies for Cross-Domain Data Fusion: An Overview," *IEEE Transactions on Big Data* (1:1), pp. 16–34.

Zhu, H., Chen, H., and Brown, R. 2018. "A Sequence-to-Sequence Model-Based Deep Learning Approach for Recognizing Activity of Daily Living for Senior Care," *Journal of Biomedical Informatics* (84), pp. 148–158.

Zook, C. J., and Moore, F. D. 1980. "High-Cost Users of Medical Care," *New England Journal of Medicine* (302:18), pp. 996–1002.

Zou, H. 2006. "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association* (101:476), Taylor & Francis, pp. 1418–1429.

Zou, H., and Hastie, T. 2005. "Regularization and Variable Selection via the Elastic-Net," *Journal of the Royal Statistical Society* (67:2), pp. 301–320.