1-1-2014

An Analytics Approach To Designing Patient Centered Medical Home

Saeede Ajorlou
Wayne State University,

AN ANALYTICS APPROACH TO DESIGNING PATIENT CENTERED MEDICAL HOME

by

SAEEDE AJORLOU

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for degree of

DOCTOR OF PHILOSOPHY

2014

MAJOR: INDUSTRIAL AND SYSTEMS ENGINEERING

Approved by:

_____

Advisor Date

_____

_____

_____

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Prof. Kai Yang for the continuous support of my Ph.D study and research. Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Boris Mordukhovich, Prof. Leslie Monplaisir , and Prof. Qingyu Yang, for their encouragement, insightful comments, and hard questions. My sincere thanks also goes to Prof. Romesh Saigal from University of Michigan for his excellent guidance and support throughout my thesis.

This study was supported by VA Center for Applied System Engineering under grant NSF CMMI 1233504 “An Allocation Model with Dynamic Updates for Balanced Workload Distribution on Patient Centered Medical Homes”.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

PCMH: Patient Centered Medical Home

VHA: Veteran Health Administration

CMS: Centers for Medicare & Medicaid Services

PHA: Progressive Hedging Algorithm

BDA: Benders Decomposition Algorithm

VSS: Value of Stochastic Solution

NCVAS: National Center for Veterans Analysis and Statistics

VAMC: VA medical centers

CBOC: Community Based Outpatient Clinic

VC: Vet Center

PCP: Primary Care Physician

DSS: Decision Support System

NPCD: National Patient Care Database

CDW: Corporate Data Warehouse

DRG: Diagnosis Related Group

ACC: Aggregated Condition Category

ICD-9-CM: International Classification of Disease, ninth version, Clinical Modification

CAN: Care Assessment Need

LOS: Length of Stay

Chapter I Introduction

Health care delivery is a complex multilevel system in which primary care is the base level and acts as a principal point of consultation for patients. The traditional format of primary care is mainly featured by primary care physicians (PCP), in which each PCP has a designated set of patients, called a patient panel. In current practices of most providers, the panel is simply decided by a predetermined maximum size; that is when the quota is reached, no more patients will be added [1,2]. Typical panel sizes range from 1200 to 1600 patients. However, this number alone cannot reflect the actual health workload generated in the panel. For example, a PCP with 1200 young and healthy patients might be generally underutilized, while one with 1200 elderly patients having multiple comorbidities may experience excessive workload, causing long delays in its panel appointment times and forcing patients to switch their PCPs.

It is found that many factors such as patient's age, gender, health status and insurance plan can influence the required healthcare workload. Ostbye and colleagues [3] find that patients with different chronic diseases regularly have different visiting frequencies to their PCPs. Naessens and colleagues [4] discover that the number of chronic conditions in a patient will significantly affect clinical workload and medical cost. Potts and colleagues [5] propose a risk-standard method to adjust the panel size for each PCP calculating disease burden of each physician panel for six chronic diseases. However, there is no description or proof about how the risk values are assigned. Balasubramanian and colleagues [6] apply classification and regression trees (CART) to classify approximately

20,000 patients at the Mayo Clinic of Rochester, Minnesota, into 28 categories by using age and gender as factors, so that each category has different workload patterns.

In recent years, the patient-centered medical home (PCMH) has been introduced as a prominent intervention to improving the US primary care systems with better-quality outcomes at lower costs [7]. This model consists of different health professionals grouped together to provide comprehensive, coordinated, accessible and cost effective care while maintaining high levels of service quality and stability. Each team consists of a group of medical professionals such as primary care provider, registered nurse, nutritionist, social worker, and medical clerk that are well poised to provide many aspects of primary care. Theoretically, medical homes are composed of "joint principles" that ideally complement one another and feed into a comprehensive vision of appropriate primary care delivery. The principles are consisted of having a personal physician with an ongoing relationship, a whole person orientation care for all stages of life, a physician-directed medical practice taking responsibilities for all of the continuing care, a coordinated and/or integrated care system across all elements of the care systems, a continuous emphasis on quality and safety, an enhanced access to care through such systems as open scheduling and expanded hours, and finally an appropriate payment system that recognize the added value provided to PCMH patients [8].

Augmented with modern health information technology, the PCMH is crafted to initiate numerous reforms in health care delivery and reimbursement systems [9].

As of 2007, there was some literature examining the prevalence and effectiveness of medical homes. For instance, Fisher [10] outlined some recommendations for the success

of medical homes such as increasing effective communication and sharing of information across health care providers, broadening the medical performance measures to include patients' experience with care and ordinary assessment of outcomes, and establishment of medical-home payment system that share savings among all providers involved. A survey by Commonwealth Fund of 3,535 US adults found that when they were provided with a medical home, racial and ethnic disparities in care access and quality were substantially reduced [11]. Furthermore, having a medical home was associated with more preventive screenings and better management of chronic conditions. The Centers for Medicare & Medicaid Services (CMS) planned to pursue Medicare pilot projects in 400 practices in 8 regional sites, and by 2009, twenty bills promoting the PCMH concept have been successfully introduced in 10 states [12]. Another study within the Group Health system in Seattle showed that a medical home prototype led to 29% fewer emergency visits, 6% fewer hospitalizations, and total savings of $10.30 per patient per month over a twenty-one month period [13]. Bates and Bitton [14] indicated seven health information technology domains deemed to be critical for the success of the PCMH model including telehealth, measurement of quality and efficiency, care transitions, personal health records, and, most importantly, registries, team care, and clinical decision support for chronic diseases.

Practically, as of December 2009, there were about 26 pilot projects involving medical home being directed in 18 states. These consist of over 14,000 physicians and approximately 5 million patients [15]. Of interest, Veterans Health Administration (VHA) launched a nationwide 3-year program in April 2010 to create PCMHs in more

than 900 primary care clinics. Early results indicated dramatic improvements such as reducing the appointment waiting time from as long as 90 days down to one day and decreasing the percentage of inappropriate emergency department visits from 52% to 12% [16].

1.1 Overview of Patient Centered Medical Home

Presently, the PCMH model has been practiced by many hospitals and medical centers, Bitton et. al. [15], and its performance has been evaluated by many studies, Nutting et. al. [17], Jaen et. al. [18], and Crabtree et. al. [19].As it is shown in Fig.1, there are two different phases in PCMH process. In this section, we describe the two phases of PCMH.

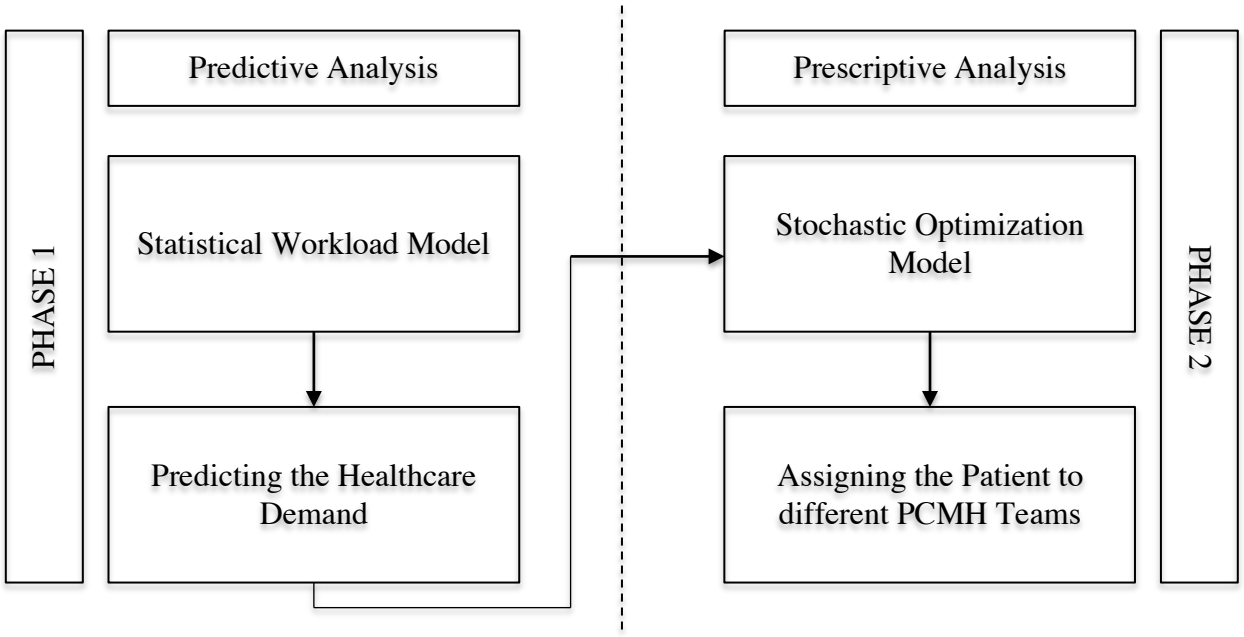


Figure 1 Two Phases in PCMH

Fig.2 shows the different benefits from Patient Centered Medical Home.

Figure 2 Patient Centered Medical Home Benefits

The first phase is, *healthcare demand estimation*. Generally, healthcare demand is the amount of time required to care for a patient over a time period which is related to patient's demographic, diagnostic, and health attributes. The goal of the first phase in PCMH process, is to develop a rigorous statistical based workload estimation model which provides a good estimate of workload healthcare demand for a relevant set of healthcare professionals for any particular patient based on his/her key attributes.

The second phase is called, *patient assignment*. This phase can be executed with the help of healthcare demand estimation from the first phase and developing an optimization

model in which each patient is assigned to only one team with the respect to balancing supply and demand policy in healthcare system. It should be noted that the healthcare supply is the total available hours of profession time within a given period (typically a year) and it can be calculated easily based on the head counts and available service hours to patients from all professional lines.

A good patient panel design and management methodology is even more critical for PCMH model than the traditional PCP model for the following reasons:

- In the traditional single PCP model, the healthcare supply is the total available hours of physician time within a given period (typically a year) by a PCP, and the healthcare demand is the total requested physician hours generated by the patients in the panel. The healthcare supply can be treated as deterministic, and the healthcare demand as a random variable. In PCMH model, the healthcare supply is a portfolio of total available hours by various members in a team within a particular period, (e.g., total physician time, total nurse time, total clerk time, etc.), the healthcare demand is a portfolio of demand requested by the patients in the patient panel to PCMH team members, the healthcare supply is in the form of a deterministic vector, while the healthcare demand is in the form of a vector of random variables.
- Even for a single PCP model, balancing healthcare supply and demand by panel design in order to optimize patient access and the continuity of care is a challenging task. For PCMH model, it is a portfolio of healthcare supply and demand that needs to be balanced in order to improve access, maintain continuity

and reduce the care cost. It is a much more challenging task and without a scientific patient panel design methodology, depending on the composition of patients in a PCMH patient panel, it is possible that some members of the PCMH team will get overly stressed, while other team members are under-utilized.

- In PCMH model, the design of the professional mix in a team (that is, who is in the team) and team members staffing level (available hours of each team member) so that they match the healthcare demand portfolio generated from the patients in the panel is the key to its success. The desirable state of a PCMH team should have the following features: a) The workload generated by the patient panel should be spread evenly on PCMH team members; b) The amount of workload for each PCMH team member should be such that the work can be accomplished in a timely manner and each team member's utilization rate is high.
- In a medical facility that practices PCMH model, all primary care is performed by numbers of PCMH teams. Designing patient panels and allocating patient population to these multiple teams is a challenge, since the professional mix and staffing level of these teams must balance well with the total workload generated by the entire patient population of the medical facility.
- In any medical facility, due to migration or death, some existing patients drop out from the patient set and some new patients add to the patient population. This necessitates that the patient panels be dynamically updated; and so too the PCMH team staffing levels (which is also susceptible to the similar migration forces).
- The PCMH model can also be used in specialty care. However, specialty cares are

usually even more expensive, and the variation in workload generated by patients with various attributes is even higher. Thus, it is critical that this mismatch of supply and demand be minimized.

In a nutshell, one of the key success factors for the patient centered medical home model is to achieve balance between supply and demand of healthcare services. The annual supply of healthcare services can be estimated relatively easily based on head counts and available service hours to patients from all professional lines. The estimation of demand of healthcare services is much more difficult and it can be estimated based on the statistical workload model in the first phase [20, 21] and finally the patients can be assigned to multiple teams by an optimization model. In fact, the PCMH is in practice even more difficult, since the optimization model has the stochastic nature and the stochasticity is due to healthcare demand source [22].

1.2. Research Objectives

1.2.1. Develop a statistical workload estimation model based on patient attributes

In this research, we develop a multivariate hierarchical based portfolio prediction model that takes into account postulated attributes from different levels such as disease types (patient-level), years of experience of the assigned provider (team-level), and zip-code based distance between the patient's home and his/her assigned facility (facility-level). We also want to propose an intensity score for panel size and staffing level adjustment used at different levels of hierarchy, as it would help decision makers on their PCMH team allocation and budget policy decisions. Finally, we seek to screen highly contributing risk factors to demand portfolio variations, since it would inform program

analysts on areas more likely affecting the care portfolio balance.

To the best of our knowledge, our work is the first attempt to develop such a clinical portfolio prediction model for medical homes within the OR/MS community. Our contributions include extending the hierarchical generalized linear model to include multivariate response variables in a Bayesian framework, presenting a Markov chain Monte Carlo algorithm with novel prior specifications to fit the model, and utilizing our proposal on real data from VHA to produce findings that have key public and medical implications. Also our approach allows for passing heterogeneous variances and unstructured covariance matrices for the nested random effects as well as their interactions with responses and covariates simultaneously.

1.2.2. Develop optimization-based models as the basis for patient allocation care

With the help of the model developed in 1.2.1, we are able to estimate the annual workload demand portfolio for each patient with given attributes. In this research together with the healthcare service supply data, and based on the principles of balancing supply and demand, we proposed a stochastic optimization model with recourse to assign the patients to different PCMH teams in second phase. Moreover, we used Progressive Hedging Algorithm (PHA) and L Shaped Benders Decomposition Algorithm (BDA) to solve the assignment problem in second phase, and finally we compare these two stochastic solutions with the deterministic solution and we reported the value of stochastic solution (VSS). Finally we compared PHA and BDA in our problem considering CPU time and their performance for different dimension of

problem.

1.3. Dissertation Organization

The dissertation is organized as follows. In chapter 2, we do literature review on Multivariate Multilevel Framework and Bayesian Variable Selection and then we discuss about STAR models and Hierarchical STAR models. Finally, in the later part of the Chapter 2 we propose our method in Multi-response Hierarchical STAR model and Bayesian Function Selection.

In Chapter 3, we first describe the assumptions and stochastic optimization models for patient assignment and then we show the extensive form of our proposed model. Next the solution approaches such as Progressive Hedging and L Shaped Benders Decomposition and their application are introduced. In chapter 4, first the data source and the attributes are described completely. With the help of proposed method in chapter 2, we did the model fitting and diagnostics on the mentioned dataset. Next, for patient assignment, we applied two solution approaches to the stochastic optimization model and then we compared the solutions with deterministic one. Finally the value of stochastic solution is reported. Conclusions and future studies are presented in Chapter 5 of the dissertation.

Chapter II Methodology for Phase I

Many kinds of health care data, including clinical data, billing/claims data, and patient specific data, involve hierarchical (nested) or clustered structure. For example, in a study of assessing differences in mortality rates across hospitals, data is randomly collected on samples of patients nested within each hospital. In this application, there are two levels of the hierarchy (level-1 for patients and level-2 for hospitals), and for each level, a set of specific covariates is existed (such as age, gender, and severity of illness at the first level; and hospital size and hospital teaching status at the second level) that might have a relationship with the outcome. To handle these hierarchically structured data, multilevel models (also known as hierarchical linear models, variance components models, random-effect models, or split-plot designs) have been proposed and applied in different fields including psychometrics, biostatistics and econometrics [23]. The basic idea is to link the covariates at higher levels to the predictor variables at lower levels by imposing another set of regressions in which the lower-level (regression) coefficients are explained by higher-level predictors.

The assumption of parametric form of covariates in the hierarchical linear model makes it rather restricted. For example, in longitudinal growth studies where repeated measures of the response variable (e.g., height) are clustered within individuals, the relation between age and height is often found to be exponential. To relax the linearity constraints, covariates with nonparametric structure (such as local regression or smoothing spline) or semi-parametric structure (such as partially linear model or

varying-coefficient models) can be incorporated in the multilevel framework at each level of the hierarchy [23]. One such extension is generalized additive mixed models, which enjoy the nonparametric properties of additive models and distributional flexibility of generalized linear mixed models. Another more recent class of this type is the hierarchical version of structured additive regression (STAR) models [24] that offers a broad and rich class of complex regression containing several important subclasses as special cases e.g., generalized additive mixed models, state-space models for longitudinal studies, geo-additive models [25], and varying-coefficient models [26].

As in many areas of statistical modeling and machine learning, the problem of variable selection (also known as feature selection, attribute selection, model selection, variable subset selection) has become an important issue in multilevel models. Variable selection often aims to choose a subset of relevant covariates from a possibly large set of candidates that might include many redundant or irrelevant features. Due to its practical importance, this problem has attracted many researchers from diverse fields, leading to a vast amount of literature on selecting predictors of regression models. Classical methods in this area basically relied on 1) p-value such as stepwise deletion or 2) information criteria like AIC, BIC, and more recently focused information criterion [27], among others. However such approaches usually suffer from lack of stability and perform poorly in selecting random effect components [28]. In addition, they involve a combinatorial optimization comparing 2^{p+q} different models (p and q are numbers of fixed and covariance parameters, respectively), which is NP-hard and might be infeasible to solve even when $p+q \ll n$ is fixed (n is sample size) [29]. To address such drawbacks,

regularization (or shrinkage) methods have been introduced that focus on selecting variables simultaneously with model estimation using some data oriented penalty functions. Popular examples may include the least absolute shrinkage and selection operator (Lasso) [30] or smoothly clipped absolute deviation (SCAD) [31] and modifications such as hierarchical or random Lasso. To get an overview of variable selection in linear models, see the review paper by [32]. Variable selection is also of great importance in high-dimensional data such as DNA microarray or functional MRI data (see [33] for a review). Likewise, various studies have been devoted to variable selection in nonparametric additive models and semi-parametric linear models (see, for example, [34] and [35]). Multivariate variable selection has also been studied in a number of researches such as [36] and [37].

Compared to classical methods that are primarily based on Bayes factors, approaches for Bayesian variable selection are mostly built around spike-and-slab priors. The basic idea is to introduce a binary latent variable I_j associated with each regression coefficient so that the variable is forced to be zero when I_j is in the spike part, or keep unchanged if I_j is in the slab part. The posterior distribution of I_j is then interpreted as marginal posterior probabilities for inclusion or exclusion of the respective covariate. See stochastic search variable selection (SSVS) of [38] and mixture of Zellner's g priors of [39] as popular examples, and a recent review paper of O'Hara and Sillanpää (2009) [40].

In multilevel models, however, the problem of selecting the random effects is more complicated since it involves boundary problems that can arise from either nonnegative constraints on fixed-effect parameters or positive semi-definite constraints on covariance

matrices. To date, approaches for variable selection in this class mainly pertain to linear (or generalized linear) mixed models such as generalized information criterion of Pu and Niu [29], and Bayesian methods of Spiegelhalter [41], among others (see [42] for a review).

In contrast to variable selection, component (or function) selection deals with selecting an appropriate subset of covariates and, at the same time, determining whether linear or more flexible functional forms of covariates have to be chosen. Research on this area has started by [43] who proposed a group SCAD penalty for regularization in wavelets approximation. [44] developed the COSSO estimator in additive smoothing spline analysis of variance (SS-ANOVA) models with a fixed number of covariates. Recently, by extending the nonnegative garrote estimator [28], [45] developed a single step shrinkage approach method for function selection in generalized additive models.

In this research, consistent with the idea of modeling multivariate outcomes in multilevel data structures [23] , we first extend hierarchical STAR models introduced in [24] to include multivariate response variables from the exponential family distribution. This way, we will be able to simultaneously model the relationship of several responses on a set of structured additive predictors accounting for possible correlation among the dependent variables. Then, we propose spike-and-slab priors for automatic variable selection and model choice within a Bayesian hierarchical framework similar to [46]. We apply our model to a real-world healthcare data obtained from the Department of Veteran Affairs (VA). The application analyzes Patient Centered Medical Home (PCMH) project data gathered from a large number of medical facilities during fiscal year 2011–12.

Separate data tables from 1) patient's health conditions and care utilization, and 2) patient's demographic information are first combined to form patient-level data. The patient-level data is further aggregated to the provider and station levels to help predict patient's total care demands on primary and non-primary care on a yearly basis. By combining these multilevel data sources together, our proposal can assist health professionals in primary care management and the assignment of predicted healthcare to providers.

2.1. STAR models based on Bayesian P-splines

Let (y_i, x_i, v_i), $i = 1, \dots, n$, denote the i-th sampled vector in data, where y_i is the response variable, $x_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right)'$ is a vector of continuous covariates, and $v_i = (v_{i1}, v_{i2}, \dots, v_{is})'$ is a vector of further (mostly categorical) predictors. Structured Additive Regression (STAR) models [46] assume that, given x_i and v_i, the distribution of y_i belongs to an exponential family $\pi(y_i|x_i, v_i, \phi) = c(y_i, \phi)\exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi}\right)$, where $b(\cdot)$, $c(\cdot)$, θ_i, and ϕ are determined by the type of distribution. The conditional expected value $\mu_i = E(y_i|x_i, v_i)$ is related to a semi-parametric additive predictor η_i by $\mu_i = g(\eta_i)$ via a fixed (known) link function $g(\cdot)$ as in generalized linear models. The additive predictor η_i has the form

$$\eta_i = f_1(x_{i1}) + \cdots + f_p\left(x_{ip}\right) + v_i'\gamma, \tag{2.1}$$

in which $f_1, \dots, f_p$ are unknown nonlinear (possibly smooth) functions of the continuous covariates, and $v_i'\gamma$ represents the usual linear part of the model. Following the Bayesian

version of P(enalized)-splines [47], the unknown functions f_j is approximated by a polynomial spline of degree r defined over a set of (not necessarily equally spaced) knots $x_j^{min} = \zeta_j^0 < \zeta_j^1 < \cdots < \zeta_j^{k_j-1} < \zeta_j^{k_j} = x_j^{max}$ within the domain of x_j. The spline can be expressed in terms of a linear combination of $M_j = k_j + r$ B-spline basis functions evaluated at the observation x_j, i.e.,

$$f_j(x_j) = \sum_{m=1}^{M_j} \beta_{jm} B_{jm}(x_j). \tag{2.2}$$

Here $B(\cdot)$'s are known basis functions and $\beta_j = (\beta_{j1}, \dots, \beta_{jM_j})'$ corresponds to a vector of unknown regression coefficients to be estimated. By defining the $(n \times M_j)$ design matrix $X_j(i,m) = B_{jm}(x_{ij})$, the predictor (2.2) can be rewritten in matrix form as

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{V}\boldsymbol{\gamma}, \tag{2.3}$$

where $\boldsymbol{V}$ is the usual design matrix for linear effects. Within this unified framework, components of (2.2) can represent various types of model terms, such as 1) linear terms ($f_j(\boldsymbol{x}) = \beta_j \boldsymbol{x}_j$); 2) nominal or ordinal predictors ($f(x_{ji}) = \beta_{x(k)}$ iff $x_{ji} = k$); 3) smooth functions of continuous covariates (splines, kriging effects, tensor product splines, etc.); 4) Markov random field or its conditional specification, e.g. the conditional autoregressive model; 5) random effect models (cluster-specific intercept or slopes); and 6) interaction terms between different effects (varying-coefficient models, effect modifiers).

For a fully Bayesian inference when selection of variables (and functions) is not considered, a diffuse prior $p(\gamma) \propto const$ is typically used for sampling from $\boldsymbol{\gamma}$. The

choice of priors for the unknown functions $f_1, \ldots, f_p$, however, depends on the type of the covariate and the prior beliefs about smoothness. To avoid over fitting of a particular function f_j, the smoothness priors can be written into a general form as

$$p(\beta_j|\tau_j^2) \propto \left(\frac{1}{\tau_j^2}\right)^{\mathrm{rk}(\boldsymbol{K}_j)/2} \exp\left(-\frac{1}{2\tau_j^2}\beta_j' K_j \beta_j\right), \tag{2.4}$$

in which $\boldsymbol{K}_j$ is a penalty matrix and τ_j^2 is the variance parameter. The goal of $\boldsymbol{K}_j$ is to shrink smoothness parameters towards zero, or penalize unexpected jumps between adjacent β_j's. In most cases such as Gaussian random field $\boldsymbol{K}_j$ is rank deficient (i.e., $\mathrm{rk}(\boldsymbol{K}_j) < M_j$), leading to partially improper prior for β_j. The variance parameter τ_j^2 , controls the amounts of smoothness and is sampled by an uninformative (conjugate) inverse Gamma hyper priors $\tau_j^2 \sim IG(a_j, b_j)$ normally with small choices for a_j and b_j.

2.2. Hierarchical STAR models

When data are hierarchically structured in some levels, STAR models can be extended in a multilevel framework to account for possible correlations within units of a cluster (or a level) in the hierarchy. Such specification is usually expressed by imposing another regression model with structured additive predictors to the coefficient $\boldsymbol{\beta_j}$ in (2.3) as

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \boldsymbol{X}_{j1}\boldsymbol{\beta}_{j1} + \cdots + \boldsymbol{X}_{jp_j}\boldsymbol{\beta}_{jp_j} + \boldsymbol{V}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j. \tag{2.5}$$

Here it is assumed that $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \tau_j^2 I)$ is a vector of i.i.d. Gaussian random variables, but more complicated forms such as Dirichlet process mixture can be applied. Modeling higher levels of the hierarchy are also straightforward by again setting another STAR equations to the parameters in (2.5) e.g., $\boldsymbol{\beta}_{jl} = \boldsymbol{\eta}_{jl} + \boldsymbol{\varepsilon}_{jl}, l = 1, \ldots, p_j, \boldsymbol{\varepsilon}_{jl} \sim N(\mathbf{0}, \tau_{jl}^2 I)$ for

level-3 regression. In this way, the whole model can be seen as a hierarchy of complicated STAR models with (possibly) nonlinear and smooth terms. In some applications of hierarchical models, observations are clustered according to their spatial (or geographical) positions. For example, in our VA medical home study, x_j may represent the district (or zip code) in which the patient lives. This way, x_j represents a group indicator taking values of $c \in \{1, \dots, C_j\}$. Then a regular way to model such cluster specific heterogeneity is to assume $f_j(c) = \beta_{jc} \sim N(0, \tau_j^2)$ with design matrix X_j being a 0/1 incidence matrix of dimension $n \times C_j$. Note that this approach is also taken when modeling random intercepts in multilevel structure. In other applications, we may like to study how the effect of a covariate is modified according to changes in the levels of a third variable. Such interactions can happen among the covariates at one given level or across multiple levels. As an instance in our case study, we are interested in how possessing a particular comorbid condition can moderate the relationship between patient's age and healthcare demand. Here, it is presumed that x_j is a two-dimensional term as $x_j = \left(x_j^{(1)}, x_j^{(2)}\right)'$. If $x_j^{(1)}$ is continuous and $x_j^{(2)}$ is categorical, their interaction is modeled by $f_j(x_j) = h\left(x_j^{(1)}\right)x_j^{(2)}$, and the associated design matrix is given by $diag\left(x_1^{(2)}, \dots, x_n^{(2)}\right)X_j^{(1)}$, in which $X_j^{(1)}$ is the usual design matrix for spline basis function evaluated at the observation $x_j^{(1)}$. If both covariates are continuous, a more flexible approach can be based on two-dimensional P-spline, in which the unknown

interaction surface can be approximated by the tensor product of the corresponding one-dimensional B-splines as:

$\left(x_j^{(1)}, x_j^{(2)}\right) = \sum_{m_1=1}^{M_{1j}} \sum_{m_2=1}^{M_{2j}} \beta_{j,m_1m_2} B_{j,m_1}\left(x_j^{(1)}\right) B_{j,m_2}\left(x_j^{(2)}\right)$.

The related design matrix X_j is then $n \times \left(M_{1j} \cdot M_{2j}\right)$ and it consists of products of basis functions. The appropriate priors for $\beta_j = \left(\beta_{j,11}, \ldots, \beta_{j,M_{1j}M_{2j}}\right)'$ are commonly found in spatial statistics.

Another common application in multilevel analysis is related to random slopes that appear when combining regression equations of higher levels with the lower levels to form a compound representation [23]. For example, in our case study of the VA medical home project, we would like to model the heterogeneity in the slope of relationship between healthcare demand and patient's age among all PCMH teams. Then, a random slope with regard to index variable $x_j^{(1)}$, which indicates the teams here, can be incorporated as $f_j\left(x_j\right) = h\left(x_j^{(1)}\right) x_j^{(2)}$ with $h\left(x_j^{(1)}\right) = \beta_{jc} \sim N\left(0, \tau_j^2\right)$. Following this, the design matrix X_j is given by $diag\left(x_1^{(2)}, \ldots, x_n^{(2)}\right) X_j^{(1)}$ where $X_j^{(1)}$ is a 0/1 incidence matrix.

2. 3. Proposed methods

2. 3. 1. Multi-response hierarchical STAR model

When we want to simultaneously study multiple response variables, a multivariate model should be developed to capture additional correlation among different measurements. One key advantage of such modeling lies in its ability to control type I error rate better as

compared to carrying out a series of univariate tests. In the context of multilevel analysis, different responses can be incorporated by placing them in a separate 'response' level at the lowest level of the hierarchy. A series of d dummy variables, one for each response, is then defined and entered into regression equations at higher levels. For simplicity, we first focus on three-level structure, $responses^{(h)}$ within $patient_i$ within (medical home) $team_j$, with regular predictors, and then show how this can be extended to STAR context. A model with more than three levels is just a straightforward extension to what we propose here.

Suppose there are H response variables in the lowest level. We define $d_{hij}^{(h')} = 1$ if response h'-th is modeled and zero otherwise [23]. Let $x_{p,ij}$ and $z_{q,j}$ denote p-th and q-th covariate in the patient level and team level, respectively. Let $u_{0,j}^{(h')}$ and $u_{p,j}^{(h')}$ represent h'-th random intercept and h'-th random slope of the p-th predictor in the patient level, one-to-one. Then we model the outcome as

$$y_{hij} = \sum_{h'} d_{hij}^{(h')} \beta_0^{(h')} + \sum_{h'} d_{hij}^{(h')} \sum_{p=1}^{P} \overline{\beta}_p^{(h')} x_{p,ij} + \sum_{h'} d_{hij}^{(h')} \sum_{q=1}^{Q} \underline{\beta}_q^{(h')} z_{q,j} + \sum_{h'} d_{hij}^{(h')} \sum_{p=1}^{P} \sum_{q=1}^{Q} \beta_{p,q}^{(h')} x_{p,ij} z_{q,j} + \sum_{h'} d_{hij}^{(h')} \sum_{p=1}^{P} u_{p,j}^{(h')} x_{p,ij} + \sum_{h'} d_{hij}^{(h')} u_{0,j}^{(h')} + \sum_{h'} d_{hij}^{(h')} \varepsilon_{ij}^{(h')} \quad (2.6)$$

$$\begin{bmatrix} u_{0,j}^{(1)} \\ \vdots \\ u_{P,j}^{(1)} \\ \vdots \\ \vdots \\ \vdots \\ u_{P,j}^{(H)} \end{bmatrix} \sim N(\mathbf{0}, \Omega_u),\ \Omega_u = \begin{bmatrix} \tau_{u_0}^{2(1)} & \cdots & & \cdots & \cdots & \cdots & \tau_{u_{0,P}}^{(1)(H)} \\ \vdots & \ddots & & & & & \vdots \\ \tau_{u_{0,P}}^{(1)} & \cdots & \tau_{u_P}^{2(1)} & & & & \tau_{u_{P,P}}^{(1)(H)} \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & & \ddots & \vdots \\ \tau_{u_{0,P}}^{(1)(H)} & \cdots & & \cdots & \cdots & \cdots & \tau_{u_P}^{2(H)} \end{bmatrix} \tag{2.7}$$

$$\begin{bmatrix} \varepsilon_{ij}^{(1)} \\ \vdots \\ \varepsilon_{ij}^{(H)} \end{bmatrix} \sim N(\mathbf{0}, \Omega_\varepsilon),\ \Omega_\varepsilon = \begin{bmatrix} \sigma_\varepsilon^{2(1)} & \cdots & \sigma_\varepsilon^{(1)(H)} \\ \vdots & \ddots & \vdots \\ \sigma_\varepsilon^{(1)(H)} & \cdots & \sigma_\varepsilon^{2(H)} \end{bmatrix} \tag{2.8}$$

The first term in (2.6) shows the grand mean for each of the response variable followed by patient level predictors and team level predictors; then cross-level interactions (effect modifiers) are included followed by random slopes and then random intercept terms; and at last patient level residuals. Note that there is no level-1 residual specified since level-1 exists only to define the multivariate structure. The random effects are defined in (2.7) with a general unstructured covariance Ω_u that contains the pairwise covariances between each set of these random effects for the intercept and slopes within each of the responses and between the response variables. The patient level residuals are defined in (2.8) with covariance structure Ω_ε that would include all variances and covariances between patient level residuals. Taking a matrix form, we can rewrite (2.6) as

$$y_{hij} = \sum_{h'} d_{hij}^{(h')} \boldsymbol{Z}_j^T \mathfrak{B}^{(h')} \boldsymbol{X}_{ij} + \sum_{h'} d_{hij}^{(h')} \boldsymbol{U}_j^{(h')^T} \boldsymbol{X}_{ij} + \sum_{h'} d_{hij}^{(h')} \varepsilon_{ij}^{(h')}, \quad (2.9)$$

where we have

$$\boldsymbol{Z}_j = \left[1, z_{1,j}, \dots, z_{Q,j}\right]^T, \boldsymbol{X}_{ij} = \left[1, x_{1,ij}, \dots, x_{P,ij}\right]^T, \boldsymbol{U}_j^{(h')} = \left[u_{0,j}^{(h')}, u_{1,j}^{(h')}, \dots, u_{P,j}^{(h')}\right]^T \quad (2.10)$$

$$\mathfrak{B}^{(h')} = \begin{bmatrix} \beta_0^{(h')} & \overline{\beta}_1^{(h')} & \dots & \overline{\beta}_P^{(h')} \\ \underline{\beta}_1^{(h')} & \beta_{1,1}^{(h')} & \dots & \beta_{P,1}^{(h')} \\ \vdots & \vdots & & \vdots \\ \underline{\beta}_Q^{(h')} & \beta_{1,Q}^{(h')} & \dots & \beta_{P,Q}^{(h')} \end{bmatrix}. \quad (2.11)$$

Note that $\overline{\beta}_1^{(h')}, \dots, \overline{\beta}_P^{(h')}$ in the first row of (11) show regression coefficients for patient level predictors while $\underline{\beta}_1^{(h')}, \dots, \underline{\beta}_Q^{(h')}$ placed in the first column of (2.11) indicate coefficients for team level variables.

To extend this within the STAR framework where the covariates are represented by a linear combination of B-splines basis functions, we simplify (2.6) for a particular outcome h' as

$$y_{hij}^{(h')} = \beta_0^{(h')} + \sum_{p=1}^{P} \left(\overline{\beta}_p^{(h')} + u_{p,j}^{(h')}\right) x_{p,ij} + \sum_{q=1}^{Q} \underline{\beta}_q^{(h')} z_{q,j} + \sum_{p=1}^{P} \sum_{q=1}^{Q} \beta_{p,q}^{(h')} x_{p,ij} z_{q,j} + u_{0,j}^{(h')} + \varepsilon_{ij}^{(h')} \quad \text{for } h' = 1, \dots, H. \quad (2.12)$$

We assume that, for response h', patient level covariate $x_p, p = 1, \dots, P$ is represented by a set of $\overline{M}_p{}^{(h')} = \overline{k}_p + \overline{r}$ polynomial spline of degree $\overline{r}$ over $\overline{k}_p + 1$ knots $\zeta_p^{0(h')} < \zeta_p^{1(h')} < \dots < \zeta_p^{k_j(h')}$. Similarly, team level predictor $z_q, q = 1, \dots, Q$ is represented by

$\underline{M}_q^{(h')} = \underline{k}_q + \underline{r}$ polynomial splines of degree $\underline{r}$ over a domain. Hence, a hierarchical STAR model with a multivariate response has the form

$$y_{hij}^{(h')} = \beta_0^{(h')} + \sum_{p=1}^{P} \sum_{m_1=1}^{\overline{M}_p^{(h')}} \left(\overline{\beta}_{m_1 p}^{(h')} + u_{m_1,p,j}^{(h')} \right) \overline{B}_{m_1 p}^{(h')}(x_{p,ij}) + \sum_{q=1}^{Q} \sum_{m_2=1}^{\underline{M}_q^{(h')}} \underline{\beta}_{m_2 q}^{(h')} \underline{B}_{m_2 q}^{(h')}(z_{q,j}) + \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{m_1=1}^{\overline{M}_p^{(h')}} \sum_{m_2=1}^{\underline{M}_q^{(h')}} \beta_{m_1 m_2 pq}^{(h')} B_{m_1 p}^{(h')}(x_{p,ij}) B_{m_2 q}^{(h')}(z_{q,j}) + u_{0,j}^{(h')} + \varepsilon_{ij}^{(h')}, \; h' = 1, \ldots, H \tag{2.13}$$

$$\begin{bmatrix} u_{0,j}^{(1)} \\ \vdots \\ \vdots \\ \vdots \\ u_{\overline{M}_P^{(H)},P,j}^{(H)} \end{bmatrix} \sim N(\mathbf{0}, \Omega_u), \; \Omega_u = \begin{bmatrix} \tau_{u_0}^{2(1)} & \cdots & \cdots & \cdots & \tau_{u_{0,\left(\overline{M}_P^{(H)},P\right)}}^{(1)(H)} \\ \vdots & \ddots & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \tau_{u_{0,\left(\overline{M}_P^{(H)},P\right)}}^{(1)(H)} & \cdots & \cdots & \cdots & \tau_{u_{\overline{M}_P^{(H)},P}}^{2(H)} \end{bmatrix} \tag{2.14}$$

$$\begin{bmatrix} \varepsilon_{ij}^{(1)} \\ \vdots \\ \varepsilon_{ij}^{(H)} \end{bmatrix} \sim N(\mathbf{0}, \Omega_\varepsilon), \; \Omega_\varepsilon = \begin{bmatrix} \sigma_\varepsilon^{2(1)} & \cdots & \sigma_\varepsilon^{(1)(H)} \\ \vdots & \ddots & \vdots \\ \sigma_\varepsilon^{(1)(H)} & \cdots & \sigma_\varepsilon^{2(H)} \end{bmatrix} \tag{2.15}$$

In (13), $B(\cdot)$ and $\beta(\cdot)$ represent basis functions and B-spline coefficient, respectively. Random effect splines are defined in (2.14). For a particular outcome, the patient level random effects present each patient's deviance from the average intercept $u_{0,j}$ and from

the average slope of each the splines $(u_{1,j}, \dots, u_{m_1,p,j})$. The patient level covariance matrix includes the pairwise covariances between each set of spline random effects for the intercept and slopes within each of the response variables as well as between the response variables. The patient level residuals are defined in (2.15) with covariance structure Ω_ε. Although covariances described in (2.14) and (2.15) are in general unstructured format, special forms such as Toeplitz or Kronecker type structure can be taken based on different applications.

Following section 2.2, the interaction effect between patient level and team level covariates is modeled with varying coefficient $h(x_{p,ij})z_{q,j}$ if z is categorical, or through nonparametric two dimensional surface fitting of $f(x_p, z_q)$ by the tensor product of two univariate B-splines as in (2.13) if z is continuous. If variable selection is not looked at, the most commonly used priors for the latter case is established on the next four nearest neighborhood on a regular lattice as

$$\beta^{(h')}_{m_1 m_2 pq} \Big| \cdot \sim N\left(\frac{1}{4}\left(\beta^{(h')}_{(m_1-1)m_2pq} + \beta^{(h')}_{(m_1+1)m_2pq} + \beta^{(h')}_{m_1(m_2-1)pq} + \beta^{(h')}_{m_1(m_2+1)pq}\right), \frac{\tau^{2(h')}_{pq}}{4}\right) \tag{2.16}$$

for $m_1 = 2, \dots, \overline{M}_p{}^{(h')} - 1, m_2 = 2, \dots \underline{M}_q{}^{(h')} - 1$, that can be seen as a direct generalization of a first-order random walk in one dimension. Other types of priors such as Kronecker product of penalty matrices of the main effects $\boldsymbol{K}_{pq,j} = \boldsymbol{K}_{p,j} \otimes \boldsymbol{K}_{q,j}$ can also be applied [48].

2. 3. 2. Relationship with a structural equation model

Here we show how the multilevel spline model with a multivariate response can equivalently be represented and estimated in the structural equation modeling framework. For simplicity we choose a model with only level-2 predictors, but this can be extended to more general cases with higher-level predictors and possible interactions such as the one we developed in previous section. In addition, we pick the linear spline model as a special case to help better understand the approach, but this can easily be generalized to other types of splines like the one we exploit in this paper.

Generally structural equation models (SEM) involve two specific parts with distinct objectives: a measurement equation and a structural equation [48]. In the measurement equation, each of the responses $y_j^{(h')}$ loads on the latent variables $f_m^{(h')}, m = 0,1,\dots,M^{(h')}$. The intercept term for response h' is $f_0^{(h')}$ and the loadings for any of the measurements $y_j^{(h')}$ on this latent variable are 1. The other $M^{(h')}$ factors serve as the slopes for each piece on domain x_p defined by the linear splines

$$s_{m,pj}^{(h')} = \begin{cases} 0 & \text{if } s_{pj} \le s_{(m-1),p}^{(h')} \\ s_{pj} - s_{(m-1),p}^{(h')} & \text{if } s_{(m-1),p}^{(h')} < s_{pj} \le s_{m,p}^{(h')} \\ s_{m,p}^{(h')} - s_{(m-1),p}^{(h')} & \text{if } s_{pj} > s_{m,p}^{(h')} \end{cases} \tag{2.17}$$

Applying the same $M^{(h')}+1$ pieces, $m = 0,1,\dots,M^{(h')}$, as above, the measurement equation can be written as

$$y_j^{(h')} = f_0^{(h')} + \sum_{m:s_{m,p}^{(h')} \le s_{pj}} \left(s_{m,p}^{(h')} - s_{(m-1),p}^{(h')}\right) f_m^{(h')} + \sum_{m:s_{(m-1),p}^{(h')} < s_{pj} < s_{m,p}^{(h')}} \left(s_{pj} - s_{(m-1),p}^{(h')}\right) f_m^{(h')} + \sum_{m:s_{(m-1),p}^{(h')} \ge s_{pj}} 0 f_m^{(h')} + \varepsilon_j^{(h')}, \text{ for } h' = 1, \dots, H; \text{for } p = 1, \dots, P. \quad (2.18)$$

It is noticed that any rescaling of (2.18) proportional to the loadings can be employed as well. To see how this is equivalent to multilevel spline model, an additional subscript showing patients, i, is included and $\left(\beta_{m,p}^{(h')} + u_{m,pj}^{(h')}\right)$ is substituted for each of the $f_m^{(h')}$. This gives

$$y_{ij}^{(h')} = \left(\beta_0^{(h')} + u_{m,pj}^{(h')}\right) + \sum_{m:s_{m,p}^{(h')} \le s_{pj}} \left(s_{m,p}^{(h')} - s_{(m-1),p}^{(h')}\right)\left(\beta_{m,p}^{(h')} + u_{m,pj}^{(h')}\right) + \sum_{m:s_{(m-1),p}^{(h')} < s_{pj} < s_{m,p}^{(h')}} \left(s_{pj} - s_{(m-1),p}^{(h')}\right)\left(\beta_{m,p}^{(h')} + u_{m,pj}^{(h')}\right) + \sum_{m:s_{(m-1),p}^{(h')} \ge s_{pj}} 0 \left(\beta_{m,p}^{(h')} + u_{m,pj}^{(h')}\right) + \varepsilon_{ij}^{(h')}$$

$$= \left(\beta_0^{(h')} + u_{m,pj}^{(h')}\right) + \sum_{m=1}^{M^{(h')}} s_{m,pj}^{(h')} \left(\beta_{m,p}^{(h')} + u_{m,pj}^{(h')}\right) + \varepsilon_{ij}^{(h')}, \text{for } h' = 1, \dots, H; \text{for } p = 1, \dots, P \quad (2.19)$$

which can be derived from (2.13) with spline $B(\cdot)$ defined in (2.17) and excluding terms that contain level-3 covariates z_q's.

The structural equation of the SEM characterizes the mutual relationships between the factors. It can be shown that the coefficient for the univariate relationship between any two factors, $f_{m_1}^{(h_1)}$ regressed on $f_{m_2}^{(h_2)}$ is identical to that between two random effects $u_{m_1,p}^{(h_1)}$ and $u_{m_2,p}^{(h_2)}$ by substituting $\left(\beta_{m,p}^{(h')} + u_{m,p}^{(h')}\right)$ for each of the $f_m^{(h')}$ as

$$\beta_{m_1 m_2}^{(h_1)(h_2)} = \frac{\mathrm{cov}\left(f_{m_1}^{(h_1)}, f_{m_2}^{(h_2)}\right)}{\mathrm{var}\left(f_{m_2}^{(h_2)}\right)} = \frac{\mathrm{cov}\left(\beta_{m_1,p}^{(h_1)} + u_{m_1,p}^{(h_1)}, \beta_{m_2,p}^{(h_2)} + u_{m_2,p}^{(h_2)}\right)}{\mathrm{var}\left(\beta_{m_2,p}^{(h_2)} + u_{m_2,p}^{(h_2)}\right)} = \frac{\mathrm{cov}\left(u_{m_1,p}^{(h_1)}, u_{m_2,p}^{(h_2)}\right)}{\mathrm{var}\left(u_{m_2,p}^{(h_2)}\right)} = \frac{\tau_{m_1 m_2,p}^{(h_1)(h_2)}}{\tau_{m_2,p}^{2(h_2)}}, \quad (2.20)$$

in which the numerator and denominator can be found from (2.14). Similarly, other regression coefficients derived from the relationship between factors can be demonstrated to be equal to those between random effects.

A number of works have investigated the equivalence of linear multilevel models and SEMs in the literature [49]. Yet, it should be pointed out that nonlinear multilevel models and generalized linear multilevel models do not always have identical parameterization within the SEM framework. Our goal here is to provide a basis for replacing a multivariate linear multilevel spline model with a standard SEM so that specific strengths of SEM analysis can be captured and they might help improve upon our multilevel analysis. Examples of such strengths may include ability to explicitly model measurement errors through multiple indicator latent factors, and testing within-level and across-level mediation, which are not straightforward in multilevel analysis. Also our attempts here can further be utilized in a way to parameterize and estimate generalized STAR models within a standard SEM framework.

2. 3. 3. Bayesian function selection

In real world data sets with complex hierarchical structure, choosing a suitable subset among many potential predictors and at the same time determining their appropriate shapes (smooth vs. linear) and interaction effects is a challenging and important task. For example in our case of the VA medical home study, we want to select a small group from

a set of 30 comorbidity indicator variables and to decide whether the effect of patient's age and patient's care assessment need score on the response variables are nonlinear or linear, whether an interaction between age and some of the comorbidities is required, and whether a district-specific heterogeneity arising from the location of medical facilities is necessary. To this end, we apply spike-and-slab prior structure for selecting single effect variables as well as grouped coefficients combined with smoothing parameters that represent particular model terms. The main idea of such an approach is to assume a mixture prior for each $\beta_j = (\beta_{j1}, \dots, \beta_{jM_j})'$ with one part being a narrow spike around the origin that imposes very strong shrinkage on the coefficients and the other part being a wide slab that forces very little shrinkage on the coefficients [50] . The posterior mixture weights for the spike (or slab) component of a specific coefficient or coefficient batch can be interpreted as the posterior probability of its exclusion from (or inclusion in) the model.

According to Section 3.1, we note that any multi-response hierarchical STAR model of form (13) can be written in a unifying form $\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\eta} = \boldsymbol{\eta}_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \cdots + \boldsymbol{X}_p\boldsymbol{\beta}_p$ with $\boldsymbol{\eta}_0$ showing offset terms (e.g., grand means of multivariate responses) and effects that are not under selection procedure. Then the conventional spike-and-slab prior structure is given by the following hierarchical Bayesian model

$$\beta_j \big| \delta_j, \rho_j^2 \overset{\text{prior}}{\sim} N\left(0, v_j^2\right) \text{ with } v_j^2 = \rho_j^2 \delta_j,$$

$$\delta_j \big| \omega \overset{\text{prior}}{\sim} \omega I_1\left(\delta_j\right) + (1-\omega) I_{v_0}\left(\delta_j\right),$$

$$\rho_j^2 \overset{\text{prior}}{\sim} \Gamma^{-1}(a_\rho, b_\rho),$$

$$\text{and } \omega \overset{\text{prior}}{\sim} \text{Beta}(a_\omega, b_\omega). \quad (2.21)$$

This structure is called Normal-mixture of inverse Gammas (NMIG) prior that places a bimodal prior on the hyper-variance v_j^2 of the coefficients that leads to a spike-and-slab type prior on the STAR coefficient themselves. $I_z(\cdot)$ is an indicator function that takes 1 in z and zero otherwise and v_0 is a very small positive constant. This way, δ will be 1 with probability ω and close to zero with probability $(1-\omega)$. Hence, the implied prior for (hyper-) variance v_j^2 is a bimodal mixture of inverse Gamma distributions, with one part focused on very small values—the spike with $\delta_j = v_0$—and a second diffuse part with more mass on larger values—the slab with $\delta_j = 1$. The mixture weights ω , in addition, follows a Beta prior that captures any prior knowledge about the sparsity of coefficient β_j [46].

It is found that prior structure (2.21) does not work well for coefficient batch in the STAR models which are associated with spline basis functions or random effects. Briefly, the problem is that a small hyper-variance for a batch of coefficient entails small coefficient values and vice versa. This problematic dependence between a vector of coefficients and their associated hyper-variances causes MCMC sampler unlikely to switch between basins of attraction around the two spike and slab modes. To reduce the dependence, a multiplicative parameter expansion for β_j is recommended that improves the mixing properties of δ_j and boosts the shrinkage characteristics of the resulting prior

compared to (21). The idea is to expand β_j as $\boldsymbol{\beta}_j = \alpha_j \boldsymbol{\Xi}_j$ where scalar $\alpha_j \overset{\text{prior}}{\sim} \text{NMIG}(v_0, \omega, a_\rho, b_\rho)$ is given as (2.21) and it is independent of $\boldsymbol{\Xi}_j$. Elements of the M_j-dimensional vector $\boldsymbol{\Xi}_j$ are then assigned as

$$\Xi_{jm}|r_{jm} \sim N(r_{jm}, 1), \quad r_{jm} \sim \tfrac{1}{2} I_1(r_{jm}) + \tfrac{1}{2} I_{-1}(r_{jm}), \ j = 1, \dots, p; m = 1, \dots, M_j \quad (2.22)$$

which corresponds to a mixture of two i.i.d Gaussian density with mean ± 1 and equal mixture weights. The current approach resolves the mixing problems of δ_j since the Markov blankets of both δ_j and ρ_j now includes only α_j of dimension one instead of the vector $\boldsymbol{\beta}_j$.

The MCMC posterior inference and component selection is performed by a block-wise Metropolis-within-Gibbs sampler which reduces to a standard Gibbs scheme when responses are Gaussian (see the Appendix). The full conditional densities (FDC) for parameters ω, ρ_j^2, δ_j, and conditional means $\boldsymbol{r} = (r_l', l: 1, \dots, \mathcal{L})$ of normal variables $\Xi|r_l \sim N(r_l, 1), r_l = \pm 1$ are given in closed form regardless the choice of exponential family for the responses (see the Appendix). The full conditionals of $\boldsymbol{\alpha}$ and $\boldsymbol{\Xi}$ are based on the conditional design matrices $\boldsymbol{X}_\alpha = \boldsymbol{X}\text{blockdiag}(\boldsymbol{\Xi}_1, \dots, \boldsymbol{\Xi}_p)$ and $\boldsymbol{X}_\Xi = \boldsymbol{X}\text{blockdiag}(\mathbf{1}_{e1}, \dots, \mathbf{1}_{ep})\boldsymbol{\alpha}$, where $\mathbf{1}_e$ is a $e \times 1$ vector of ones and $\boldsymbol{X} = (\boldsymbol{X}_1, \dots \boldsymbol{X}_p)$ is the concatenation of the designs for the model terms as in (2.2). Under the Gaussian assumption of the responses, these are given as follows

$$\boldsymbol{\alpha}| \cdot \sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \text{ where}$$

$$\boldsymbol{\Sigma}_\alpha = \left(\frac{1}{\phi}\boldsymbol{X}_\alpha^T \boldsymbol{X}_\alpha + \text{diag}(\boldsymbol{\delta}\boldsymbol{\rho}^2)^{-1}\right)^{-1}, \boldsymbol{\mu}_j = \frac{1}{\phi}\boldsymbol{\Sigma}_\alpha \boldsymbol{X}_\alpha^T \boldsymbol{y} \quad (2.23)$$

and

$$\boldsymbol{\Xi}|\cdot \sim N(\boldsymbol{\mu}_{\Xi}, \boldsymbol{\Sigma}_{\Xi}) \text{ where}$$

$$\boldsymbol{\Sigma}_{\Xi} = \left(\frac{1}{\phi}\boldsymbol{X}_{\Xi}^{T}\boldsymbol{X}_{\Xi} + \mathbf{I}\right)^{-1}, \boldsymbol{\mu}_{j} = \boldsymbol{\Sigma}_{\Xi}\left(\frac{1}{\phi}\boldsymbol{X}_{\Xi}^{T}\boldsymbol{y} + \boldsymbol{r}\right). \quad (2.24)$$

If the response variables are not Gaussian, the penalized iteratively reweighted least squares (P-IWLS) is used within a Metropolis-Hastings iteration to sample from $\boldsymbol{\alpha}$ and $\boldsymbol{\Xi}$ [46]. The posterior inclusion probability $P(\delta_j = 1|\boldsymbol{y})$ can then be employed to decide upon insignificant, intermediate, and important model terms.

2. 4. Model Specification

The PCMH data is hierarchically organized into three nested levels as shown in Fig.3, where patients are grouped within PCMH teams, and teams are in turn nested within VA facilities. Note that PCMH teams are tied to facilities, i.e., a specific team cannot work at different facilities (teams are nested within facilities). Risk factors can be associated with the response variables at each level while patients from the same team (facility) may have more similar outcomes than patients chosen at random from different teams (facilities). For example, we can study the effects of age (patient-level), PCMH assigned provider's experience (team-level), and type of hospital (facility-level) on the outcomes with nested sources of variability. This setting, in addition to health services research, may happen in many other applications such as educational studies where students are nested within schools and successively within school district. It has been shown that ignoring a level of hierarchy in a data can greatly influence the estimated variances and sensitivity, can seriously inflate Type I error rates [51], and also can result in errors in interpreting the

results of statistical significance tests [52]. As such, multilevel statistical models have been proposed to appropriately analyze the hierarchical (correlated) nesting of data, taking into account the variability associated with each level of the hierarchy [52].

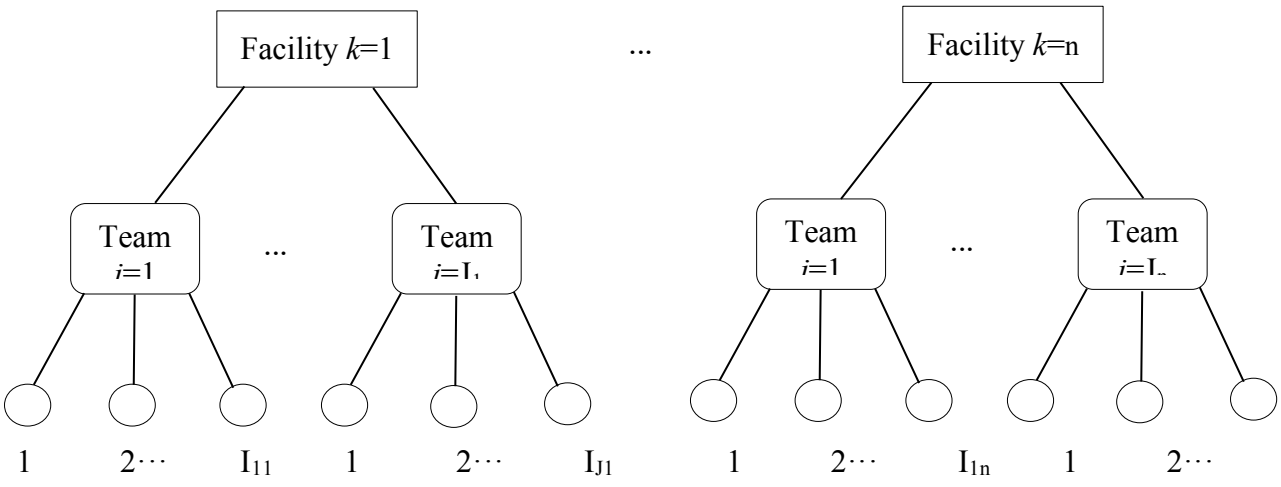


Figure 3 Data structure for PCMH hierarchical model

To simplify, we begin by creating a univariate 2-level generalized linear model (GLM) that predicts the primary care RVU (PCRVU) in each PCMH team with one patient-level (age) and one team-level (assigned provider's experience) predictors. The level-1 model would look like

$$y_{ij} = \beta_{0j} + \beta_{1j}\mathrm{X}_{ij} + e_{ij} \tag{2.25}$$

where y_{ij} is the PC workload for patient i in PCMH team j with an exponential family density of form $f(y \mid x,\phi) = c(y,\phi)\exp\left\{\dfrac{y\theta - b(\theta)}{\phi}\right\}$, β_{0j} is the average PC workload generated in team j, X_{ij} is the patient-level predictor (age) for patient i in team j, and

β_{1j} is its coefficient or slope. The parameters θ and ϕ are called canonical (natural) parameter and scale (dispersion) parameter, respectively. Also $c(\cdot)$ and $b(\cdot)$ are determined by the type of (conditional) distribution under study. This way, we assume that each team has a different (varying) intercept coefficient and a different (varying) slope coefficient. These team-specific coefficients can be specified as either fixed effects or random effects. Treating them as fixed effects, however, leads to a large number of parameters with often very poor estimation results. A more conservative way is to think of them as random variables being modeled by some (level-2) *hyperparameters*. The last term, e_{ij}, is the patient-level error term which is assumed to be normally distributed with covariance structure R. Unlike most methods in the literature, which suppose that the residual variation is the same at the 2-level (teams) and/or the upper levels of hierarchy, we allow unequal variations of the residual to be passed not only on various levels of the hierarchy but also on different response variables.

The next step is to explain the variation of the (level-1) regression coefficients introducing explanatory variables at the team level like

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{oj} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \quad .\end{aligned} \tag{2.26}$$

In this equation, γ_{00} is the grand mean of PC workload across patients and across PCMH teams, γ_{10} is the average effect of the patient-level predictor (age) across all teams, Z_j is the team-level predictor (assigned provider's experience) for team j, γ_{01} and γ_{11} are its (level-2) intercept and slope regression coefficient, and the u-terms are random errors at

the team level, which are assumed to be normally distributed with covariance G. Similar to the R-side covariance matrix, we let these level-2 random errors have unequal variances and also leave them free to be correlated with each other. It is worth pointing out that Z_j in the second line of (2.26) acts as a *moderator* for the relationship between workload and patient age at level-1 analysis; that is, the relationship varies according to the value of the moderator variable. Following the same logic, we can extend this model to add further hierarchies at the facility-level, at the regional level, and so on.

Now a multivariate generalization of this hierarchical GLM is proposed in which both PC and Non-PC workloads are predicted simultaneously. There are several advantages of using a multivariate approach instead of univariate method [21]. One is that the multivariate analysis can better control the type I error rate compared to carrying out a series of univariate statistical tests. Second, this approach can shrink the prediction interval of the dependent variables to a large extent when compared to predicting one of them in isolation. Also using a multivariate scheme, the covariance structure of the responses can be decomposed over the separate levels of hierarchy, which can be of much value for multilevel factor analysis.

Suppose we have *P* response variables and let Y_{hijk} be the workload on outcome *h* (PC or Non-PC workload here) of patient *i* in PCMH team *j* and facility *k*. Here we put the measures (responses) on the lowest level of hierarchy, and represent the different outcome variables by defining *P* dummy variables like

$$d_{phijk} = \begin{cases} 1 & p = h \\ 0 & p \neq h \end{cases} . \qquad (2.27)$$

Then we formulate the lowest level as

$$Y_{hijk} = \pi_{1ijk} d_{11ijk} + \pi_{2ijk} d_{22ijk} + \ldots + \pi_{pijk} d_{ppijk} \quad , \tag{2.28}$$

in which neither the usual intercept nor the error term exists as before. The reason for this is that we solely serve the lowest level as a way to define the multivariate structure using dummy variables. Then following (2.25), we may use π-terms to employ regression equations at the patient level

$$\pi_{pijk} = \beta_{p0jk} + \beta_{p1jk} X_{pijk} + e_{pijk} \tag{2.29}$$

in which a separate index is utilized for denoting the dependent variable of interest. It is noted that with this approach one can fit different intercepts and slopes for different response variables and allow them to vary across any levels of hierarchy. Following (2.26), at the team level, we can have

$$\begin{aligned} \beta_{p0jk} &= \gamma_{p00k} + \gamma_{p01k} Z_{jk} + u_{p0jk} \\ \beta_{p1jk} &= \gamma_{p10k} + \gamma_{p11k} Z_{jk} + u_{p1jk} \quad , \end{aligned} \tag{2.30}$$

where we introduce our 2-level predictors (level-1 moderators) along with random intercepts and slopes and finally link them to the facility level equations by

$$\begin{aligned} \gamma_{p00k} &= \lambda_{p000} + \lambda_{p001} W_k + u_{p00k} \\ \gamma_{p01k} &= \lambda_{p010} + \lambda_{p011} W_k + u_{p01k} \\ \gamma_{p10k} &= \lambda_{p100} + \lambda_{p101} W_k + u_{p10k} \\ \gamma_{p11k} &= \lambda_{p110} + \lambda_{p111} W_k + u_{p11k} \quad . \end{aligned} \tag{2.31}$$

Keeping on this way, one can straightforwardly extend the model to include more predictors at each level and study the effects of fixed and random parameters at any given

point. Another advantage of such modeling is that we can impose an equality constraint across all response variables to build a specific relation with certain effects. For example, we can force level-1 regression coefficients for p=1 (PC workload) and p=2 (Non-PC workload) to be equal by adding the constraint $\beta_{1\,1\,j\,k} = \beta_{2\,1\,j\,k}$. This makes the new model nested within the original model, and thus we can test whether simplifying the model is justified, using a chi-square test on deviances. Plus, if the predictor has random components attached to it, a similar approach would apply to the random part of the model.

At this point, we specify the structure of random components in the model. As shown, we have two random parts in our method: first is the level-1 residual errors as appear in (2.29) by e-terms, and second relates to (higher level) varying intercepts and slopes introduced by u-terms in (2.30) and (2.31). We denote the covariance matrix of the former as R and the latter as G and then assume that both are normally distributed with

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$
$$\mathrm{Var}\begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix} \quad . \tag{2.32}$$

As illustrated, the residual and random parameters are independent having zero means. Generally G and R matrices are large and square with dimensions equal to the number of random coefficients and residuals. While several structures such as spatial or compound symmetry can be thought to formulate those, here we propose an unstructured

parameterization tactic by taking the Kronecker product of their decomposed matrices, named **Parametric** and **Structured**, as

$$\mathbf{G} = \begin{bmatrix} \mathbf{P}_1 \otimes \mathbf{S}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}_2 \otimes \mathbf{S}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathrm{O} \end{bmatrix} \quad . \tag{2.33}$$

At the moment we focus on G decomposition, but a same logic is applied to R. In (2.33), $\otimes$ shows the Kronecker (direct) product; $\mathbf{P}$-terms represent the Parametric part, which is low dimension and needs to be estimated by data; $\mathbf{S}$-terms stands for Structured part, which is typically high dimensional and assumed as known; and zero-off diagonals express the independence among components (see [53] for use of Kronecker product in modeling covariance structures). Note that in its simplest case such as general linear models, where the Parametric matrix is reduced to scalars and the Structured part is taken as identity matrices, equation (2.33) will reduce to the previously known formula $\mathbf{G} = \mathbf{P} \otimes \mathbf{S} = \sigma^2 \mathbf{I}$. Thus we can imply (2.33) as a generalization for covariance functions of other linear statistical models.

To better describe the structure in (2.33), we present examples from our case study. Suppose that we are interested to know whether the identity of a VA facility introduces dissimilar amounts of workload variations. Thus we may construct the top left part of (2.33) like

$$\mathbf{P}_{\text{Facility}} \otimes \mathbf{S}_{\text{Facility}} = \begin{bmatrix} \sigma^2_{\text{PCRVU}} & \sigma_{\text{PCRVU,Non-PCRVU}} \\ \sigma_{\text{Non-PCRVU,PCRVU}} & \sigma^2_{\text{Non-PCRVU}} \end{bmatrix} \otimes \mathbf{I} \tag{2.34}$$

which permits heterogeneous variances across workloads (main diagonal) along with their possible correlation (off-diagonal), and further postulates that the facilities are independent to each other (with the identity matrix). So at the worst case for fitting (2.34), we need 3 degree-of-freedom (DF) to estimate three different elements from the parametric matrix. Further, we may suspect that it is better to fit age (level-1 predictor) with varying intercept and slopes presented by different teams as

$$\mathbf{P}_{\text{Team}} \otimes \mathbf{S}_{\text{Team}} = \begin{bmatrix} \sigma^2_{\text{(Intercept)}} & \sigma_{\text{(Intercept),Age}} \\ \sigma_{\text{Age,(Intercept)}} & \sigma^2_{\text{Age}} \end{bmatrix} \otimes \mathbf{I} \tag{2.35}$$

where the (1:1) element is the amount of variation in regression intercepts among different teams, the (2:2) element is the amount of variation in regression slopes introduced by the patient age across teams, and as before the identity matrix expresses the independence among PCMH teams. Here the model specification is completed and in the next part we explain the model fitting and inference in a Bayesian framework.

2. 5. Estimation and Inference

Before describing model inferences, we give another but equivalent description of our proposal. By substituting equation (2.26) into equation (2.25) and rearranging the terms, we have

$$y_{ij} = \gamma_{00} + \gamma_{10}\mathrm{X}_{ij} + \gamma_{01}Z_j + \gamma_{11}\mathrm{X}_{ij}Z_j + u_{1j}\mathrm{X}_{ij} + u_{oj} + e_{ij} \tag{2.36}$$

in which two distinct segments can be implied: the first is $\left[\gamma_{00} + \gamma_{10}\mathrm{X}_{ij} + \gamma_{01}Z_j + \gamma_{11}\mathrm{X}_{ij}Z_j\right]$, which we call the deterministic part, and the second

is $\left[u_{1j}\mathrm{X}_{ij} + u_{oj} + e_{ij}\right]$, which we call the stochastic part. That way, the moderator effect of (26) is expressed as *cross-level* interaction $\mathrm{X}_{ij}Z_j$ and the multiplication $u_{1j}\mathrm{X}_{ij}$ directly reveals that the error is different for different values of X_{ij} (heteroscedasticity). Taking a matrix form, we may rewrite the right-hand-side of (2.10) as $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\varepsilon}$, where $\mathbf{X}$ and $\mathbf{W}$ are the design matrices for deterministic and stochastic parts. Then the left-hand-side of (2.10), conditional on the stochastic, shapes a GLM response of $g\left(E\left[\mathbf{Y}\middle|\boldsymbol{\varepsilon}\right]\right)$, where $g(\cdot)$ is a differentiable monotonic link function that allows the outcomes to possess any member of the exponential class of distributions. Now assuming a density function of $q^p(\boldsymbol{\varepsilon}^p;\upsilon^p)$ for the stochastic part of the p^{th} response variable ($p = 1,2,\cdots,P$), we can make inferences about the unknown parameters by maximizing the marginal likelihood

$$L(\boldsymbol{\gamma},\upsilon,\phi \mid \mathbf{Y}) = \int\prod_{p=1}^{P} f^p(\mathbf{Y}^p \mid \theta^p,\phi^p)\; q^p(\boldsymbol{\varepsilon}^p;\upsilon^p)\; d\boldsymbol{\varepsilon}^p \quad , \tag{2.37}$$

where $\boldsymbol{\gamma} = \left[\boldsymbol{\gamma}^1,\boldsymbol{\gamma}^2,\cdots,\boldsymbol{\gamma}^P\right]$ is the vector of deterministic coefficients, $q^p(\boldsymbol{\varepsilon}^p;\upsilon^p)$ is a multivariate Gaussian distribution of dimension P with mean zero and variance-covariance υ^p, and ϕ^p and θ^p are the GLM scale and canonical parameters, respectively.

Generally two basic methodologies have been expressed in the literature for optimizing a univariate version of (2.37): the first one tries to approximate the model based on linearization and pseudo-data with fewer nonlinear components, such as the

pseudo-likelihood technique [54]. The second category consists of integral approximation methods that attempt to approximate the log likelihood of (2.37), such as adaptive Gaussian quadrature [55]. But both approaches have some key drawbacks that, we think, cause them inappropriate for our study context. For example, a true objective function for the overall optimization does not exist in the first class; thus it potentially produces estimates that are inconsistent under standard (small domain) asymptotic assumptions. Additionally, the bias size can be substantial in the case of major variance components or few observations per participant. Similarly, methods in the second approach cannot accommodate R-side covariance structure such as over dispersion parameter. These problems also become more crucial when more than one outcome needs to be estimated [54].

Due to this, we decide to put forward a Bayesian framework that utilizes an exact maximum likelihood approach by numerical integration techniques. To this end, we need to first determine suitable priors for the parameters of interest then employ a simulation-based integration technique, such as Metropolis-Hastings or slice sampling, to iteratively sample the posterior until convergence. Afterwards, generated samples are used to estimate the approximate expectations of quantities of interest. However, setting up the appropriate priors can greatly affect inference about posteriors, because in many cases, *diffuse* priors and/or *improper* priors lead to improper posteriors upon which no valid inference can be made [56]. Accordingly, for the deterministic coefficient vector $\boldsymbol{\gamma}^{p}$, we use a Gaussian prior of form $N\left(\boldsymbol{\gamma}_0,\boldsymbol{\Gamma}\right)$. Moreover, to sample from $\boldsymbol{\eta}$, since its distribution cannot be identified, we apply the Metropolis-Hastings update of Damlen et

al. [57]. In summary, the method is updating $\boldsymbol{\eta}$ in some blocks; each consists of groups of residuals expected to have some form of residual co-variation as defined by the R structure. That way, the conditional density of $\boldsymbol{\eta}^p$ is formulated as

$$f(\boldsymbol{\eta}_l^p \mid \mathbf{Y}^p;\boldsymbol{\gamma}^p,\boldsymbol{\varepsilon}^p) \propto \prod_{i\in l} p_i(\mathbf{Y}_i^p \mid \eta_i^p) f_N^p(\mathbf{e}_l \mid \mathbf{0},\mathbf{R}_l) \tag{2.38}$$

where l stands for blocks of $\boldsymbol{\eta}^p$ with non-zero residual covariances, f_N^p indicates a conditional multivariate normal distribution for the linear predictor residuals, and $p_i(\mathbf{Y}_i^p \mid \eta_i^p)$ is the probability of data point $\mathbf{Y}_i^p$ (from p^{th} outcome) with linear predictor $\boldsymbol{\eta}_l^p$.

In order to update the parameter vector $\boldsymbol{\rho}^p = [\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\varepsilon}^{\mathrm{T}}]^{\mathrm{T}}$, the single-block Gibbs sampler of García-Cortés, Sorensen [26] is applied. Essentially, the method solves the sparse linear system of $\hat{\boldsymbol{\rho}} = \mathbf{A}^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{R}^{-1}(1 - \mathbf{M}\boldsymbol{\rho}_*^p - \mathbf{e}_*^p)$ using Cholesky decomposition technique. In the formula, $\mathbf{A}$ is the coefficient matrix of form $\mathbf{A} = \mathbf{M}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{M} + \begin{bmatrix} \boldsymbol{\Gamma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} \end{bmatrix}$, in which $\mathbf{M} = [\mathbf{X}\ \mathbf{W}]$ is the whole design matrix, $\boldsymbol{\Gamma}$ is the prior (co)variance matrix for the deterministic part, and $\left\{\boldsymbol{\rho}_*^p, \mathbf{e}_*^p\right\}$ are random realizations drawn from multivariate normal distributions $\boldsymbol{\rho}_*^p \sim N\left(\begin{bmatrix} \boldsymbol{\gamma}_0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{bmatrix}\right)$ and $\mathbf{e}_*^p \sim N(\mathbf{M}\boldsymbol{\rho}_*^p, \mathbf{R})$ respectively. Based on these, the desired prior sample of $f(\boldsymbol{\rho}^p \mid \boldsymbol{\eta}^p;\mathbf{M},\mathbf{R},\mathbf{G})$ is given by $\hat{\boldsymbol{\rho}} + \boldsymbol{\rho}_*^p$.

For taking samples of the variance structures $\mathbf{R}$ and $\mathbf{G}$, we need the sum of squares matrix associated with each diagonal component of (2.33). This is given by $\mathbf{H} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{\Phi}$, where $\mathbf{\Phi}$ is a stochastic matrix in which each column is related to the relevant row/column of Parameteric matrix $\mathbf{P}$ and each row is associated with the related row/column of Structured matrix $\mathbf{S}$. In this way, $\mathbf{P}$ can be Gibbs sampled in one block from the *Inverse-Wishart* (IW) distribution $\mathbf{P} \sim IW((\mathbf{H}_p + \mathbf{H})^{-1}, n_p + n_\Phi)$, where n_Φ is the number of rows in $\mathbf{\Phi}$, $\mathbf{H}_p$ is the prior sum of squares, and n_p is its degrees of freedom. It should be noted that IW is a conjugate prior for the covariance matrix of a multivariate normal distribution.

Usually the goodness-of-fit of Bayesian models can be assessed using the deviance information criterion (DIC), which is a Bayesian alternative to AIC and Schwarz criterion. The DIC can be calculated at different levels of hierarchy and a smaller amount indicates a better fit to the data while compensating for model complexity. Here, we adopt the method of Spiegelhalter et al. [57] and define the *deviance* as $D = -2\log(\Pr(\mathbf{Y} \mid \mathbf{\Omega}))$, where $\mathbf{\Omega}$ are some parameters of the model. We calculate this probability for the lowest level of the hierarchy at each iteration. In the formula, in case of Gaussian responses we have $\mathbf{\Omega} = \{\boldsymbol{\rho}, \mathbf{R}\}$ and the likelihood would be the normal density $f_N(\mathbf{Y} \mid \mathbf{X}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\varepsilon}, \mathbf{R})$. On the other hand, when the responses are not normal, $\mathbf{\Omega} = \boldsymbol{\eta}$ and the likelihood would change to $\prod_i f_i\left(\mathbf{Y}_i^p \mid \boldsymbol{\eta}_i^p\right)$, where the argument denotes the conditional probability of the i^{th} data point (lowest level of hierarchy). In other words,

for non-Gaussian responses, deviance is obtained by the probability of the data given the linear predictor $\boldsymbol{\eta}$, whereas in normal responses, it is calculated using the probability of the data given the parameters. The DIC can then be attained by $\mathrm{DIC} = 2\bar{D} - D(\bar{\boldsymbol{\Omega}})$, where $\bar{D}$ is the mean deviance of all iterations and $D(\bar{\boldsymbol{\Omega}})$ is the deviance evaluated at the mean estimates of the parameters.

Chapter III Methodology on Phase II

In section 1.1, we described two different phases for implementing Patient Centered Medical Home (PCMH), and in this section we present a stochastic programming model with recourse for the second phase, patient assignment.

3.1 Model Assumptions

Here, we presume the following assumptions in our proposed model:

Assumption A1 Patient assignment in PCMH teams is completed in two stages.

We begin by an initial panel including anticipated patients who, we think, ask for care from the PCMH teams. Since we do not know this panel is fixed within the planning horizon, the assignment in the first stage is provisional with tentative cost ($\tilde{c}$). Then, the second stage is started in which actual patient's care demands become known. If the demands for each profession line cannot be met by the available capacity, some of patients are reassigned to under loaded/backup members at a specific cost.

Assumption A2 Excessive workload for each PCMH profession beyond a given limit is supported by backup professionals.

Excess workload is the difference between the demanded workload and the time available for care. Some researches indicate that excessive workload decreases the quality of patient care [58].x In our case, there is a threshold on excessive workload for each profession line specified by the VHA. The extra workload beyond this limit results in other backup professionals assisting overloaded professionals while a penalty is paid for each unit of extra workload below the limit.

Assumption A3: The patient panel to be assigned to PCMH teams includes unforeseen outpatients, so the number of patients is fixed.

During a planning horizon, patients may leave the panel by switching to other healthcare systems or death. Also new patients may enter the panel by transferring from other healthcare systems or direct admissions, and will need an assignment to a team. One way to deal with this situation is to periodically update the patient pool and remaining hours on personnel. Then the model has to be solved at regular time intervals with new patient pool. Instead, we can represent the actual panel by multiplying the number of registered patients by a factor to account for unforeseen patients. This latter approach is taken in our present work.

Assumption A4: The composition of medical home team is known *a priori*.

Veterans are at the center of their medical homes, which also includes their families and caregivers. In Veteran Health Administration (VHA), healthcare professionals on the team include a primary care physician (PCP) or nurse practitioner (NP), a registered nurse (RN) who serves as the care manager, a clinical staff assistant who is usually a licensed practical nurse (LPN), and an administrative clerk. When additional services are needed to meet the Veteran's goals and needs, another care team may be called in. These may include social workers, dietitians, pharmacists, mental health practitioners, specialists, and other non-VA health care professionals.

Assumption A5: Staffing levels in the medical home teams can be regarded as constant and purely exogenous.

The supply of health services can be given relatively easy based on head counts and available service hours from all professional lines within a specific planning period. That is, the healthcare supply is a portfolio of total available hours by various members in a team within a particular period which includes total PCP time, total RN time, total clerk time, etc. This includes hiring and firing of all temporary personnel too. The counted service times incorporate the productivity factor per full-time equivalent (FTE) hours. Here we assume a deterministic ratio of productive and contractual FTE per professional. Also capacity levels are assumed to be independent of demand for care, quality of care provided or any other controlled variables in the model.

Assumption A6: Healthcare demands on professional lines are dependent random variables.

In medical home model, demands are in form of a portfolio, i.e. a vector of continuous random variables composed of stochastic demands on each team member. Demands are measured on a yearly basis in relative value unit [59]. We assume that, for a given patient, demands generated on different team members are dependent. This means that, for a particular instance of the problem, demand realizations on PCP and RN may be correlated. Thus, to generate problem instances for our computational study, we use a multivariate workforce prediction algorithm that can take into account such dependencies among demands [60].

Assumption A7: Shortages on health services depend only on demand and contracted capacities.

Assumption A8: There is no difference in efficiency among similar professionals.

It is assumed that professional lines (such as RNs or clerks) among teams are identical in terms of efficiency and quality of care provided.

Assumption A9: No coordination exists among identical professionals.

We assume that care needed by a patient cannot be split among multiple professionals in different teams. In other words, identical professionals in different teams (e.g., two LPNs in two medical home teams) cannot coordinate their tasks among each other.

3.2 Stochastic model for patient assignment

Our proposed model consists of finding the optimum allocation of a set of patients to medical home teams such that each patient is assigned exactly to one team, subject to resource constraints limiting teams' workload capacity to handle patients. The problem is modeled as a two-stage stochastic program with mixed 0-1 recourse. The first-stage decisions involve assigning an initial panel to the PCMH teams well ahead in time. These assignments are associated with a tentative cost that can be estimated or given as constant. The second-stage decisions are related to the adjustments that are made after first-stage decisions and once we get closer to the actual demand realizations. These include possible patient reassignments and overtime capacities used in excess of available service hours on each professional line in each team under each possible scenario.

Our first two-stage stochastic patient assignment model is presented by the following notations where bold face represents vectors throughout:

Indices

$i \in I$: index for patients

$j \in J$: index for medical home teams

$r \in R$: index for professional lines within each team

$\omega \in \Omega$: index for scenarios, Ω denotes the sample space of the underlying probability triple

Fixed Model Parameters

c_{ij}: cost for team j to deliver care to patient i

b_{jr}: regular time available on profession r in team j

β_r: marginal overtime penalty of profession r per each unit of excess time usage

Scenario Dependent Model Parameters

$d_{ijr}(\omega)$: time required by profession r in team j to deliver care to patient i in scenario ω, $\mathbf{d} \in \mathbb{R}_{\geq 0}^{|I||J||R|}$

First Stage Decision Variable

x_{ij}: binary assignment variable representing whether patient i is assigned to medical home team j ($x_{ij} = 1$) or not ($x_{ij} = 0$)

Second Stage Decision Variable

u_{jr}: amount of overtime on professional r within team j used in excess of its available capacity

Note that in all stochastic models presented throughout this paper, we assume dependency exists between second stage random parameters and ω, but the dependence of the second stage decision variable on ω is suppressed. Also we presume that Ω has a finite support on the underlying probability space. Using the above notations, the mathematical formulation of the first model (TSSPA1) can be written as follows.

(TSSPA1 model)

First Stage Problem:

$$\min_{\mathbf{x}} \ \sum_{i\in I}\sum_{j\in J} c_{ij}x_{ij} + \ \mathcal{Q}(x) \tag{3.1a}$$

$$\text{s.t.} \ \sum_{j\in J} x_{ij} = 1 \quad \forall\, i \in I \tag{3.1b}$$

$$x_{ij} \in \{0,1\} \qquad \forall\, i \in I, \ \forall\, j \in J, \tag{3.1c}$$

where

$$\mathcal{Q}(x) := \ \mathbb{E}_{\mathbf{d}}[Q\big(x, d(\omega)\big)]. \tag{3.2}$$

Second Stage Recourse Problem:

$$Q\big(x, d(\omega)\big) = \min_{\mathbf{u}} \ \sum_{j\in J}\sum_{r\in R} \beta_r u_{jr} \tag{3.3a}$$

$$\text{s.t.} \ \sum_{i\in I}\sum_{j\in J} d_{ijr}(\omega)\, x_{ij} \leq \ b_{jr} + \ u_{jr} \ \ \forall\, j \in J, \ \forall\, r \in R \tag{3.3b}$$

$$u_{jr} \geq 0 \quad \forall\, j \in J, \ \forall\, r \in R. \tag{3.3c}$$

Note that $\mathbb{E}$ stands for mathematical expectation. Generally, the expectation can take the form of utility functions or it may include risk measures. Here $Q\big(x, d(\omega)\big)$ equals the total capacity violation cost given assignment x, the realized healthcare demands $d(\omega)$, and a given recourse policy. The objective function then minimizes the sum of patient initial assignment costs and the expected mismatch cost. Constraints (3.1b) ensure that each patient is assigned to only one medical home team. Binary restrictions on the first-stage variables are defined by (3.1c). The expected recourse function is given in (3.2). The objective function in the second-stage is to minimize the sum of overtime penalties incurred by allocating patients in excess of professional capacities for given first-stage assignment policy x and random vector $\mathbf{d}$. Constraints (3.3b) are the healthcare supply-demand constraints stating that, on each profession within each medical home team, demands can exceed supplied capacities by u_{jr} unit at the cost of β_r per unit. Non-negativity restrictions on the second stage decision variables are defined by (3.3c).

In summary, the TSSPA1 model first assigns patients to medical home teams without full information on the healthcare demands subject to constraints (3.1b) and (3.1c) and with associated cost c_{ij}. Later, when full information about demands become available, we observe a realization of $d(\omega)$ and penalize the sum of supply-demand mismatches over all professionals in all PCMH teams with respective unit penalty $\beta_r \geq 0$. The TSSPA1 model is in the framework of the classical stochastic generalized assignment problem [61] extended for the multi-recourse case when violations on capacity-constraints are allowed.

We can also write the TSSPA1 model as a nonlinear stochastic program like

$$\min_{\mathbf{x}} \ \sum_{i\in I}\sum_{j\in J} c_{ij}x_{ij} + \ \mathbb{E}\left[\sum_{j\in J}\sum_{r\in R}\beta_r\left(\sum_{i\in I}\sum_{j\in J} d_{ijr}(\omega)\, x_{ij} - b_{jr}\right)^{+}\right]$$

$$\text{s.t. (3.1b) and (3.1c)},$$

where $(\cdot)^{+} = \max(\cdot\,, 0)$. Here, the expected mismatch penalty, i.e., the expected values of the recourse function for a given patient assignment $\hat{\boldsymbol{x}}$, is given as

$$\mathbb{E}\left[\sum_{j\in J}\sum_{r\in R}\beta_r\left(\sum_{i\in I}\sum_{j\in J} d_{ijr}(\omega)\,\hat{x}_{ij} - b_{jr}\right)^{+}\right] =$$

$$\sum_{j\in J}\sum_{r\in R}\beta_r\,\mathbb{E}\left[\left(\sum_{i\in I}\sum_{j\in J} d_{ijr}(\omega)\,\hat{x}_{ij} - b_{jr}\right)^{+}\right],$$

in which $\mathbb{E}\left[\left(\sum_{i\in I}\sum_{j\in J} d_{ijr}(\omega)\,\hat{x}_{ij} - b_{jr}\right)^{+}\right]$ is the expected overtime required by profession r in medical home team j.

In our second proposed model we extend the recourse subproblem by allowing reassignments of some patients at a pre-specified cost, provided that the capacity of a professional is violated. To this end, we introduce new binary variable $y_{ij}, (i,j) \in I\times J$ that determines the final patient assignment with actual cost c_{ij}. This cost differs from the first stage provisional cost which we here denote by $\tilde{c}_{ij}$. We also define $z_i, i \in I$ as an auxiliary variable referring to those patients with a non-zero healthcare demand that have been reassigned, and $\alpha_i, i \in I$ as recourse costs for such reassignments. Further, we let $w_i, i \in I$ as a binary parameter whose values is conditional on the value of demand such that patient i brings a non-zero demand ($d_{ijr}(\omega) > 0$) only if $w_i = 1$. The new model (TSSPA2) can then be formulated using the extended recourse function as follows.

(TSSPA2 model)

First Stage Problem:

$$\min_{\mathbf{x}} \ \sum_{i\in I}\sum_{j\in J}\tilde{c}_{ij}x_{ij} + \ \mathcal{Q}(x) \tag{3.4}$$

s.t. (3.1b) and (3.1c),

where $\mathcal{Q}(x) := \ \mathbb{E}_{\mathbf{d}}[Q\big(x, d(\omega)\big)]$, and $Q\big(x, d(\omega)\big)$ is the value of second stage recourse function obtained by solving the following optimization problem:

Second Stage Recourse Problem:

$$Q\big(x, d(\omega)\big) = \min_{\mathbf{y},\mathbf{u},\mathbf{z}} \ \sum_{i\in I}\sum_{j\in J} c_{ij}y_{ij} + \sum_{i\in I}\alpha_i z_i + \sum_{j\in J}\sum_{r\in R}\beta_r u_{jr} \tag{3.5a}$$

$$\text{s.t.} \quad y_{ij} + z_i \geq w_i x_{ij} \qquad \forall\, i \in I, \ \ \forall\, j \in J \tag{3.5b}$$

$$\sum_{j\in J} y_{ij} \geq \ w_i \qquad \forall\, i \in I \tag{3.5c}$$

$$\sum_{i\in I}\sum_{j\in J} d_{ijr}(\omega)\, y_{ij} \leq \ b_{jr} + \ u_{jr} \quad \forall\, j \in J, \ \forall\, r \in R \tag{3.5d}$$

$$y_{ij} \in \{0,1\} \quad \forall\, i \in I, \ \ \forall\, j \in J \tag{3.5e}$$

$$z_i \in \{0,1\} \qquad \forall\, i \in I \tag{3.5f}$$

$$u_{jr} \geq 0 \qquad \forall\, j \in J, \ \forall\, r \in R. \tag{3.5g}$$

The first stage problem in TSSPA2 model is similar to the first stage problem in

TSSPA1 except that the assignment costs in TSSPA1 are replaced with provisional assignment costs $\tilde{c}_{ij}$. The second-stage problem, however, intends to minimize the expected total costs of actual (final) assignments, reassignments, and overtime penalties on professional lines. Constraints (3.5b) set z_i to 1 whenever patient i has a non-zero demand and it is not assigned to the same team it was assigned to *a priori*. In other words, z_i equals to one if ($w_i x_{ij} = 1$ and $y_{ij} = 0$), otherwise it takes zero. Constrains (3.5c) ensure that all patients with non-zero demand are indeed assigned to a medical home team. Integrality requirements on the actual assignment variable y_{ij} and the auxiliary variable z_i are defined in (3.5e) and (3.5f), respectively. All other constraints are the same as the second stage problem in TSSPA1 model. In what follows we focus on the stochastic program defined by the TSSPA2 model as it manages handling patient reassignments in the PCMH workforce planning.

3. 3. Solution Approaches

The stochastic problem TSSPA2 is difficult to solve since it has binary variables in stage one and mixed binary variables in stage two. This implies that the recourse function $Q(x)$ is generally non-convex and lower semi-continuous [62]. In addition, since the model handles excess healthcare demands with a unit overtime penalty and enough capacity in the second stage, it follows that TSSPA2 model is a two-stage program with *relatively complete recourse*. A popular approach to proceed consists of approximating the uncertainty in the healthcare demand by a finite set of scenarios that leads to a decision tree representation of the stochastic model. Here the set of scenarios can be

assumed available [63] or has to be generated by proper methods satisfying special statistical features [64]. Then the model can be reformulated into a large-scale deterministic equivalent program. Commercial solvers have been used directly to such models but due to the amount of memory required, such packages cannot solve practical real-world problems in reasonable times. Thus, decomposition techniques have been proposed in the literature to obtain operational running time by exploiting the block structure of the feasible region defining the extensive form models [65].

There are two general classes of decomposition strategies for stochastic programs: vertical or stage-based, and horizontal or scenario-based [66] The well-known example of the former is the L-shaped method or Bender decomposition [67], and exemplars of the later include progressive hedging algorithm [68] and dual decomposition [69]. Solution approaches based on the L-shaped method approximate the non-linear recourse function by outer linearization using an alternative formulation of (3.4) and (3.1b)-(3.1c). This way, a *master problem* is solved at each iteration, k, to achieve a feasible solution, (x^k, θ^k), where θ^k lower bounds the recourse function $Q(x)$. The solution of the first stage problem is transferred to the second stage subproblems that are solved independently to get the dual solutions. Then optimality cuts are generated from the dual solutions and added to the master problem. These cuts help to lower bound hyperplanes of the recourse function $Q(x)$. The algorithm stops once the optimal solution is found or some pre-specified tolerance is met. A multicut version of the L-shaped method has also been proposed and applied to solve two stage and multistage stochastic programs. It differs from the classical L-shaped method in that it creates an optimality cut for every

single scenario in the second stage. As a result the size of the master problem increases rapidly and it becomes computationally expensive to tackle.

Solution methods based on scenario decomposition, on the other hand, use an alternative formulation of the two-stage problem in which the first stage decision variables $\boldsymbol{x}$ are temporarily indexed by random scenarios. The (augmented) Lagrangian relaxation is applied to all non-anticipativity (or implementability) constraints that ensure feasible solutions are scenario-invariant at each node of the decision tree. The original stochastic problem is then decomposed per ω and the resulting subproblems are independently solved to obtain a general lower bound. One main advantage of such methods over the variants of L-shaped approaches is that no limitations are existed on the number of stages and also on the type of decision variables allowed in each stage – as the case in many proposed mixed-integer stochastic algorithms. Another benefit of these decomposition approaches is that, given a set of scenarios, the difficulty of handling subproblems is more uniform because the underlying partition strategy is scenario-based. This is favorable in parallel computing since the distribution of workload among parallel processing elements would be more stable. However, the L-shaped methods depend highly on convexity assumptions and more importantly the computational burden of the master problem can increase significantly with the number of iterations, while the subproblems are regularly easy to solve.

In our initial numerical experiments we found that the progressive hedging algorithm (PHA) is suitable for our problem. We decided to propose our solution approach based on the PHA since (1) the core PHA strategy is based on augmented Lagrangian, which is not

restricted to the issues of convexity; (2) the subproblems only need to be solved approximately in the algorithm; (3) it is easily implemented and customized in environments where packages for solving deterministic equivalent model already exist; (4) it is proved to be an efficient scalable approach to large-scale mixed-integer stochastic programs in a number of real-world instances [70,71,72];(5) valid lower bounds (and quality of the solutions) for the mixed-integer case can be obtained by using dual prices of the non-anticipativity constraints in any iteration [73]; and (6) it can be easily parallelized [74].

In this paper we consider the following reasonable assumption: healthcare demands are team-independent, that is, for each $i \in I$ and $r \in R$, $d_{ijr}(\omega) = d_{ir}(\omega)\ \forall\, j \in J$. In the next section, inspired by the PHA scheme, we first apply a scenario decomposition technique to separate the stochastic problem TSSPA2 per scenarios of the demand realization. Then we define a reference patient assignment policy and modify the first-stage cost of scenario subproblems that can reflect the difference between each assignment and the reference point. Following this, we present our proposed primal-dual algorithm iteratively computes a reference assignment, updates the fixed costs to seek for a consensus design, performs upper bounding and lower bounding. The last step of the algorithm repeats the same procedure for a reduced problem with the assignments that have not converged to a consensus policy.

3.3.1. Scenario Decomposition

Stochastic program TSSPA2 is generally an infinite dimensional optimization problem. To deal with this, we approximate the problem by considering a finite set of possible

scenarios $\mathcal{S} \subseteq \Omega$ of the random event with corresponding probability mass $p_s, s \in \mathcal{S}$. Doing this, the mathematical expectation is expressed by a probability weighted sum and the problem is represented by a multiscenario deterministic model called the extensive form (EF), as follows.

(EF):

$$\min \quad \sum_{i\in I}\sum_{j\in J}\tilde{c}_{ij}x_{ij} + \sum_{s\in\mathcal{S}} p_s \left(\sum_{i\in I}\sum_{j\in J} c_{ij}y_{ij}^s + \sum_{i\in I}\alpha_i z_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r u_{jr}^s\right) \quad (3.6a)$$

$$\text{s.t.} \quad \sum_{j\in J} x_{ij} = 1 \quad \forall\, i \in I \quad (3.6b)$$

$$y_{ij}^s + z_i^s \geq w_i x_{ij} \quad \forall\, i \in I,\ \forall\, j \in J,\ \forall\, s \in \mathcal{S} \quad (3.6c)$$

$$\sum_{j\in J} y_{ij}^s \geq w_i \quad \forall\, i \in I,\ \forall\, s \in \mathcal{S} \quad (3.6d)$$

$$\sum_{i\in I}\sum_{j\in J} d_{ir}^s\, y_{ij}^s \leq\ b_{jr} + u_{jr}^s \quad \forall\, j \in J,\ \forall\, r \in R,\ \forall\, s \in \mathcal{S} \quad (3.6e)$$

$$x_{ij} \in \{0,1\} \quad \forall\, i \in I,\ \forall\, j \in J \quad (3.6f)$$

$$y_{ij}^s \in \{0,1\} \quad \forall\, i \in I,\ \forall\, j \in J,\ \forall\, s \in \mathcal{S} \quad (3.6g)$$

$$z_i^s \in \{0,1\} \quad \forall\, i \in I,\ \forall\, s \in \mathcal{S} \quad (3.6h)$$

$$u_{jr}^s \geq 0 \quad \forall\, j \in J,\ \forall\, r \in R,\ \forall\, s \in \mathcal{S}, \quad (3.6i)$$

in which the reassignment and overtime variables as well as, now deterministic, healthcare demands are scenario specific. The constraint matrix defining model

(3.6a)-(3.6i) is sparse and exhibits a block-diagonal structure, each block associated to a single scenario $s \in \mathcal{S}$ in the tree. By solving problem (**EF**) one can finds an assignment policy that minimizes the sum of fixed costs and expected second-stage costs over all scenarios. Constraints (3.6c) link the first- and second-stage variables and allow the reassignments to happen only if they are different from the assignments in the first stage.

By making copies of the first-stage decision variables, $x_{ij}^s = \{0,1\}, \forall\, i \in I,\ \forall\, j \in J$, for each scenario $s \in \mathcal{S}$, problem (**EF**) can be reformulated as

(P1):

$$\min \quad \sum_{s\in\mathcal{S}} p_s \left(\sum_{i\in I}\sum_{j\in J} \tilde{c}_{ij} x_{ij}^s + \sum_{i\in I}\sum_{j\in J} c_{ij} y_{ij}^s + \sum_{i\in I} \alpha_i z_i^s + \sum_{j\in J}\sum_{r\in R} \beta_r u_{jr}^s\right) \quad (3.7a)$$

$$\text{s.t.} \quad \sum_{j\in J} x_{ij}^s = 1 \quad \forall\, i \in I,\ \forall\, s \in \mathcal{S} \quad (3.7b)$$

$$y_{ij}^s + z_i^s \geq w_i x_{ij}^s \quad \forall\, i \in I,\ \forall\, j \in J,\ \forall\, s \in \mathcal{S} \quad (3.7c)$$

$$x_{ij}^s = x_{ij}^t \quad \forall\, s,t \in \mathcal{S},\ s \neq t \quad (3.7d)$$

$$x_{ij}^s \in \{0,1\} \quad \forall\, i \in I,\ \forall\, j \in J,\ \forall\, s \in \mathcal{S} \quad (3.7e)$$

$$(6d), (6e), \text{and } (6g)\text{-}(6i),$$

where equations (3.7d) enforce non-anticipativity or implementability constraints. These constraints correspond to a large non-separable block with nonzero coefficients and they

make sure initial patient assignments (first-stage decisions) are not tailored for each specific scenario that might happen. Since the number of such constraints may be too large to affect the convergence rate, for each $s \in \mathcal{S}$ they are replaced with a "global" solution $\bar{x}_{ij} \in \{0,1\}, \forall\, (i,j) \in (I \times J)$. Then augmented Lagrangian relaxation is applied on $x_{ij}^s = \bar{x}_{ij}, \forall\, (i,j) \in (I \times J), \forall s \in \mathcal{S}$ penalizing quadratically any violations of it. The resulting objective function becomes

$$\min \quad \sum_{s\in\mathcal{S}} p_s \Big(\sum_{i\in I} \sum_{j\in J} \tilde{c}_{ij} x_{ij}^s + \sum_{i\in I} \sum_{j\in J} c_{ij} y_{ij}^s + \sum_{i\in I} \alpha_i z_i^s + \sum_{j\in J} \sum_{r\in R} \beta_r u_{jr}^s + \sum_{i\in I} \sum_{j\in J} \lambda_{ij}^s \left(x_{ij}^s - \bar{x}_{ij} \right) + \frac{1}{2} \sum_{i\in I} \sum_{j\in J} \rho \left(x_{ij}^s - \bar{x}_{ij} \right)^2 \Big), \tag{3.8}$$

in which $\lambda_{ij}^s, \forall\, (i,j) \in (I \times J), \forall s \in \mathcal{S}$ denotes the dual variables for the relaxed constraints, and $\rho > 0$ is an external penalty ratio that aims to achieve a consensus among the scenario solutions. In other words, the last two components in (3.8) construct additional costs we pay for the differences between the scenario solutions and the "global" first-stage policy. Given the fact that x_{ij}^s and $\bar{x}_{ij}$ are binary, the objective can be reduced to

$$\min \quad \sum_{s\in\mathcal{S}} p_s \Big(\sum_{i\in I} \sum_{j\in J} \Big(\tilde{c}_{ij} + \lambda_{ij}^s - \rho \bar{x}_{ij} + \frac{\rho}{2} \Big) x_{ij}^s + \sum_{i\in I} \sum_{j\in J} c_{ij} y_{ij}^s + \sum_{i\in I} \alpha_i z_i^s + \sum_{j\in J} \sum_{r\in R} \beta_r u_{jr}^s \Big) - \sum_{i\in I} \sum_{j\in J} \lambda_{ij}^s \bar{x}_{ij} + \sum_{i\in I} \sum_{j\in J} \frac{1}{2} \rho \bar{x}_{ij}. \tag{3.9}$$

Note that the relaxed problem defined with the objective function (3.9) is not separable. However, if the global solution $\bar{x}_{ij}$ is given and fixed, the relaxed formulation

can be decomposed according to each scenario. For a scenario s, the corresponding subproblem, given $\bar{x}_{ij}$, can be expressed as

(P2):

$$\min \quad \left[\sum_{i\in I}\sum_{j\in J}\left(\tilde{c}_{ij}+\lambda_{ij}^{s}-\rho\bar{x}_{ij}+\frac{\rho}{2}\right)x_{ij}^{s}+\sum_{i\in I}\sum_{j\in J}c_{ij}y_{ij}^{s}\right.$$
$$\left.+\sum_{i\in I}\alpha_i z_i^s+\sum_{j\in J}\sum_{r\in R}\beta_r u_{jr}^{s}\right] \tag{3.10a}$$

$$\text{s.t.} \quad \sum_{j\in J}x_{ij}=1 \quad \forall\, i\in I \tag{3.10b}$$

$$y_{ij}^{s}+z_i^s\geq w_i x_{ij} \quad \forall\, i\in I,\ \forall\, j\in J \tag{3.10c}$$

$$\sum_{j\in J}y_{ij}^{s}\geq w_i \quad \forall\, i\in I \tag{3.10d}$$

$$\sum_{i\in I}\sum_{j\in J}d_{ir}^{s}\,y_{ij}^{s}\leq\ b_{jr}+\ u_{jr}^{s} \quad \forall\, j\in J,\ \forall\, r\in R \tag{3.10e}$$

$$x_{ij}\in\{0,1\} \quad \forall\, i\in I,\ \forall\, j\in J \tag{3.10f}$$

$$y_{ij}^{s}\in\{0,1\} \quad \forall\, i\in I,\ \forall\, j\in J \tag{3.10g}$$

$$z_i^s\in\{0,1\} \quad \forall\, i\in I \tag{3.10h}$$

$$u_{jr}^{s}\geq 0 \quad \forall\, j\in J,\ \forall\, r\in R, \tag{3.10i}$$

which takes the form of a deterministic mixed 0-1 formulation identical to (**EF**) with a perturbed first stage cost $\left(\tilde{c}_{ij}+\lambda_{ij}^{s}-\rho\bar{x}_{ij}+\frac{\rho}{2}\right)x_{ij}^{s}$. For addressing these subproblems we rely on the branch and bound algorithm in CPLEX (with the clique cuts and feasibility pump heuristic), though a variety of other exact or heuristic methods exploiting special combinatorial structures can be applied. By solving the subproblem (**P2**), solutions for

different scenarios may be dissimilar. To cope with this issue and to obtain a consensus solution among the subproblems, we proposed the following algorithm.

3.4. Proposed Algorithm

Let ν denote the iteration index of our proposed algorithm. To get an overall solution that is served as a reference point for all scenarios, the average operator given the scenario probabilities was originally suggested by [67]. Let $X(s) \coloneqq \{(7b), (7c), (7e), (6d), (6e), (6g), (6h), (6i)\}$ be the feasible set for scenario $s \in \mathcal{S}$. Having solved subproblems $\eta(s) = \{\min \sum_{i\in I} \sum_{j\in J} \tilde{c}_{ij} x_{ij}^s + \sum_{i\in I} \sum_{j\in J} c_{ij} y_{ij}^s + \sum_{i\in I} \alpha_i z_i^s + \sum_{j\in J} \sum_{r\in R} \beta_r u_{jr}^s \big| (x_{ij}^s, y_{ij}^s, z_i^s, u_{jr}^s) \in X(s)\}$ for all $s \in \mathcal{S}$, the suggested overall assignment yields

$$\bar{x}_{ij}^{\nu} = \sum_{s\in\mathcal{S}} p_s \, x_{ij}^{s\nu}, \qquad \forall\, (i,j) \in (I \times J), \tag{3.11}$$

which is useful to recognize the local trends and characteristics among scenario solutions. Here two situations can happen: (1) $\bar{x}_{ij}^{\nu} \in \{0,1\}$, which means that consensus occurs and the overall solutions has been retained; (2) $0 < \bar{x}_{ij}^{\nu} < 1$, which means that the overall assignment is not feasible for the original problem. Provided that case (2) happens, a value of $\bar{x}_{ij}^{\nu}$ that is close to one imply a tendency toward assigning a given patient i to PCMH team j, and vice versa. Because case (1) is rarely occurred, to produce a feasible overall solution in iteration ν of the algorithm, one can pick one solution among $x_{ij}^{1\nu}, x_{ij}^{2\nu}, \dots, x_{ij}^{\mathcal{S}\nu}$. In the current study, we pick a worst-case scenario solution with the maximum objective value, that is, $x_{ij}^{WC,\nu} = \text{argmax}\left\{\eta^{\nu}\left(x_{ij}^{s\nu}\right): s \in \{1, 2, \dots \mathcal{S}\}\right\}$. This

solution is feasible for problem (**EF**) and $\sum_{i\in I}\sum_{j\in J}\tilde{c}_{ij}x_{ij}^{WC,\nu}+\sum_{s\in\mathcal{S}}p_s\,Q\left(x^{WC,\nu},d(s)\right)$ provides an upper bound for its optimal value. Although $x_{ij}^{WC,\nu}$ may bias the search process, we calculate it iteratively to keep a best upper bound.

In order to gradually obtain consensus among scenario solutions, the Lagrangian multiplier λ and the penalty parameter ρ are iteratively updated (similar ideas are suggested in Rockafellar and Wets (1991)). This way, we dynamically adjust for the differences between the scenario solutions and the overall solution generated, thus the scenario solutions are forced to converge to a reference solution. If $\lambda_{ij}^{s\nu}$ defines the Lagrangian multiplier associated with the nonanticipativity constraint for assignment of patient i to PCMH team j for scenario s at iteration ν, and ρ^{ν} denotes the quadratic penalty at iteration ν, we then update the parameters as follows

$$\lambda_{ij}^{s\nu}\leftarrow\lambda_{ij}^{s\nu-1}+\rho^{\nu-1}\left(x_{ij}^{s\nu}-\bar{x}_{ij}^{\nu-1}\right),\qquad\forall\,(i,j)\in(I\times J) \tag{3.12}$$

$$\rho^{\nu}\leftarrow\partial\rho^{\nu-1}\,. \tag{3.13}$$

Here we set the initial value ρ^0 to a small positive real number and $\partial>1$, which requires a gradual increase in the penalty parameter. Following these updates, variants of the scenario subproblems are solved which are augmented with a linear term in $\boldsymbol{x}$ proportional to $\boldsymbol{\lambda}^{s\nu}$ and a quadratic proximal term penalizing diversion of $\boldsymbol{x}^{s\nu}$ from $\overline{\boldsymbol{x}}^{\nu-1}$. The algorithm proceeds until the following both conditions are met: (1) the differences between scenario solutions and overall assignment get sufficiently small, that is,

$\sum_{s\in\mathcal{S}}\left(x_{ij}^{s}-\bar{x}_{ij}\right)^{2}\leq\varepsilon$; and (2) there are 10 consecutive nonimproving iterations. The statement of the entire procedure is given in Algorithm 1.

Remark 1. The idea behind the adjustment scheme (3.12) is intuitive. For any patient assignment $x_{ij}^{s\nu}$ in a scenario subproblem s at iteration ν, two cases might occur: (1) $x_{ij}^{s\nu}<\bar{x}_{ij}^{\nu-1}$ which corresponds to the case that patient i is not assigned to PCMH team j in this scenario (or $x_{ij}^{s\nu}=0$) but it is assigned to it in the overall assignment (remember that $0<\bar{x}_{ij}^{\nu-1}<1$). Then the idea is to decrease its cost in the scenario subproblem in order to encourage assigning patient i to team j. Also the modification is more powerful when $\bar{x}_{ij}^{\nu-1}$ is near one. (2) $x_{ij}^{s\nu}>\bar{x}_{ij}^{\nu-1}$ which means patient i is assigned to PCMH team j in the scenario (or $x_{ij}^{s\nu}=1$) but not all other scenarios agree upon this assignment. Then first stage cost $\left(\tilde{c}_{ij}+\lambda_{ij}^{s}-\rho\bar{x}_{ij}+\frac{\rho}{2}\right)$ is adjusted within the scenario to trigger not assigning patient i to PCMH team j. Again the adjustment is stronger when $\bar{x}_{ij}^{\nu-1}$ is near zero.

Lower Bounds

We present the following proposition showing that implicit lower bounds on the optimal cost can be obtained in any iteration of the proposed algorithm for the two-stage stochastic mixed 0-1 problem TSSPA2. Let z^{*} represents the optimal objective value of (**P1**) and suppose that (**P1**) is feasible with $-\infty<z^{*}<+\infty$, and $X(s)\neq\emptyset,\forall s\in\mathcal{S}$.

Proposition 1. *The dual price system $\lambda^{s},s\in\mathcal{S}$ define implicit lower bounds on z^{*}.*

Proof. Let $\lambda^s \in \mathbb{R}^n$ meet $\sum_{s\in\mathcal{S}} p_s\lambda^s = 0$ $(component\text{-}wise)$. Define

$$\mathcal{D}_s(\lambda^s) \coloneqq \min_{\left(x_{ij}^s, y_{ij}^s, z_i^s, u_{jr}^s\right)\in X(s)} \sum_{i\in I}\sum_{j\in J}\left(\tilde{c}_{ij} + \lambda_{ij}^s\right)x_{ij}^s + \sum_{i\in I}\sum_{j\in J} c_{ij}y_{ij}^s + \sum_{i\in I}\alpha_i z_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r u_{jr}^s \tag{3.14}$$

Then we have to prove $\mathcal{D}_s \coloneqq \sum_{s\in\mathcal{S}} p_s\mathcal{D}_s(\lambda^s) \le z^*$. Let $\left\{\left(\hat{x}_{ij}^s, \hat{y}_{ij}^s, \hat{z}_i^s, \hat{u}_{jr}^s\right), s \in \mathcal{S}\right\}$ be the optimal solution to (**P1**). Following the feasibility conditions, we have

$$\mathcal{D}_s(\lambda^s) \le \sum_{i\in I}\sum_{j\in J}\left(\tilde{c}_{ij} + \lambda_{ij}^s\right)\hat{x}_{ij}^s + \sum_{i\in I}\sum_{j\in J} c_{ij}\hat{y}_{ij}^s + \sum_{i\in I}\alpha_i\hat{z}_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r\hat{u}_{jr}^s$$

Then it follows

$$\mathcal{D}_s \le \sum_{s\in\mathcal{S}} p_s\left(\sum_{i\in I}\sum_{j\in J}\left(\tilde{c}_{ij} + \lambda_{ij}^s\right)\hat{x}_{ij}^s + \sum_{i\in I}\sum_{j\in J} c_{ij}\hat{y}_{ij}^s + \sum_{i\in I}\alpha_i\hat{z}_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r\hat{u}_{jr}^s\right)$$

$$= \sum_{s\in\mathcal{S}} p_s\left(\sum_{i\in I}\sum_{j\in J}\tilde{c}_{ij}\hat{x}_{ij}^s + \sum_{i\in I}\sum_{j\in J} c_{ij}\hat{y}_{ij}^s + \sum_{i\in I}\alpha_i\hat{z}_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r\hat{u}_{jr}^s\right) + \sum_{i\in I}\sum_{j\in J}\sum_{s\in\mathcal{S}} p_s\,\lambda_{ij}^s\,\hat{x}_{ij}^s$$

$$= \sum_{s\in\mathcal{S}} p_s\left(\sum_{i\in I}\sum_{j\in J}\tilde{c}_{ij}\hat{x}_{ij}^s + \sum_{i\in I}\sum_{j\in J} c_{ij}\hat{y}_{ij}^s + \sum_{i\in I}\alpha_i\hat{z}_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r\hat{u}_{jr}^s\right) = z^*.$$

It is worth noting that argument $\sum_{s\in\mathcal{S}} p_s\lambda^s = 0$ is maintained in every iteration ν of the algorithm. To see this, we check for $\nu = 1$ and extend the results by induction in

every ν: for $\nu = 1$, we have $\lambda^{s,1} = \rho\left(x_{ij}^{s,1} - \bar{x}_{ij}^{1}\right) = \rho\left(x_{ij}^{s,1} - \sum_{s\in\mathcal{S}} p_s\, x_{ij}^{s,1}\right)$, thus the sum $\sum_{s\in\mathcal{S}} p_s\lambda^{s,1}$ is equal to $\rho\sum_{s\in\mathcal{S}} p_s\left(x_{ij}^{s,1} - \bar{x}_{ij}^{1}\right)$ which is zero since $\sum_{s\in\mathcal{S}} p_s\left(x_{ij}^{s,1} - \bar{x}_{ij}^{1}\right)$ is always zero. So $\sum_{s\in\mathcal{S}} p_s\lambda^{s\nu} = 0$ for all ν. Furthermore, we can see that dual subproblems $\mathcal{D}_s(\lambda^s)$ are roughly identical in structure to those solved by the algorithm, except that quadratic penalty terms $\left(-\rho\bar{x}_{ij} + \frac{\rho}{2}\right)$ are absent. This observation is very helpful in efficiently obtaining lower bounds within the proposed algorithm. Also $\sum_{s\in\mathcal{S}} p_s\lambda^s = 0$ can be taken as "dual" feasibility constraints for the "primal" non-anticipativity constraints $x^s = \bar{x}$, since their subspaces are orthogonal to each other— a primal-dual optimality condition in the convex case.

To assess the quality of the lower bounds obtained, we consider the standard Lagrangian method for (**P1**) and relax the non-anticipativity constraints (3.7d) using multipliers γ^s. We let $\mathcal{W}$ as the feasible set defined by constraints set (**P1**) except (3.7d), and for $w = \left(\bar{x}_{ij}, \left(x_{ij}^s, y_{ij}^s, z_i^s, u_{jr}^s\right)_{s\in\mathcal{S}}\right) \in \mathcal{W}$ we define objective

$$L(w,\gamma) :=$$
$$\sum_{s\in\mathcal{S}} p_s \left[\sum_{i\in I}\sum_{j\in J} \tilde{c}_{ij}x_{ij}^s + \left(\sum_{i\in I}\sum_{j\in J} c_{ij}y_{ij}^s + \sum_{i\in I}\alpha_i z_i^s + \sum_{j\in J}\sum_{r\in R}\beta_r u_{jr}^s\right) + \right.$$
$$\left.\sum_{i\in I}\sum_{j\in J}\gamma_{ij}^s\, x_{ij}^s - \sum_{i\in I}\sum_{j\in J}\gamma_{ij}^s\,\bar{x}_{ij}\right].$$

Then the relaxation problem can be expressed as $F(\gamma) = \min_{w\in\mathcal{W}} L(w,\gamma)$, and the Lagrangian dual problem is given by

$$z_{LD} := \sup_{\gamma} F(\gamma). \tag{3.15}$$

Further, based on Theorem 1 [66], the value of z_{LD} in the mixed 0-1 problem equals to the optimal objective value of the following *linear* program.

Theorem 1.

$$z_{LD} =$$

$$\min \Big\{ \sum_{s\in\mathcal{S}} p_s \Big[\sum_{i\in I} \sum_{j\in J} \tilde{c}_{ij} x_{ij}^s + \Big(\sum_{i\in I} \sum_{j\in J} c_{ij} y_{ij}^s + \sum_{i\in I} \alpha_i z_i^s + \sum_{j\in J} \sum_{r\in R} \beta_r u_{jr}^s \Big) \Big] \Big| \Big\{ \Big(x_{ij}^s, y_{ij}^s, z_i^s, u_{jr}^s \Big) \in \mathbb{C}\big(X(s)\big), p_s x_{ij}^s - p_s \bar{x}_{ij}, \forall s \in \mathcal{S} \Big\} \Big\} \qquad (3.16)$$

in which the closure of the convex hull of s, $\mathbb{C}\big(\mathrm{X(s)}\big)$, is a closed polyhedral set. In most practical cases the requirements by which $\mathbb{C}\big(X(s)\big)$ is a closed polyhedral set are met. Examples include situations when the set defined by the linear constraints is bounded or the cost coefficients are rationals.

Since it is a stochastic mixed 0-1 problem, a duality gap is usually existed between (**P1**) and problem (3.16). This gap can be closed by a branch-and-bound approach where bounding is obtained by solving either the dual problem (3.15) as in [67] or the primal problem (3.16) as in the work by [72]. In the following proposition we show that our proposed algorithm can yield both primal and dual optimal solutions to (3.16) and (3.15). In addition it is shown that the lower bound $\mathcal{D}(\lambda)$ from (3.14) is as tight as best bounds from the dual decomposition, z_{LD}.

Proposition 2. *Assume that Algorithm 1 is used to the primal problem (3.16). In each iteration ν, and for each scenario s, a scenario subproblem of the form*

$$\min_{\left(x_{ij}^{s}, y_{ij}^{s}, z_{i}^{s}, u_{jr}^{s}\right)\in\mathbb{C}(X(s))} \left\{ \sum_{i\in I}\sum_{j\in J} \tilde{c}_{ij} x_{ij}^{s} + \left(\sum_{i\in I}\sum_{j\in J} c_{ij} y_{ij}^{s} + \sum_{i\in I} \alpha_{i} z_{i}^{s} + \sum_{j\in J}\sum_{r\in R} \beta_{r} u_{jr}^{s} \right) \right.$$

$$\left. + \sum_{i\in I}\sum_{j\in J} \lambda_{ij}^{s,\nu}\, x_{ij}^{s} + \frac{\rho}{2} \sum_{i\in I}\sum_{j\in J} \left(x_{ij}^{s} - \bar{x}_{ij}^{\nu}\right)^{2} \right\}$$

is solved. Then in the limit, a solution set $(\bar{x}^{*}, \lambda^{s*})$ is obtained where $\bar{x}^{*}$ solves the primal problem (16), and $\lambda^{s*}, \forall s \in \mathcal{S}$ solves the dual problem (15). In addition, in the limit, the lower bound $\mathcal{D}(\lambda)$ from (14) is equal to z_{LD}.

Proof. Since $\mathbb{C}\big(X(s)\big), \forall s \in \mathcal{S}$ define closed convex polyhedral sets, the problem (3.16) is a linear program. Then the proof is given by Theorem 5.2 in [67].

Therefore, we can interpret Algorithm 1 as a primal-dual algorithm where primal solutions $\left\{\bar{x}_{ij}^{\nu}\right\}_{\nu=1}^{\infty}$ and dual solutions $\left\{\lambda_{ij}^{s,\nu}\right\}_{\nu=1}^{\infty}, \forall s \in \mathcal{S}$ are generated during the running time. Moreover, the above sequences converge to a saddle point of the standard Lagrangian.

Chapter IV: Application

4.1. Data Source and Study Variables

According to National Center for Veterans Analysis and Statistics (NCVAS), VA operates the largest health care system in the USA with 23 geographically different regions (known as VISNs, or Veterans Integrated Service Networks) separated hierarchically within each VISN by level of care or type into different facilities such as VA medical centers (VAMC), Community Based Outpatient Clinic (CBOC), Vet Center (VC), and so forth. Within each facility, every VA primary care enrollee was assigned to an independent physician or non-physician PCP by a standard process-VA Primary Care Management Module. To ensure sufficient staffing and quality of care, each PCP was appointed a target panel size, taking into account the intensity of primary care visits and availability of resources such as supporting staff and capital.

In this study we collected outpatient data from a random sample of 888 different facilities (which corresponds to 130 VAMCs of all 23 VISNs) during FY11 quarter 3 to FY12 quarter 2. The period of one year is appropriate; according to the VA program professionals, the primary care population at each practice site is not subject to drastic change from one year to the next. The Decision Support System (DSS) and National Patient Care Database (NPCD) files of the VA Corporate Data Warehouse (CDW) were employed to extract demographic, socioeconomic, and other types of variables. In addition, due to its rigorous data validity and availability, we chose DRG (Diagnosis

Related Group, 29th version) and its ACC (Aggregated Condition Category) codes for patient case-mix and risk adjustment measures in our predictive analytics [75].

Initially there were 82,000 randomly selected patients with 48 independent attributes coded. All patient visits to primary care and women's health are assembled for a total capture period of one year. Visits from other primary care related clinics, such as Internal Medicine or Geriatric Primary Care, are excluded from the analysis. The two dependent variables are total primary care (PC) and non-primary care (Non-PC) Relative Value Units (or RVUs), and for each unique SSN, they are calculated by converting the primary care and non-primary care Current Procedural Terminology (or CPT) codes from all patient visits during the fiscal year (according to the Centers for Medicare and Medicaid Services model). Simply, the Non-PCRVU refers to all of the non-primary care workload during the year, which could be from one or many visits to outpatient specialty care, and the PCRVU is the primary care workload during the year from outpatient primary care. One advantage of using RVUs in our approach, as opposed to simple face-to-face visit counts, lies in its ability to further accommodate workloads generated by telephone encounters at the VHA. It is noted that the RVU can be seen as a comparable measure of value for care services used in the US Medicare reimbursement and is determined by assigning weight to factors such as personnel time, level of skill, and sophistication of equipment required to render patient services. The predictor variables include baseline demographic and socioeconomic attributes along with some medical factors such as whether the patient has insurance, to which VA facility the patient has been admitted, and so on. Before presenting descriptive of the independent variables we perform some

data-preprocessing activities to prevent unexpected errors during model fitting phase. These include: 1) discarding and imputing (by unconditional mode imputation) missing values of such features as 'VISN' and 'CAN Score' (will be introduced shortly), 2) removing outliers from such variables as 'Age' and 'Assigned provider experience' thus focusing on the first through ninety-ninth percentiles, and 3) binning multimodal, highly skewed features such as 'Distance' and 'Length of stay' into discrete factors. Following this preprocessing, the number of records was reduced to 81,190 patients.

To achieve a better picture of the data environment, we tentatively arranged all independent attributes into five groups as summarized in Table 1. It should be noted that these variables remain the same for a patient during the fiscal year. Note that SD stands for standard deviation and % denotes the percentages of the subgroup in the population. 'Priority' levels range from 1-8 and are assigned based on the veteran's severity of service-connected disabilities and VA income means test (VHA Handbook 1601A.03). 'Distance' is calculated in miles between patient's home zip code and the zip code of the facility he/she admitted, considering the latitude and longitude of the two locations. Records with a calculated distance greater than 240 miles were excluded and the remaining were converted into three levels. 'Changed provider count' denotes the number of times during the year that the patient changed his/her assigned provider. As mentioned earlier this variable could be a marker of unbalanced workload among PCPs and discontinuity of care received by patients. 'Length of stay' (LOS) displays the number of days spent admitted at a VA hospital. 'CAN Score' is the care assessment need score, which reflects the likelihood of admission or death within a specified time period. This

score is commonly expressed as a percentile ranging from 0 (lowest risk) to 99 (highest risk) and it indicates how a VA patient is compared with other patients in terms of the likelihood of hospitalization or death. Each PCMH team has a unique 10-digit code throughout all VA medical systems nationwide. Currently all teams have the same number of professions within all VA centers. The number of PCMH teams and VA facilities in our data set are 6,051 and 287 respectively. ACC categories are determined based on the various ICD-9-CM (International Classification of Disease, ninth version, Clinical Modification) codes assigned to the patient at each visits during the whole fiscal year. They basically indicate the occurrence of a specific disease group, and they are not mutually exclusive categories, meaning that a patient may have more than one ACC during the fiscal year and most actually do.

As shown in Table 1, the mean age of patients is 62.42 years (SD = 15.26) and about half of the cohort was over age 63 (median = 63). Not surprisingly, near 94% of our veteran population was male and approximately 61% of all were insured. Over half of the patients were married but lower than one third of all were reported as actively employed. The most frequently enrolled patients are the low income and Medicaid group followed by >50% connected disability, and non-service connected patients with income above HUD (Housing and Urban Development). The majority of patients (93%) did not spend a day as an inpatient admitted to the hospital, and most of them travelled only a short distance to receive care from the VA hospitals. The mean care assessment score is roughly 47 with a great variation (SD = 28.88). Also, on average, most of patient's assigned providers are well-experienced working rather full time in their roles.

Table 1 Baseline characteristics of patient factors (*n* = 81190)

Group	Attribute	Mean (SD)	*n* (%)
Demographic	Gender		
	Male		76247 (93.91)
	Female		4943 (6.09)
	Age (as of 7/1/2011, years)	62.42 (15.23)	
	Marital status		
	Married		46634 (57.44)
	Previously married		22520 (27.74)
	Never married		11559 (14.24)
	Unknown		477 (0.58)
Socioeconomic	Insurance (of any types)		
	Yes		49551 (61.03)
	No		31639 (38.97)
	Employment status		
	Active Military Service		134 (0.16)
	Employed Full-Time		17008 (20.95)
	Employed Part-Time		4013 (4.94)
	Not Employed		28619 (35.25)
	Retired		28517 (35.12)
	Self Employed		2039 (2.51)
	Unknown		860 (1.07)
Enrollment	Priority		
	1 (service connected disability > 50%)		18404 (22.67)
	2 (service connected disability 30%–40%)		6548 (8.07)
	3 (service connected disability 20–30%)		9859 (12.14)
	4 (catastrophically disabled)		2285 (2.82)
	5 (low income or Medicaid)		21258 (26.18)
	6 (Agent Orange or Gulf War illness)		3697 (4.55)
	7 (non-service connected, income below HUD)		2243 (2.76)
	8 (non-service connected, income above HUD)		16896 (20.81)

Next, we provide two schematic views of the mean annual care demand and disease *prevalence* of multiple patient groups. In Fig. 4, the average RVU demands of the primary and non-primary care generated are displayed across different priority groups with insurance status nested. Not unexpectedly, the non-primary care effort is always more than the primary care workload and its ratio changes from 1.8 in group 8-insured to 6.6 in group 4-uninsured. In all priority groups, uninsured VA patients compared to

insured ones produce, on average, more workload in terms of both primary and non-primary care. In addition, the biggest (lowest) workload demands for both primary and specialty care services are associated with group 8-uninsured patients (group 6-insured patients).

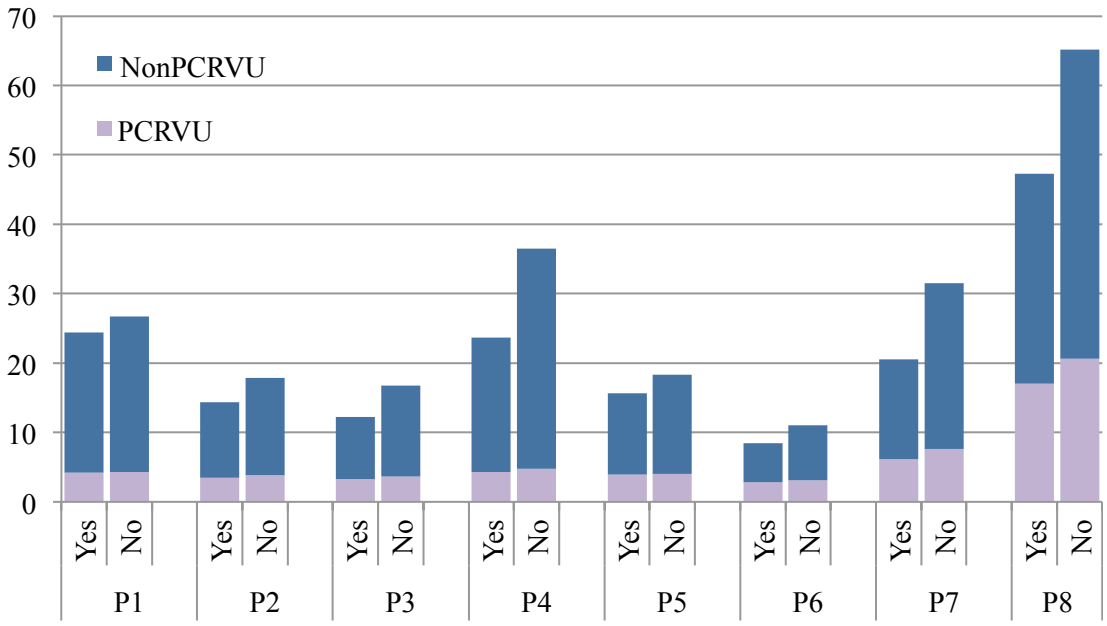


Figure 4 Average annual primary care and non-primary care relative

Fig. 5 displays a mosaic plot of illness types along with patients' gender and their marital status. We excluded ACC 28 (neonate's diseases) and 'unknown' marital category from these analyses because of either the absence or rarity in our sample study. Note that letters P, N, and M above the marital bar denote 'Previously married', 'Never married', and 'Married' groups. The ACC labels are given in Table 6. As shown, the most commonly occurring conditions among all patient clusters is ACC 30 (Screening)

followed by ACCs 5 (nutritional and metabolic) and 16 (heart). However, the least prevalent illnesses among the VA patients are ACC24 (pregnancy-related), ACC13 (developmental disability), and ACC15 (cardio-respiratory arrest). Plus, in almost all disease types, married males are more at risk than two other male groups.

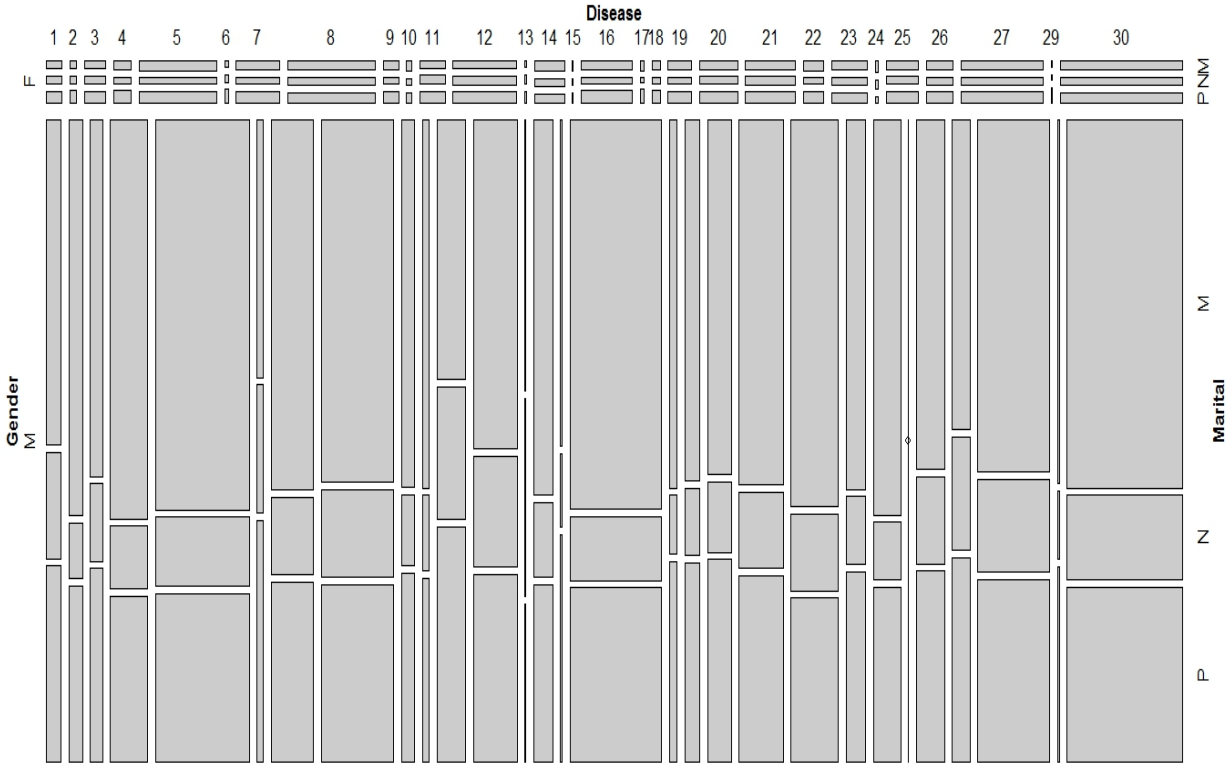


Figure 5 Mosaic plot of disease prevalence across patient gender and marital

4.2. Analytics

4.2.1. Model Fitting and Diagnostics

We conduct multiple analyses to estimate the effect of different patient factors such as disease types (ACCs) on the mean annual primary and non-primary care. To employ our method we first determine the appropriate distributions for the two responses. Here

the standard Quantile-Quantile plot along with maximum likelihood method is used, but one can also employ non-parametric techniques such as kernel density estimation. We examine different base densities such as gamma, lognormal, beta, and Cauchy, then judge the best choice as having the best graphical pattern in QQ plot and the biggest likelihood value simultaneously. Based on these criteria, the lognormal distribution is found the most proper case for both RVUs. Fig. 6 shows the QQ plots along with bootstrapped point-wise confidence envelopes at 0.95 accuracy rate. As shown, the PCRVU (left panel) displays a perfect linear pattern, and even for Non-PCRVU (right panel), almost all points lie within the confidence band. We also get the minimum value of the minus log-likelihood based on ML fitting when the lognormal distribution is taken.

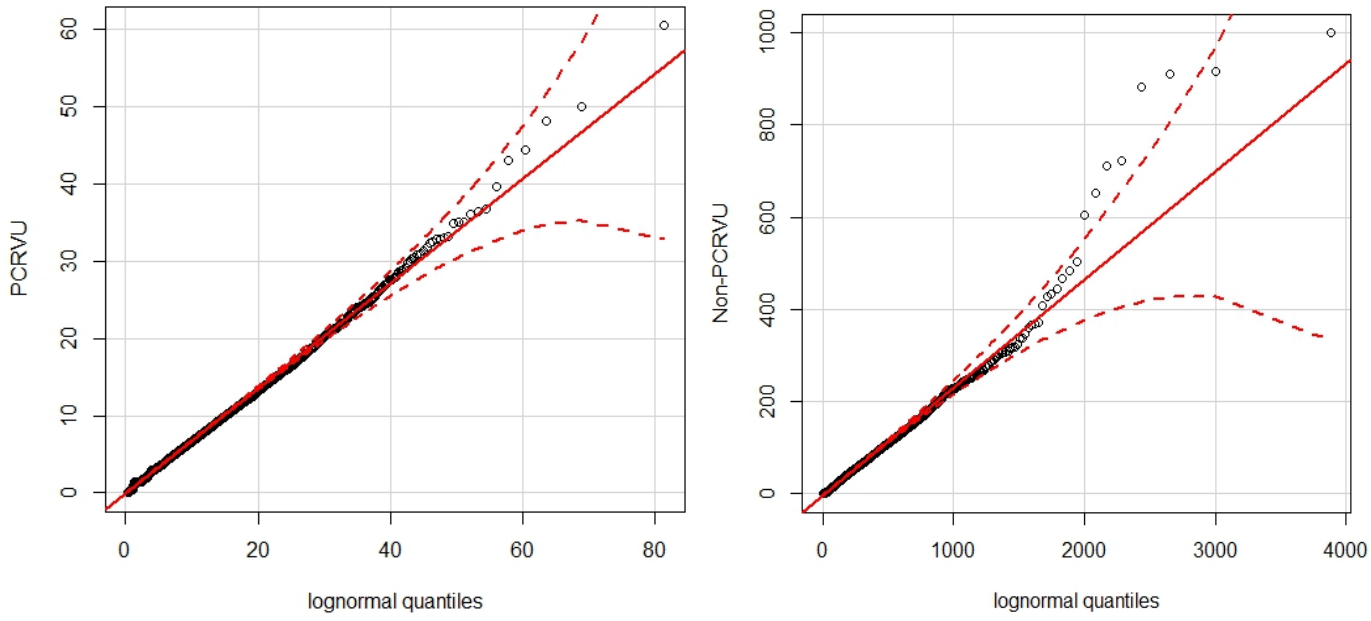


Figure 6 Quantile-Quantile plots of primary care relative value unit

To determine the appropriate link function $g(\cdot)$, a range of classical options including log link and inverse link are evaluated by two goodness-of-fit measures, namely DIC and modified Hosmer-Lemeshow test [76]. Based on the results (not shown here) we observe that the (default) identity link does estimate the upper and lower tails of both RVUs, accounted for the covariates, more properly than other links, and thus it is chosen for our study.

Since failing to specify the suitable probability density for priors can result in inferential and numerical problems as discussed in chapter2, for the deterministic parameters we pick a multivariate normal density with zero vector for the mean $\boldsymbol{\gamma}_0$ and a diagonal matrix of large variances (1e+10) for $\boldsymbol{\Gamma}$. This way we can make sure that the prior is always *proper*. However, for each (decomposed) *block* of the G-(R-) side, we are required to specify the hyperparameters through the IW distribution, which takes two scalars; the expected (co)variance at the limit and the degree of belief parameter. We configure several prior specifications not only for these two parameters but also for different shapes (degrees of freedom) the decomposed matrices can take, then assess the impacts on the DIC measure and their posterior distributions (with MCMC diagnostics). A few such comparisons are discussed in chapter 2, but now for the first step of our modeling strategy (discussed later), we choose a diagonal matrix of 1/3 for all three hierarchies (patients, PCMH teams, and facilities) with 2 degrees of freedom. Scaling outcomes to have a unit variance before the analysis, this prior implies that the total variance is equally split across all three levels together with *a priori* independence of PC and Non-PC workloads.

Although different modeling strategies could be selected for estimating our multilevel model, we focus on the most parsimonious and best-fitting approach for the given data and our specific research questions. To this end, six models (Table 2) from basic to comprehensive are run sequentially and the outputs are reported for each step in order to provide insights for a particular objective. Further, to avoid overfitting within each step, we perform stepwise selection for the deterministic covariates with probabilities to enter and stay of 0.15 and 0.1 respectively.

Table 2 Regression modeling strategy and results for 3-level hierarchical model

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
No predictors, just residual and random intercepts (Unconditional)	Model 1 + patient-level predictors	Model 2 + random slopes for patient-level predictors	Model 3 + team-level predictors	Model 4 + random slopes for team-level predictors	Model 5 + facility-level predictors
Results used to compute Interclass Correlation Coefficient (ICC) which assesses the degree of clustering among subsets of cases in the data.	Results show the relationships between patient-level predictors and outcomes	Model 2 results + findings that show if the associations between patient-level predictors and the outcomes vary across team-level and facility- level units	Model 3 results + results that reveal the relationships between team-level predictors and the outcomes	Model 4 results + findings that shows if the associations between team-level predictors and the outcomes vary across facility-level units	Model 5 results + results that indicate the relationships between team-level predictors and the outcomes.

Alternatively, one can employ a Bayesian selection to determine a variable subset. Different functional forms of covariates, such as logarithmic and power relations, as well

as within-level interactions are evaluated too at each step but only the statistically significant ones are included. As an example, we analyze 12 pairs of ACC interactions that are notable for co-occurring in patients with multiple chronic illnesses and/or an acute disease combined with a chronic condition [77].

The improvement in model fit is evaluated by DIC over all iterations after the burn-in phase of MCMC simulations. Based on a rule of thumb, we favor the model with lower DIC when the DIC reduction of more than 10 units is observed. Depending on the goodness-of-fit and significance tests, sometimes intermediate models, such as a reduced version of model 3 with only one significant random slope, are also examined. Performing this strategy, we seek to answer the following three research questions:

- How much of the variance in PC and Non-PC workload is associated with patients, PCMH teams, and VA facilities?
- Does the effect of any patient-level predictor change among PCMH teams or VA facilities? And does the effect of any team-level predictor vary among VA facilities?
- What is the impact of patient *non-adherence* (as measured by "Changed provider count") on PC workload, controlling for patient, PCMH team, and VA facility characteristics?

Setting the significance level at 0.05, we run the models with 50,000 iterations, a burn-in period of 10,000, and a thinning interval of 25. All analyses and computations are done in R version 3.0.2 [78]. In order to address the first question, we fit the unconditional model as summarized in Tables 3-5. Note that the first (third) row in each

table shows PC (Non-PC) intercept variance along with its 95% Highest Posterior Density interval, and the second row corresponds to the workload correlations. The team interclass correlation coefficient for the PC outcome is computed as $\frac{0.168}{0.609 + 0.168 + 0.218}$. Note that the numerator is the PCRVU variance at the team level obtained from Table 4, and the denominator is the sum of PCRVU variances in all levels obtained from Tables 4-6. Simply put, we find that about 17% of the variation in PC workload exists between PCMH teams and 22% is there between VA facilities, leaving near 61% of the variance to be accounted for by patients. Thus a practically meaningful proportion of all variation happens at higher levels, providing support for our use of a 3-level hierarchical model. These percentages are 5%, 16%, and 79% for Non-PC workload respectively. Other useful points can be made by interpreting the correlations among PC and Non-PC at different levels. First, the results of a joint conditional independence test Gueorguieva [79] show that the RVUs (at the patient level) are positively associated which confirms the fact that a simultaneous modeling of both primary and non-primary care is more reasonable than using one of them in isolation. Second, we infer that the correlation is not significant when it comes to the team level, and it is poorly significant at the facility level.

We continue our modeling effort to include predictors and random components at all levels, and then answer other research questions based on the outputs from the best fitting model. For brevity we will not walk through all detailed outputs at each stage, and instead summarize them in Table 3,4,5. Also note that level-2 and level-3 predictors are

displayed italic. In each row, the first number is for PC and the second is for Non-PC outcome, with (′), (″), () displaying significance at 0.05, <0.001, and non-significance respectively. It worth noting that we suppress the overall intercept since otherwise, the parameter estimates associated with PC are translated as contrasts with Non-PC. Also for team-level, facility-level, and interactions, we only include those factors that are significant in at least one of the six models.

Graphically assessing the relation of age with the outcomes, we observe that both responses have a sigmoidal trend at team levels thus we decide to fit its nested random components with covariance matrix like

$$\mathbf{P}_{\mathbf{Team}} \otimes \mathbf{S}_{\mathbf{Team}} = \begin{bmatrix} \sigma^2_{\mathbf{(Intercept)}} & 0 & 0 \\ 0 & \sigma^2_{\mathbf{Age}} & 0 \\ 0 & 0 & \sigma^2_{\mathbf{Age}^2} \end{bmatrix} \otimes \mathbf{I}$$

Table 3 Coefficient estimates model for joint PC and Non-PC workloads

	Model 1	Model 2	Model 3	Model 4
Deterministic Effect				
Gender, Male		0.41″, 0.02″	0.43″, 0.03″	0.42″, 0.01″
Age		1.02″, 1.04″	1.03, 1.03	1.03, 1.03
Age × Age		0.92″, 0.94″	0.9′, 0.91′	0.91′, 0.93′
Insurance, Yes		0.95′, 0.92	0.94′, 0.9	0.95′, 0.91
LOS, Zero		1.07, 0.74″	1.06, 0.71″	1.08, 0.73″
CAN Score		1.12″, 1.07″	1.08, 1.02	1.09, 1.03
SQRT (CAN Score)		1.15′, 1.19″	1.1′, 1.12′	1.12′, 1.13′
Priority (ref = 8)				
1 (disability > 50%)		0.96″, 1.25″	0.97″, 1.22″	0.96″, 1.23″
2 (disability 30%–40%)		1.02′, 1.32″	1.02′, 1.28′	1.03′, 1.29′
3 (disability 20–30%)		0.94′, 1.01″	0.92′, 1.04″	0.92′, 1.03″
4 (catastrophically dis.)		1.03″, 1.17″	1.04″, 1.14″	1.03′, 1.15″
5 (Medicaid)		1.05″, 1.03″	1.04″, 1.05″	1.05″, 1.04″
6 (Agent Orange, …)		1.06″, 1.34′	1.03″, 1.32″	1.03″, 1.33″
7 (below HUD)		1.09″, 1.1″	1.08″, 1.07″	1.09″, 1.07″
ACC001–Infectious and Parasitic		1.07″, 1.22″	1.05″, 1.23″	1.04″, 1.24″
ACC002–Malignant Neoplasm		1.04″, 1.33″	1.04″, 1.3″	1.03″, 1.31″
ACC003–Benign/In Situ/Uncertain Neoplasm		1.07″, 1.65″	1.06″, 1.65″	1.06″, 1.64″
ACC004–Diabetes		1.53″, 0.98′	1.52″, 0.97′	1.53″, 0.96′
ACC005–Nutritional and Metabolic		1.18″, 1.02	1.19″, 1.03	1.2″, 1.02
ACC006–Liver		1.13″, 1.04′	1.11″, 1.05′	1.12″, 1.05′
ACC007–Gastrointestinal		1.09″, 1.13″	1.07″, 1.14″	1.07″, 1.14″
ACC008–Musculoskeletal and Connective Tissue		1.18″, 1.27″	1.17″, 1.27″	1.16″, 1.28″
ACC009–Hematological		1.09″, 1.05″	1.08″, 1.06″	1.07″, 1.06″
ACC010–Cognitive Disorders		1, 1.12″	0.98, 1.1″	1, 1.11″
ACC011–Substance Abuse		1.06″, 0.88″	1.06″, 0.9″	1.05″, 0.9″

Table 4 Coefficient estimates model for joint PC and Non-PC workloads

	Model 1	Model 2	Model 3	Model 4
ACC012–Mental		1.03′, 1.73″	1.04′, 1.7″	1.03′, 1.71″
ACC013–Developmental Disability		0.99, 1.24″	1.01, 1.23″	1.01, 1.22″
ACC014–Neurological		1.07″, 1.15″	1.06″, 1.14″	1.07″, 1.16″
ACC015−Cardio-Respiratory Arrest		1.07′, 1.02	1.03′, 1.04	1.05′, 1.03
ACC016−Heart		1.15″, 1.05′	1.14″, 1.06′	1.16″, 1.04′
ACC017−Cerebrovascular		1.05, 1.02	1.05, 1.03	1.04, 1.01
ACC018−Vascular		1.08″, 1.26″	1.1″, 1.26″	1.09″, 1.27″
ACC019−Lung		1.09″, 1.11″	1.07″, 1.12″	1.08″, 1.12″
ACC020−Eyes		1.08″, 1.12″	1.09″, 1.13″	1.09″, 1.14″
ACC021−Ears, Nose, and Throat		1.11″, 1.40″	1.12″, 1.38″	1.1″, 1.39″
ACC022−Urinary System		1.06″, 1.01	1.07″, 1.02	1.08″, 1.02
ACC023−Genital System		1.09″, 1.07″	1.09″, 1.04″	1.1″, 1.06″
ACC025−Skin and Subcutaneous		1.11″, 1.42″	1.13″, 1.43″	1.12″, 1.43″
ACC026−Injury, Poisoning, Complications		1.1″, 1.28″	1.11″, 1.29″	1.12″, 1.3″
ACC027−Symptoms, Signs, and Ill-Defined Conditions		1.17″, 1.45″	1.15″, 1.41″	1.16″, 1.42″
ACC029−Transplants, Openings, Amputations		0.9″, 1.01	0.94″, 0.98	0.92″, 0.99
ACC030−Screening/History		1.22″, 2.01″	1.23″, 1.98″	1.2″, 1.98″
Changed provider count				1.11″, 1.09″
Distance (ref = Far) Middle Near				
Diabetes × Liver		1.02, 1.13′	1.03, 1.15′	1.03, 1.16′
Diabetes × Cardio-Respiratory Arrest		1.12′, 1.11″	1.1′, 1.13″	1.13′, 1.12″
Diabetes × Heart		1.03, 1.1″	1.04, 1.12″	1.03, 1.11″
Diabetes × Cerebrovascular		1.07′, 1.17′	1.06′, 1.14′	1.06′, 1.15′
Diabetes × Urinary System		1.04, 1.12′	1.06, 1.1′	1.05, 1.1′

Table 5 Coefficient estimates model for joint PC and Non-PC workloads

	Model 1	Model 2	Model 3	Model 4
Diabetes × Transplants, Openings, Amputations		1.08′, 1.09	1.09′, 1.07	1.09′, 1.08
Substance Abuse × Mental		1.04″, 1.20″	1.03″, 1.21″	1.04″, 1.21″
Heart × Cerebrovascular		1.12′, 1.14″	1.09′, 1.13″	1.1′, 1.15″
Heart × Vascular		1.06, 1.04′	1.07, 1.05′	1.05, 1.05′
Cerebrovascular × Vascular		1.01, 1.12″	1.03, 1.13″	1.02, 1.14″
Male × Diabetes		1.06′, 1.12′	1.05′, 1.14′	1.04′, 1.14′
Male × Neurological		1.08″, 1.11′	1.09″, 1.13′	1.1″, 1.12′
Age × Heart		1.11″, 1.21′	1.09″, 1.19′	1.09″, 1.2′
Age × Nutritional and Metabolic		1.14′, 1.07″	1.15′, 1.09″	1.14′, 1.08″
Age × Gastrointestinal		1.05′, 1.1′	1.07′, 1.12′	1.06′, 1.12′
Priority 4 × Neurological		1.13′, 1.17″	1.14′, 1.17″	1.11′, 1.16″
Priority 6 × Cardio-Respiratory Arrest		1.14″, 1.06′	1.14″, 1.07′	1.13″, 1.07′
Variance Component				
Residual	0.609′, 0.79′	0.446′, 0.55′	0.357′, 0.46′	0.352′, 0.44′
Intercept (team)	0.168′, 0.05′	0.093′, 0.04′	0.076′, 0.04′	0.064′, 0.04′
Intercept (facility)	0.218′, 0.16′	0.125′, 0.1′	0.106′, 0.08′	0.091′, 0.08′
Slope (age: team)			0.088′, 0.09′	0.081′, 0.09′
Slope (age^2: team)			0.042′, 0.06	0.047, 0.07′
Slope (CAN Score: team)			0.078′, 0.09′	0.072′, 0.1′
Slope (CAN Score^(0.5): team)			0.037, 0.05′	0.042′, 0.04′
Slope (insurance: facility)			0.051′, 0.06′	0.047′, 0.07′
Slope (changed provider count: facility)				
Model Fit				
DIC	461019.6	227245.2	225469.7	225411.4

A similar structure is fitted for CAN Score as well, but with square root instead of second power relation. For 'Changed provide count' we first test a structure with both random intercept and slope at the facility-level, but after failing to reject the null hypothesis of intercept, we reduce it to random slope only. For fitting insurance covariance, again we first try $\begin{bmatrix} \sigma^2_{\text{Insured}} & \sigma_{\text{Insured,Un-insured}} \\ \sigma_{\text{Un-insured,Insured}} & \sigma^2_{\text{Un-insured}} \end{bmatrix}$, and then drop the correlation after the significance test.

According to the DIC index shown at the bottom of Table 6, we realize that each forward model exhibits a better fit to the data, so we take model 6 to answer the remaining research questions. In order to further validate the final model, we apply model 6 to FY12 quarter 3 data and find almost identical results. We repeat the joint independence test of Gueorguieva [79] for model 6 and reaffirm the positive correlation of responses at the patient level. Put differently, we find that after controlling for all sources of variation, if the primary care workload is increased from one patient to another, on average we will expect an increase in the related non-primary care. In Table 6, the estimates for deterministic effects are interpreted as prevalence ratios but variance components are reported in natural scale. Also note that the data is scaled to have a unit variance before analysis.

It is worth to highlight that some estimates are changed in terms of significance among models. For example, age, insurance, and CAN Score are significant in Model 2 but no longer significant in later models once their related random slopes are introduced

in Model 3. Examining other random components in these models, we figure out that significant variability exists in their nested random intercepts and slopes, even after controlling for these patient-level predictors. Hence, we can say that the association between these variables and the outcomes varies considerably among PCMH teams. Thus we expect that the influence of patient oldness on care demands may be stronger or weaker from one PCMH team to another within a VA facility. The same thing happens in terms of effect magnitude for 'Changed provider count' between Models 4 and 5; the relationship between this variable and both workloads changes meaningfully among different VA facilities. By these statements, we tackle our second research question.

To answer the last research question, we look at the deterministic effect of 'Changed provider count' in Model 6. As shown, for each time that a patient switches assigned provider, we will expect an average of 6% more workload in his/her primary care, after accounting for variations of his/her non-primary care demands. Other selected key findings from Model 6 can be summarized as below:

- Adjusting for the contributions of all other variables, female VA patients tend for produce about 57% more PC (98% more Non-PC) compared to males. This is not unexpected due to gender imbalance issue existed in VA patients.
- Inpatient cohort generally creates 28% more workload in non-primary care compared to outpatients, after accounting for variations of their primary cares.
- Catastrophically ill veterans (P4) have 1.15 times the Non-PC demands of the P8 comparison group. The increase rates are about 35% and 23% for veterans exposed to Agent Orange (or other herbicides) and >50% for disabled veterans. Having been

exposed to such chemicals also notably affects the increased caress for cardio-respiratory arrest.

- Change rates in primary cares range from 7% decrease for ACC29 (Transplant) to 52% increase for ACC4 (Diabetes). For non-primary cares, this varies from 11% reduction for ACC11 (Substance Abuse) to 99% rise for ACC30 (Screening).
- Both team-level (patient non-adherence) and facility-level (distance) predictors are significantly associated with the outcomes: Patients travelling more miles to VA hospitals are likely to generate a larger amount of care than closely located patients.
- In co-occurring diseases studies, diabetes greatly interacts with some acute and chronic conditions. For instance, in patients with cardio-respiratory arrest, having diabetes is associated with a 13% (14%) increase in primary care (Non-PC) workload. Another comorbid condition that poses a similar pattern is heart disease, especially for cerebrovascular patients.
- Risk adjustment for disease types and their interactions improves the model fit to a great extent (about 160K reduction in DIC) and makes most of their related effects statistically significant.

Now we present some diagnostic tests for verifying the accuracy of Model 6. First, to assess the Markov chain convergence and mixing properties, trace plots and smoothed posterior densities are provided for each parameter of interest. As an illustration, Fig.7 shows the plots for age and gender across both outcomes and Fig.8 displays them for R-side covariance components. As depicted in Fig.5, the traces are trendless and the

chains are mixing well travelling quickly to the target distribution with small autocorrelations.

Nearly same patterns are observed in Fig.8, but chains are now mixing marginally at a bit slower traverse rate, which can easily be tackled by increasing the MCMC iterations. Nonetheless, the densities do smoothly estimate the mean posterior for residual variances as reported in Table 3,4,5. For deterministic terms in Fig.7, however, the posterior histogram is plotted in log scale. We additionally perform Gelman, Rubin [80] and effective sample size tests to all posterior estimates (not shown here) and no violations are found therein.

Figure 7 Trace plot and posterior density estimates

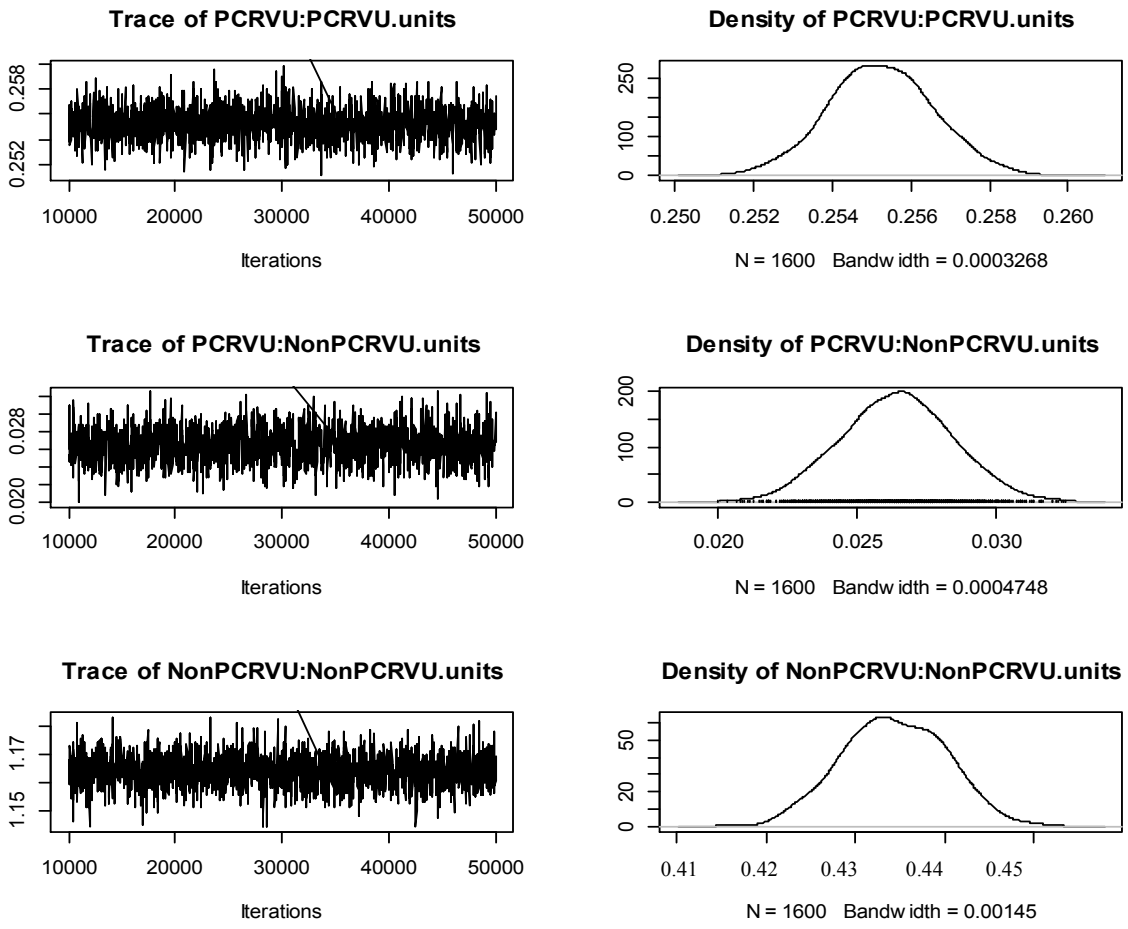
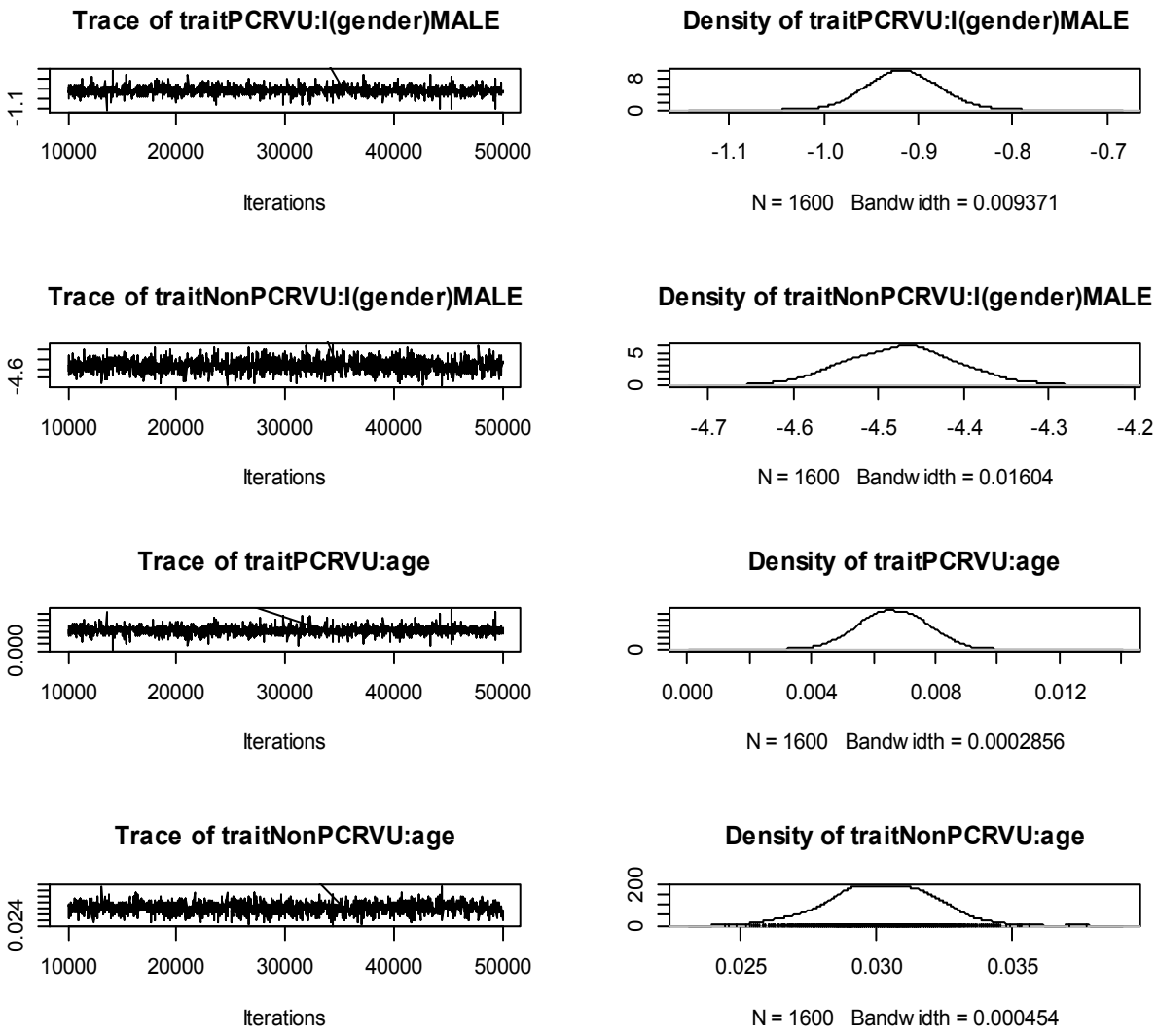


Figure 8 Trace plot and density estimates for (co)variance components of PC



After additional validation steps such as Copas [81] test of overfitting and posterior correlation diagnostics of estimated parameters, we develop two operational indices, namely, hospital level normalized intensity score (NIS) and hospital level risk-standardized utilization rate (RSUR) as

$$NIS_j = \frac{\text{Sum of the total predicted workload of all patients at the given facility}}{\text{Number of patients in the facility} \times \text{Median predicted workload}}$$

and

$$RSUR_j = \frac{\text{Sum of the total predicted workload of all patients with random components}}{\text{Sum of the total predicted workload of all patients without random components}} .$$

The NIS can be used to adjust the panel size up or down for a given hospital, or even for a specific PCMH team within a hospital. Note that the random components are implicitly included in the formula. On the other hand, RSUR indicates the ratio of *predicted* (technically called shrinkage estimate) to *expected* utilization; the numerator computes the PC/Non-PC workload when patients are treated as the specific hospital and denominator calculates the workload as if patients are treated at a so called 'reference' (or normal) hospital. Thus values greater than one reveals that the hospital is over-utilized as compared with the national average range.

4.2 .2 Numerical Comparisons

In this section we design three comparison studies to demonstrate some novel aspects of our proposal. First, we evaluate an alternative variance structure with the one applied in Model 6 in terms of the goodness-of-fit measure. Particularly, for patient (residual), team, and facility random intercepts in scenario (1), we change the Parametric matrix to have the same diagonal elements with zero off-diagonals then compare the results with the structure used in Model 6. We run each model twice to take control of the Monte Carlo error and keep all other factors constant among different fittings. As shown in

Table 6, the best fit is corresponding to the first row in which the proposed variance structure is applied at all levels of hierarchy.

Table 6 Goodness-of-fit values for the two scenarios

Facility	Team	Patient	Deviance information criterion
2	2	2	225337.8 – 227448.1
2	2	1	225491.7 – 225494.1
2	1	2	225401.1 – 225396.9
2	1	1	225582.5 – 225580.3
1	2	2	225378.5 – 225375.7
1	2	1	225444.9 – 225441.2
1	1	2	225457.8 – 225460.5
1	1	1	225550.7 – 225554.0

Second, we investigate the impacts of the random component's prior specification on MCMC diagnostics and posterior distributions. To this end, the DF is kept fixed, and then two alternatives for the expected limit (co)variance (1.one 2.restricted maximum likelihood estimates), as well as other values for the IW degree of belief $\{0.002, 0.02, 0.2, \text{and } 1\}$ are assessed. The values used in Model 6 for these two are 1/3 and zero. Results (available from authors) denote that almost no change occurs in deterministic estimates, DIC measure, and directions of (co)variance components. However, the absolute range of alternations in variance estimates is around 2.3% that the base values in Model 6. We detect that better chain convergence and mixing property is observed when using priors with smaller limit (co)variances and larger (near one) degree of the belief parameter. Further, the posterior correlation estimates remain reasonably unchanged

while examining different types of priors, which provide some reassurance that our priors do not dominate the model to an unacceptable extent.

Lastly, we perform comparisons between our proposal and the situations when one employs a series of univariate (multilevel) GLMs for predicting the outcomes. To this end, we keep Model 6 settings constant and consider two scenarios: 1) A bivariate 3-level GLM with joint primary and non-primary care workloads, and 2) Two univariate 3-level GLMs one for primary care (PC) and one for non-primary care (Non-PC) workload predictions. Fitting both models, we aggregate the credible intervals for the mean outcomes and then compare them with the actual values. Interestingly, the probability of *joint* correct prediction (for both responses) is about 67% for the first scenario and about 58% for the second. Then we pick those correct intervals, compute their ranges $\{\max - \min\}$, and calculate basic statistics for the ranges in Table 7. As displayed, the credible intervals are substantially narrowed when applying the multivariate approach. Thus we can conclude that a joint modeling of primary and non-primary care workloads would provide more robust and realistic predictions for medical home practices.

Table 7 Summary statistics for the range of joint correct intervals

	Multivariate		Univariate	
	Primary Care	Non-Primary Care	Primary Care	Non-Primary Care
Mean	0.431	1.023	0.514	1.083
Median	0.381	0.977	0.439	1.058

4.2. Modeling

We use natural logarithm transformation for the both response variables (primary care relative value unit or 'pcrvu' and non-primary care relative value unit or 'npcrvu') in order to convert them into Gaussian. We distinguish four levels of hierarchy: responses (level-1) are nested in $patient_i$ (level-2), patients are nested in PCMH $team_j$ (level-3), and PCMH teams are nested in VA medical $facility_k$ (level-4). The following four level hierarchical STAR model is suggested:

$$\text{Level}-1: y_{ijk}^{(h')} = d_{ijk}^{(1)}\,\mathbf{ln}(\boldsymbol{pcrvu}) + d_{ijk}^{(2)}\,\mathbf{ln}(\boldsymbol{npcrvu})$$

$$\text{Level}-2: \mathbf{ln}(\boldsymbol{pcrvu}) =$$

$$\mathbf{1}\boldsymbol{\eta}_0^{(1)} + \boldsymbol{f}_1^{(1)}(age) + \boldsymbol{f}_2^{(1)}(los) + \boldsymbol{f}_3^{(1)}(can) + \boldsymbol{f}_4^{(1)}(age)acc1 + \cdots +$$

$$\boldsymbol{f}_{32}^{(1)}(age)acc30 + \boldsymbol{f}_{33}^{(1)}(can, los) + \cdots + \boldsymbol{V}^{(1)}\boldsymbol{\gamma}^{(1)} + \boldsymbol{\varepsilon}^{(1)}$$

$$= \mathbf{1}\boldsymbol{\eta}_0^{(1)} + \boldsymbol{X}_1^{(1)}\boldsymbol{\beta}_1^{(1)} + \cdots + \boldsymbol{V}^{(1)}\boldsymbol{\gamma}^{(1)} + \boldsymbol{\varepsilon}^{(1)}$$

$$\text{Level}-2: \mathbf{ln}(\boldsymbol{npcrvu}) = \mathbf{1}\boldsymbol{\eta}_0^{(2)} + \boldsymbol{X}_1^{(2)}\boldsymbol{\beta}_1^{(2)} + \cdots + \boldsymbol{V}^{(2)}\boldsymbol{\gamma}^{(2)} + \boldsymbol{\varepsilon}^{(2)}$$

$$\text{Level}-3: \boldsymbol{\eta}_0^{(1)}$$

$$= \mathbf{1}\boldsymbol{\eta}_{0,0}^{(1)} + \boldsymbol{f}_{0,1}^{(1)}(prov.exp) + \boldsymbol{f}_{0,2}^{(1)}(prov.fte) + \boldsymbol{f}_{0,3}^{(1)}(prov.chng)$$

$$+ \boldsymbol{f}_{0,4}^{(1)}(prov.pos) + \boldsymbol{f}_{0,5}^{(1)}(prov.exp)prov.pos + \cdots + \boldsymbol{V}_0^{(1)}\boldsymbol{\gamma}_0^{(1)} + \boldsymbol{\varepsilon}_0^{(1)}$$

$$= \mathbf{1}\boldsymbol{\eta}_{0,0}^{(1)} + \boldsymbol{X}_{0,1}^{(1)}\boldsymbol{\beta}_{0,1}^{(1)} + \cdots + \boldsymbol{V}_0^{(1)}\boldsymbol{\gamma}_0^{(1)} + \boldsymbol{\varepsilon}_0^{(1)}$$

$$\text{Level}-3: \boldsymbol{\eta}_0^{(2)} = \mathbf{1}\boldsymbol{\eta}_{0,0}^{(2)} + \boldsymbol{X}_{0,1}^{(2)}\boldsymbol{\beta}_{0,1}^{(2)} + \cdots + \boldsymbol{V}_0^{(2)}\boldsymbol{\gamma}_0^{(2)} + \boldsymbol{\varepsilon}_0^{(2)}$$

$$\text{Level}-3: \boldsymbol{\beta}_1^{(1)} = \boldsymbol{f}_{3,1}^{(1)}(prov.exp) + \boldsymbol{V}_3^{(1)}\boldsymbol{\gamma}_3^{(1)} + \boldsymbol{\varepsilon}_3^{(1)} = \boldsymbol{X}_{3,1}^{(1)}\boldsymbol{\beta}_{3,1}^{(1)} + \boldsymbol{V}_3^{(1)}\boldsymbol{\gamma}_3^{(1)} + \boldsymbol{\varepsilon}_3^{(1)}$$

$$\text{Level}-3: \boldsymbol{\beta}_1^{(2)} = \boldsymbol{f}_{3,1}^{(2)}(prov.exp) + \boldsymbol{V}_3^{(2)}\boldsymbol{\gamma}_3^{(2)} + \boldsymbol{\varepsilon}_3^{(2)} = \boldsymbol{X}_{3,1}^{(2)}\boldsymbol{\beta}_{3,1}^{(2)} + \boldsymbol{V}_3^{(2)}\boldsymbol{\gamma}_3^{(2)} + \boldsymbol{\varepsilon}_3^{(2)}$$

$$\text{Level}-4: \boldsymbol{\eta}_{0,0}^{(1)} = \boldsymbol{V}_{0,0}^{(1)}\boldsymbol{\gamma}_{0,0}^{(1)} + \boldsymbol{\varepsilon}_{0,0}^{(1)}$$

$$\text{Level} - 4: \boldsymbol{\eta}_{0,0}^{(2)} = \boldsymbol{V}_{0,0}^{(2)}\boldsymbol{\gamma}_{0,0}^{(2)} + \boldsymbol{\varepsilon}_{0,0}^{(2)} \quad (4.1)$$

The top level equation contains the two responses. The level-2 equations are STAR models for logged primary and non-primary care workloads that are regressed on possibly nonlinear effects of patient's age, care assessment need score, and length of stay using P-splines. We also include interaction effects between age, CAN score, priority, and all disease types, and between CAN score and length of stay with a two dimensional surface. The categorical covariates on the patient level along with their possible interactions are encoded as dummy variables and subsumed in $\boldsymbol{V}^{(\cdot)}$ with parameters $\boldsymbol{\gamma}^{(\cdot)}$. Note that here we use the same set of effects for the both response regression, but this may change in other applications with a bivariate response. The first and the second level-3 equations model patient-specific offset by the team level covariates such as provider experience and its interaction with provider position plus random intercepts $\boldsymbol{\varepsilon}_0^{(\cdot)}$. In addition, the linear or index terms on this level such as provider position are included in $\boldsymbol{V}_0^{(\cdot)}$. The third and the fourth level-3 equations model slope-specific heterogeneity of age plus additional linear terms $\boldsymbol{V}_3^{(\cdot)}$, and random slopes $\boldsymbol{\varepsilon}_3^{(\cdot)}$. Finally team-specific intercepts are modeled through level-4 equations containing the logarithm of average facility distance $\boldsymbol{V}_{0,0}^{(\cdot)}$ and facility random intercepts $\boldsymbol{\varepsilon}_{0,0}^{(\cdot)}$.

4.3. Analyses

We perform sensitivity analysis for component selection with regards to different hyperparameter settings, i.e. $v_0 = 0.00025, 0.005, 0.01$ and $(a_\rho, b_\rho) = (5, 25), (5, 50), (10, 35)$. We also evaluate the prediction performance of models with and without having higher level hierarchies based on deviance values obtained for a test subset containing 1,000 observations.

4.4. Results

The maximal model contains approximately 121 model terms with 640 coefficients in total. The hyperparameters are set to $(a_\omega, b_\omega) = (1, 1)$, $(a_\rho, b_\rho) = (5, 25)$, and $v_0 = 0.00025$. Since we convert our responses to Gaussian, a very flat hyperprior $\phi \sim \Gamma^{-1}(10^{-4}, 10^{-4})$ is chosen for the error variance. The estimates are constructed on MCMC samples from ten parallel chains with a burn-in run of 1,000 iterations each, followed by a sampling phase of 15,000 iterations, with every tenth iteration used. For modeling smooth terms we use cubic P-spline basis functions with 20 equidistant inner knots over the range of the covariates plus second-order difference penalties penalizing deviations from linearity. For linear/polynomial terms we use orthogonal bases functions of the associated degree without an intercept. For modeling index effects we employ dummy variables with sum-to-zero contrasts. The correlation structures of the random effects ('team-ind' and 'fac-ind') are set to identity here, but more complex classes such as autoregressive or spatial correlation can be applied.

The model terms with posterior inclusion probability $P(\delta_j = 1|\boldsymbol{y})$ greater than 0.10 are listed in Table 2, for the primary care relative value unit, and in Table 3 for the non-primary care value unit. Compared with the non-primary care RVU, the model for the primary care RVU is rather sparse with only 10 terms with inclusion probability larger than 0.10. In both models, the team and facility random intercepts accounted for hierarchical heterogeneity turn out to be very imperative. Four other terms are also common in the two models, i.e., linear part of CAN score, marital status, whether the patient has been diagnosed with a musculoskeletal or connective tissue condition, and whether the patient has had a screening or history of disease. In terms of disease variables, the non-primary care additive predictor is almost dominated by cancer, eye, mental, skin, ear/nose/throat, and injury/poisoning, while nutrition/metabolic and heart diseases are more prominent in the primary care additive predictor. The posterior mean of the nonparametric additive predictor η associated with a number of selected effects along with 90% credible intervals are illustrated in figures11-13 for the primary care RVU, and

in figures 14-17 for the non-primary care RVU. As shown in figure 9, the care assessment need score effect on the primary care RVU is increasing from about -0.2 to +0.2 with a zero effect around 50. However, on the non-primary care RVU, the CAN score has a greater effect changing from -1 to +1 (figure 12). The effects of comorbidities are shown in figures 10 and 14. As expected, having a comorbid condition is always associated with greater clinical workload in both primary and non-primary care settings. The interaction effect of CAN score and priority on the non-primary care RVU (figure 16) shows that, patients in priority groups such as 5 and 6 are likely to generate more workloads as their CAN score increase. Yet, patients in other groups like 8 and 2 have a decreasing trend with regards to increasing in CAN score. The interaction effect of age and provider position on the non-primary care workload is oscillating with a direction change around the age of 58. The effect of length of stay on the non-primary care workload is much higher than all other covariates. Its interaction with priority group is also illustrated in figure 17 showing a large positive effect in group 7 and a large negative one in group 6.

In the test set containing 1,000 independent patients, the selected covariate set is the same as in Table 2 for the primary care RVU, except that there is no interaction effect identified; for the non-primary care workload prediction, the model includes exactly the same terms as shown in Table 3. This finding assures the stability of our approach and reinforces its internal validity (or reproducibility) with related samples underlying a same population.

We then perform predictive performance evaluation with different hyperparameter settings. To this end, the mean posterior deviance $\frac{1}{T}\sum_t^T -2l\left(\boldsymbol{y}|\boldsymbol{\eta}^{(t)}, \boldsymbol{\phi}^{(t)}\right)$, the average of twice the negative log-likelihood of the observations over the saved MCMC iterations, is calculated and saved. Results confirm that the prediction accuracy is very robust across all the parameter combinations for both primary care and non-primary care workloads. However, variable selection is a little bit sensitive to the varying hyperparameters especially to the choice of v_0. Generally we observe that very small values of v_0 allow small effects to be included in the model, while larger values of v_0 do more

conservatively. The model sparsity is found to be more sensitive with regard to v_0 than toward (a_ρ, b_ρ).

Examining the hierarchical versus nonhierarchical modeling, we notice that the mean posterior deviance is much smaller when we include random intercepts from level-3 and level-4 hierarchies. Specifically for the primary care RVU in the test set the reductions in deviance are 186 and 53 units with regards to the team and facility intercepts, respectively with the null deviance equal to 1932. For the non-primary care workload these cuts are found to be 197 and 64 units. Thus we perceive that ignoring the hierarchical structure of data, which introduces nested correlations among observations, can result in a biased prediction of both outcomes.

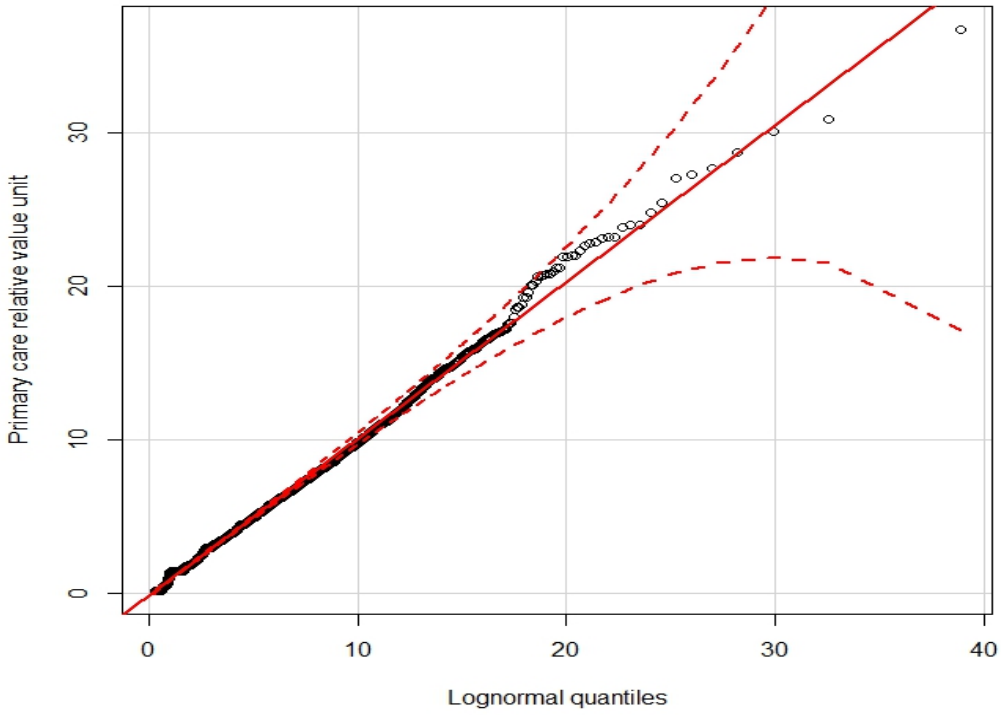


Figure 9 QQ plot of primary care relative value unit with 95% confidence bands

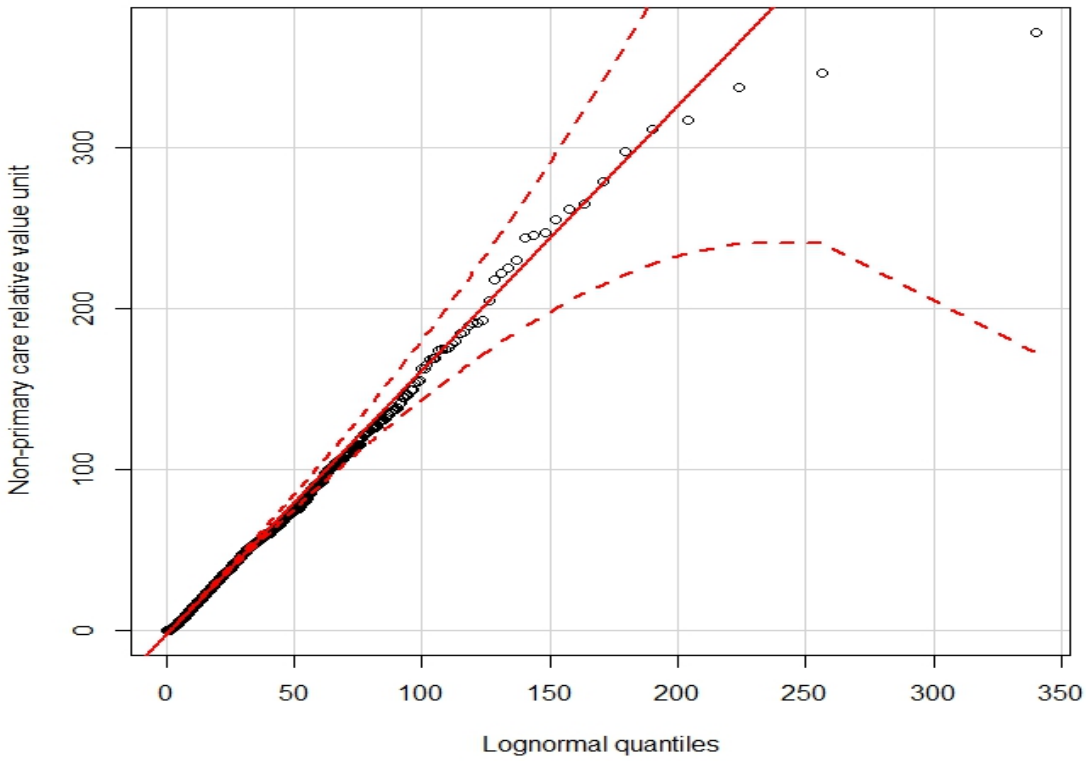


Figure 10 QQ plot of non-primary care relative value unit with 95% confidence bands

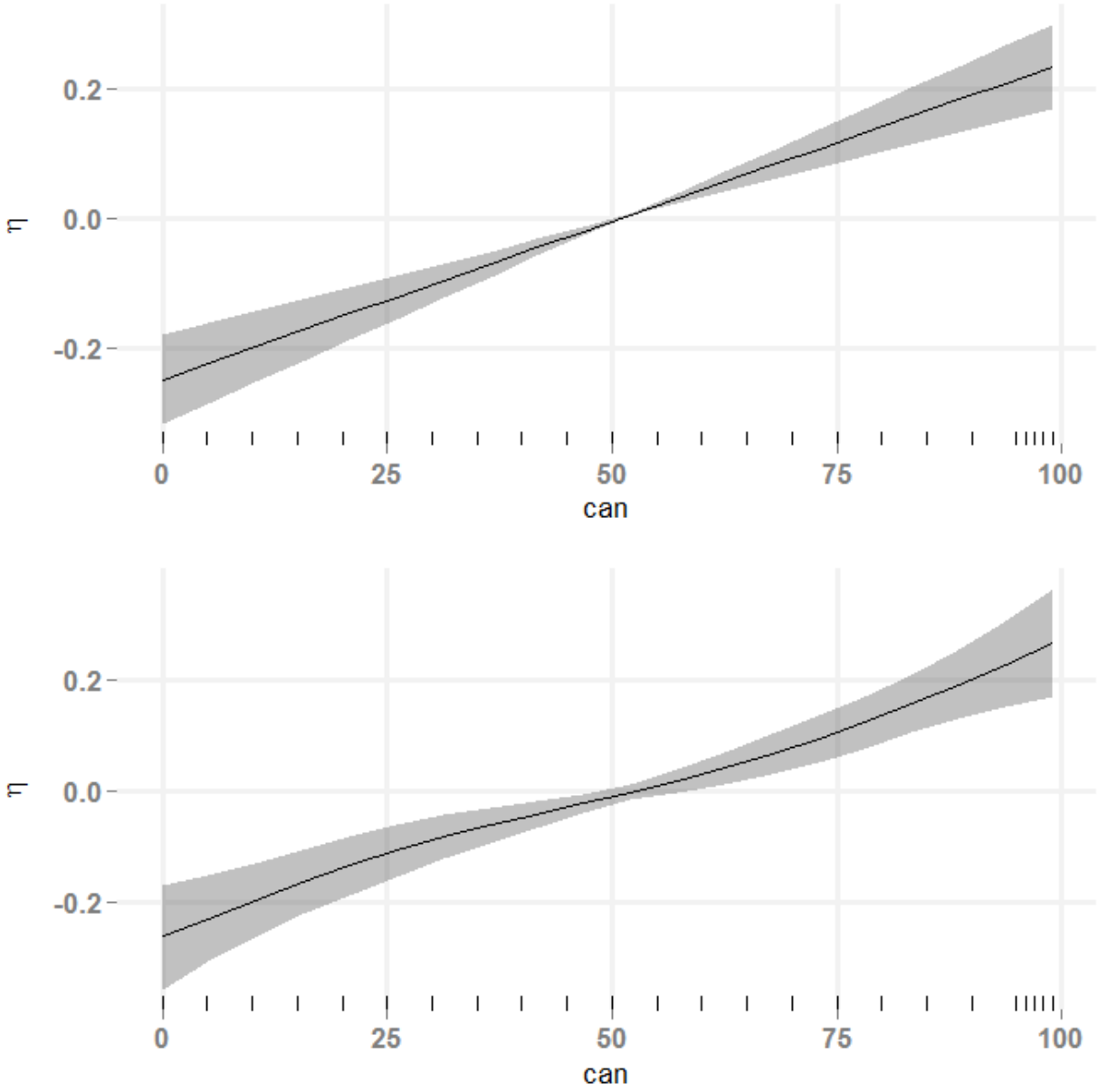


Figure 11 Linear (top) and nonlinear (down) effects of care assessment

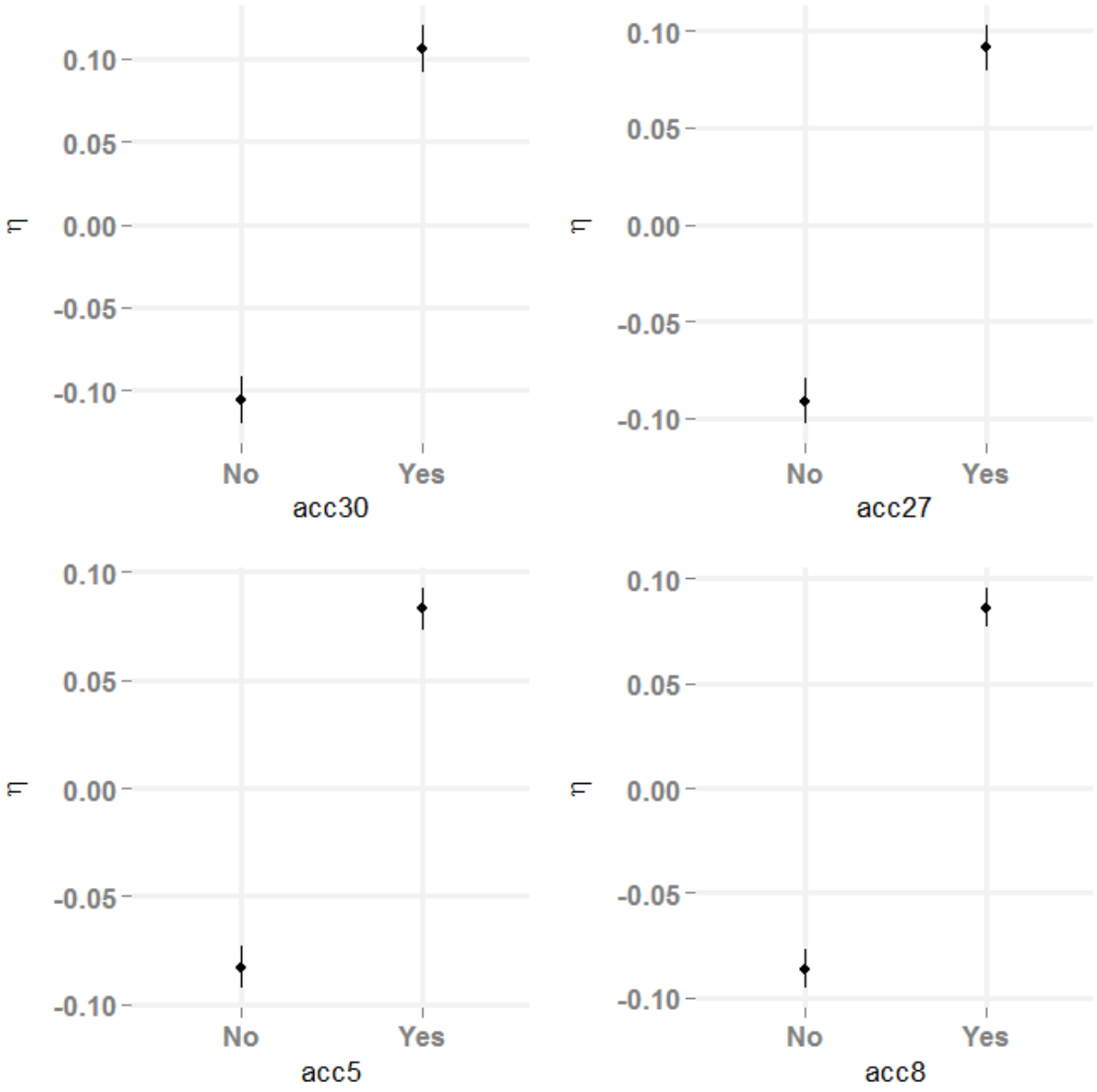


Figure 12 Effects of different comorbid conditions

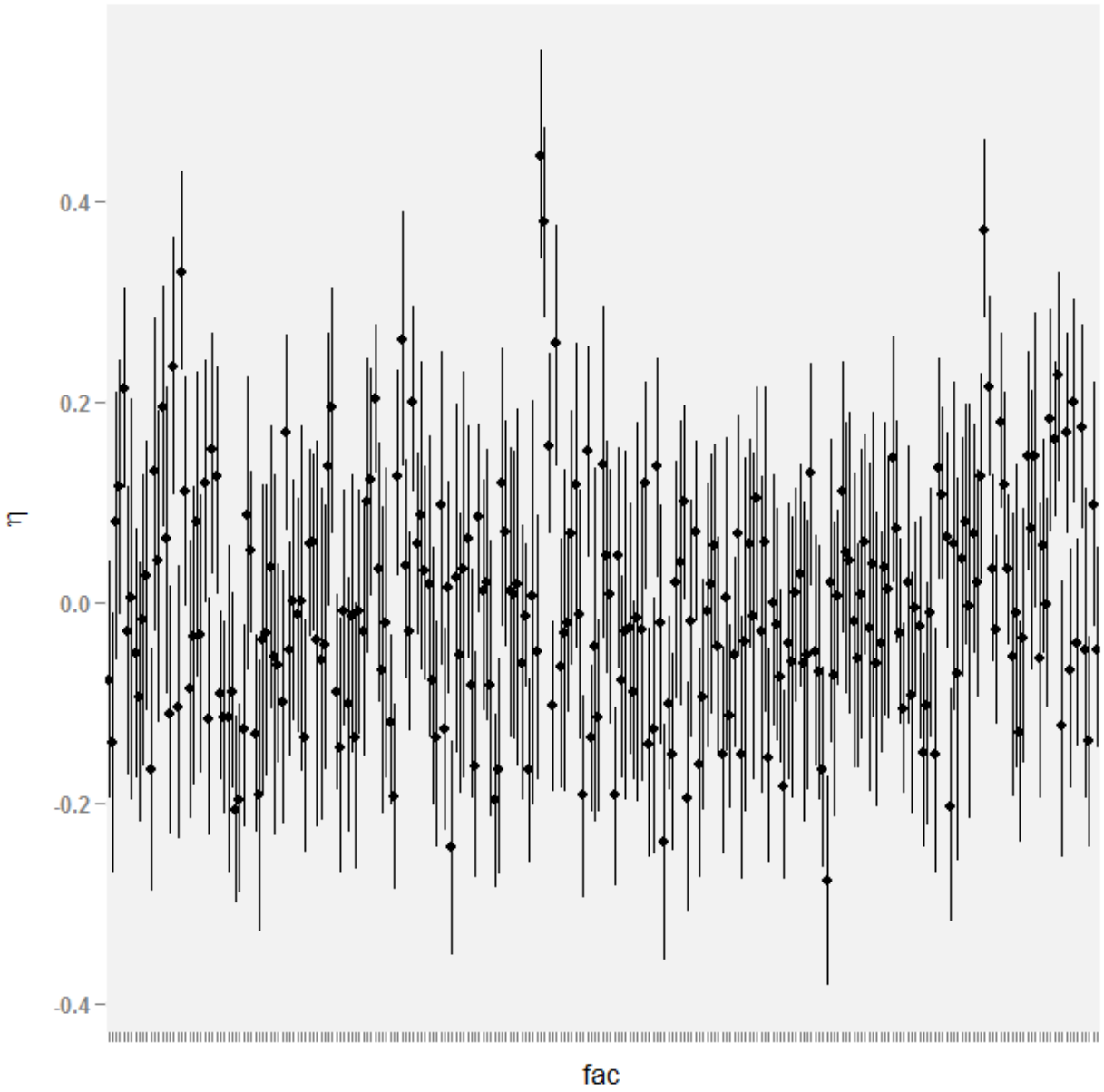


Figure 13 Effects of different facility on the primary care

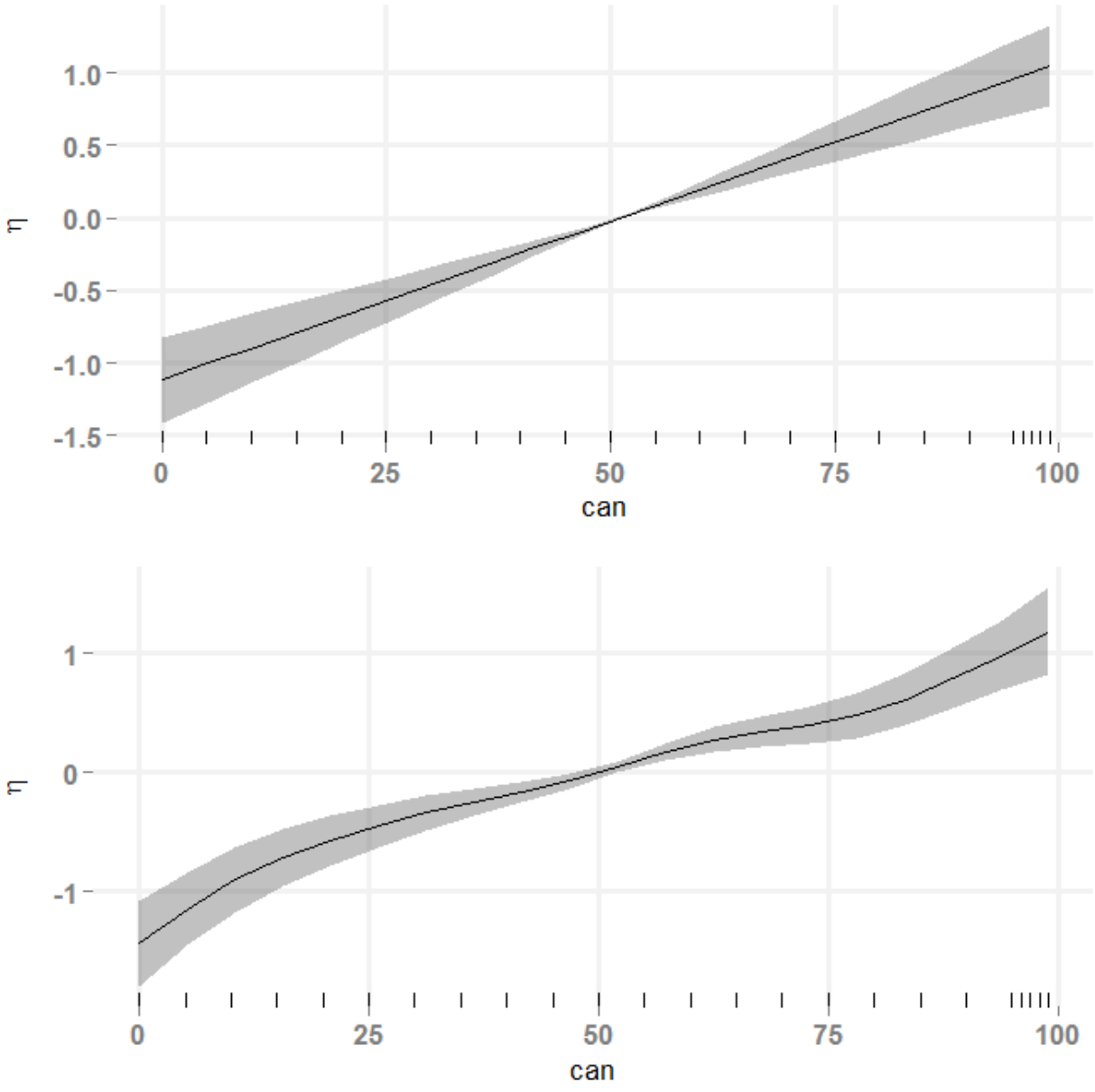


Figure 14 Linear (top) and nonlinear (down) effects of care assessment

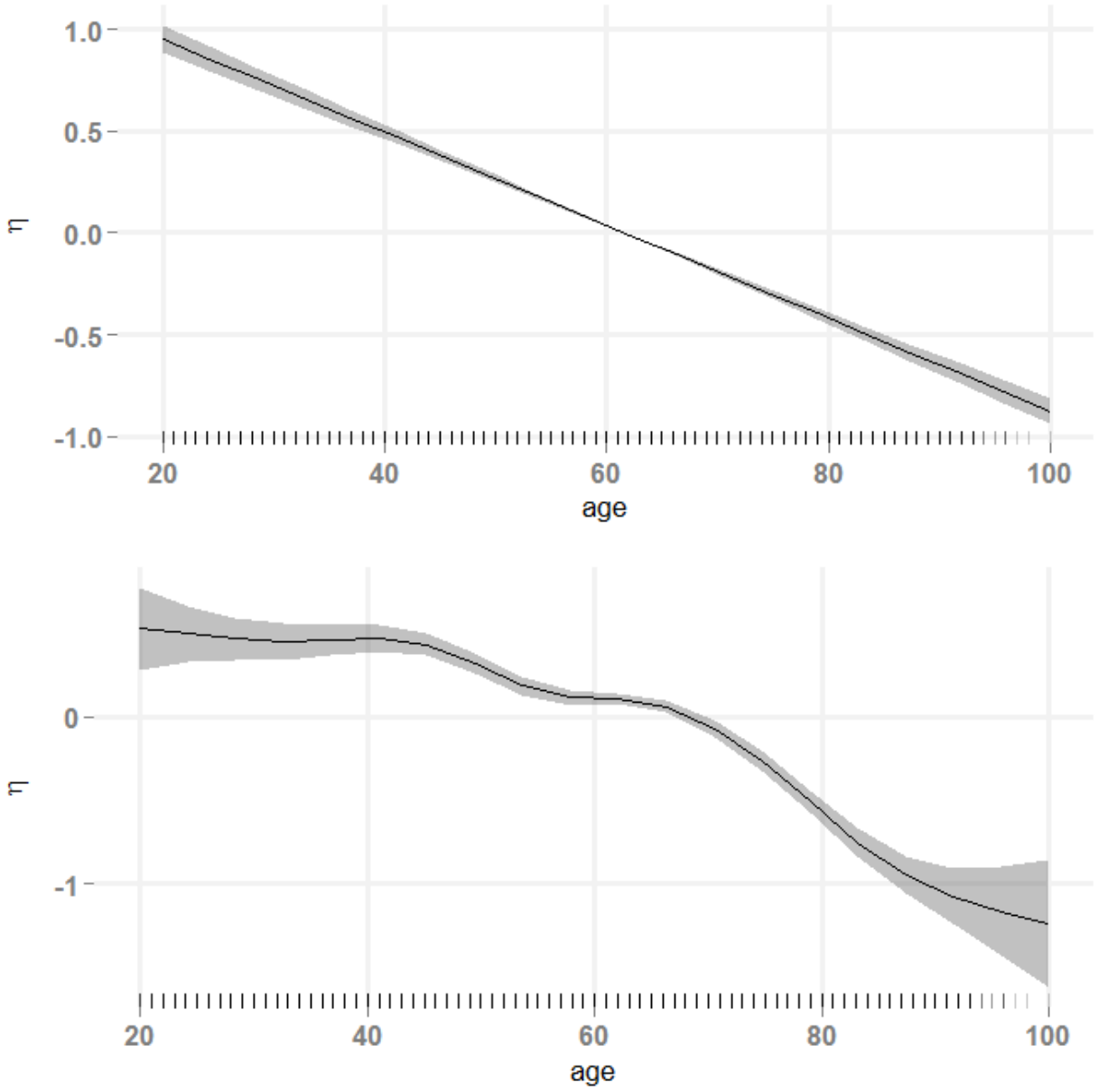


Figure 15 Linear (top) and nonlinear (down) effects of age on NPC

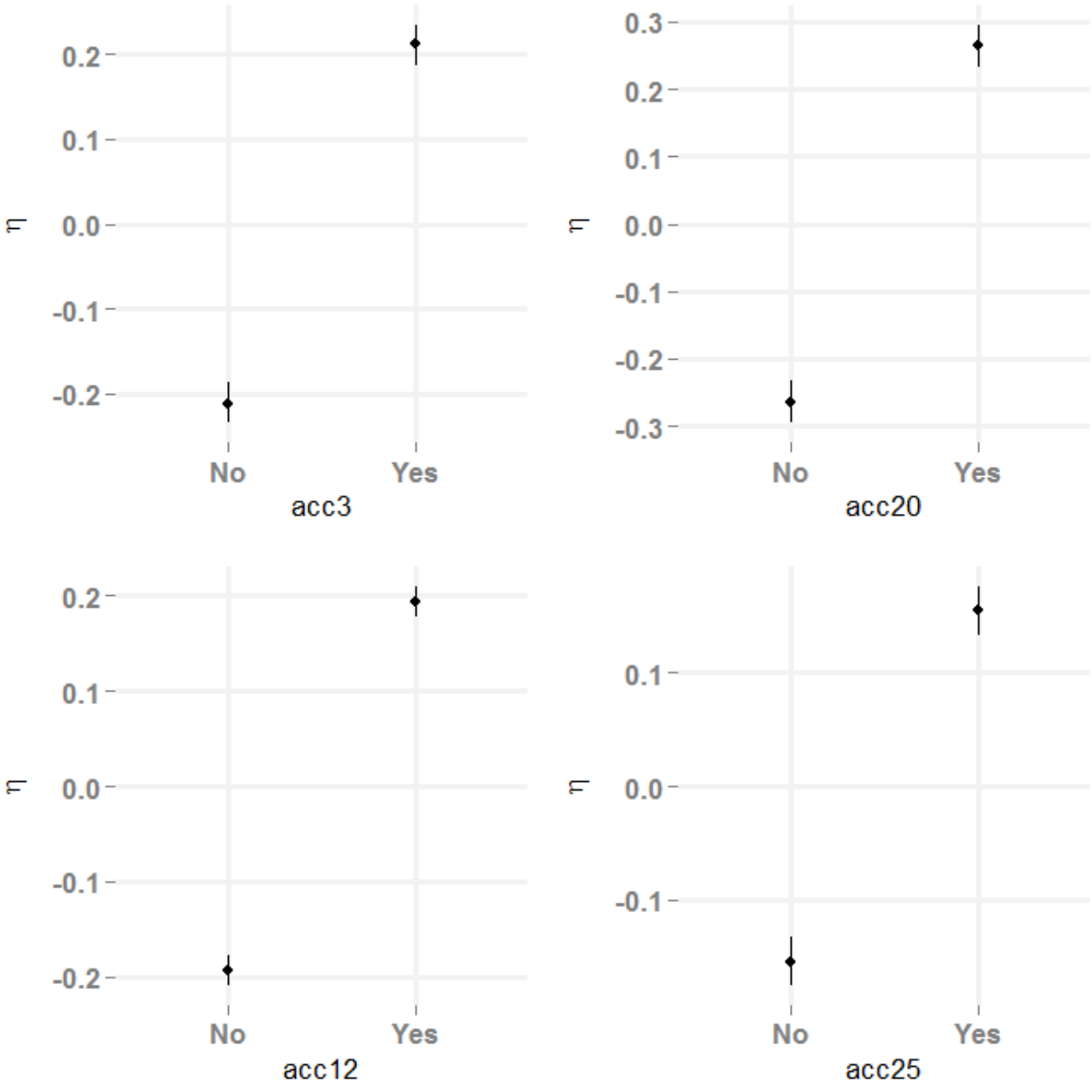


Figure 16 Effects of different comorbid conditions on the non-primary care

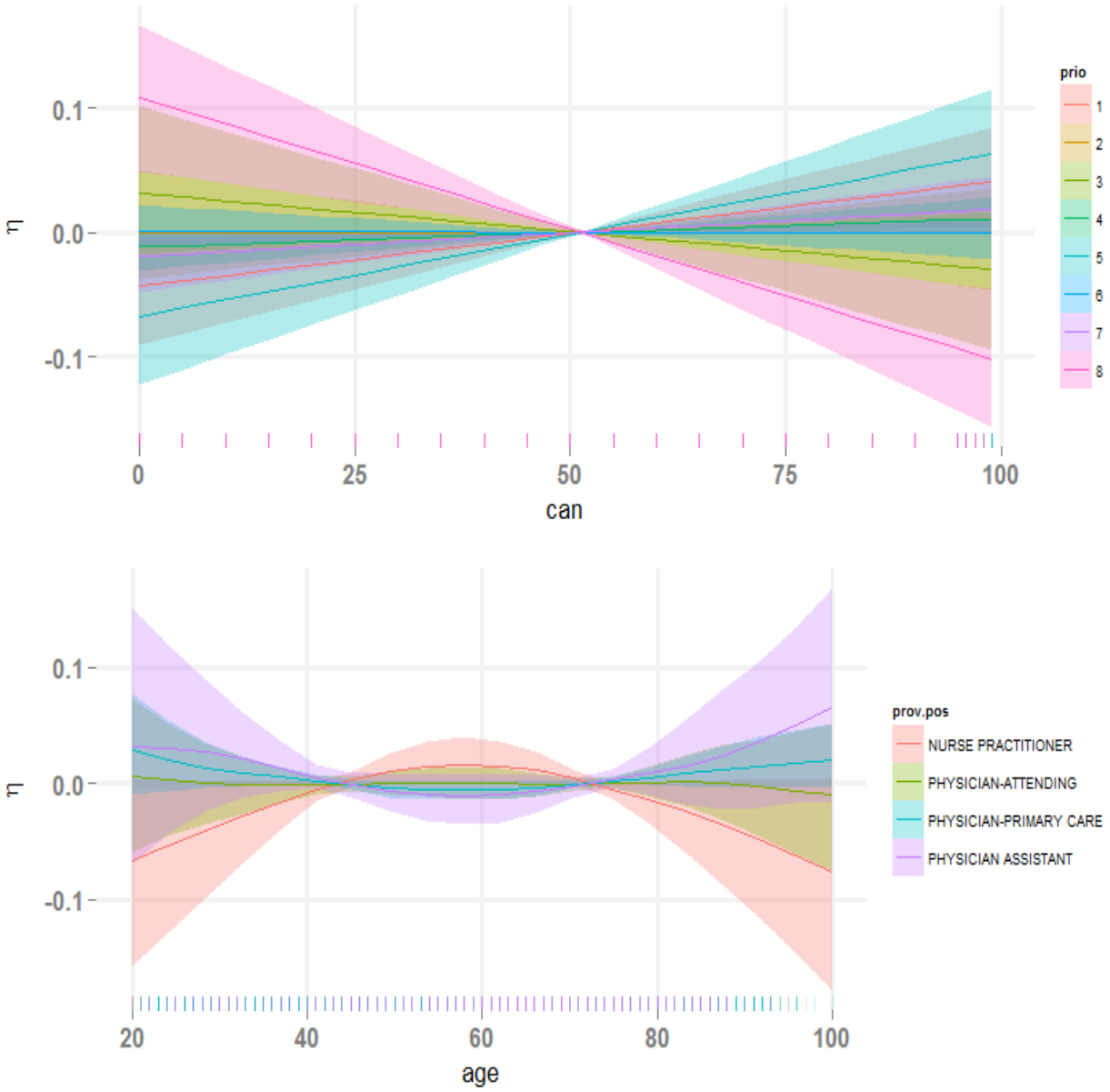


Figure 17 Interaction effects of care assessment need score and enrollment priority

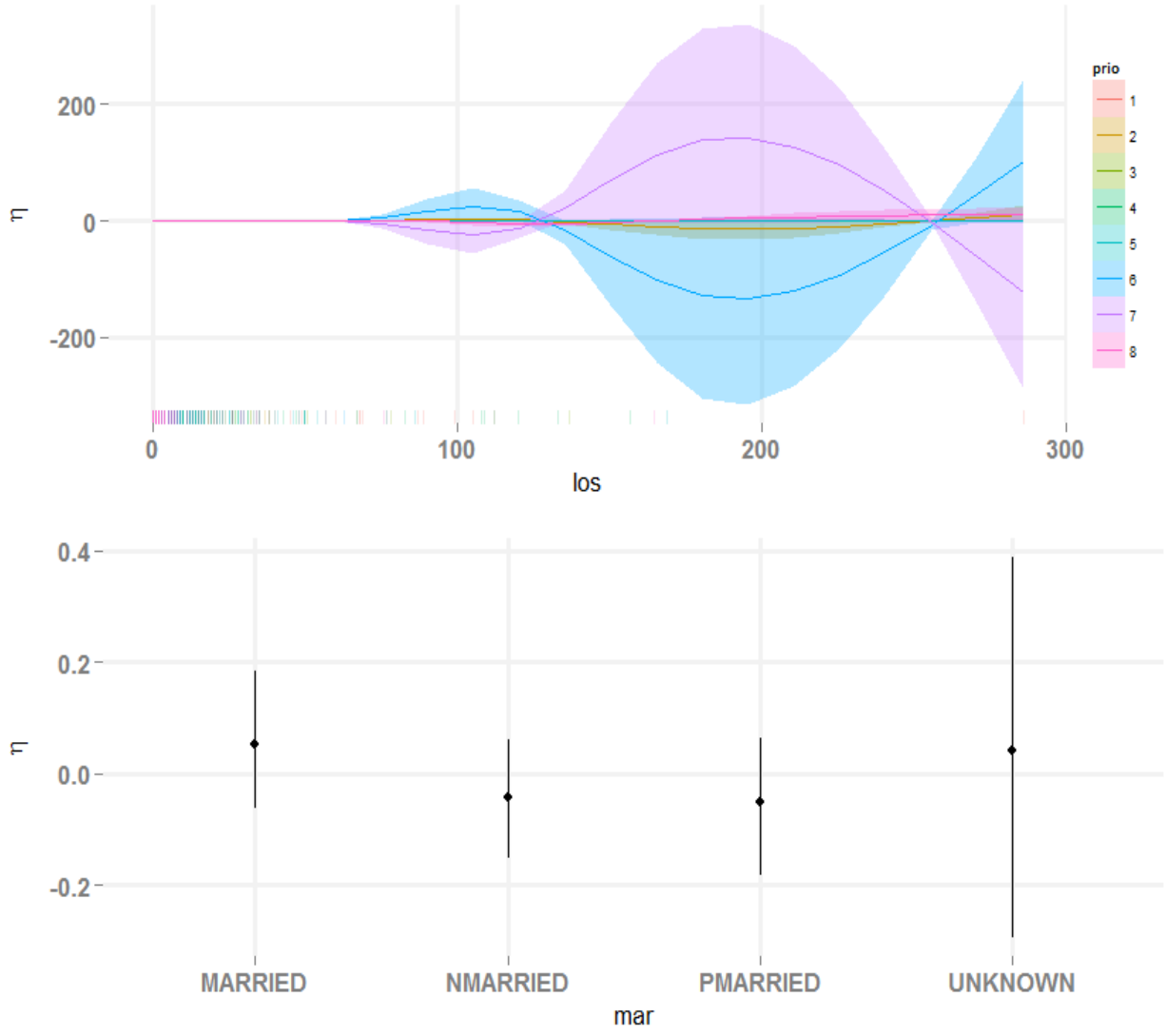


Figure 18 Interaction effects of length of stay and enrollment priority (top

4. 2. 3. Computational study for phase II

In this section we present results from an empirical study for outpatient assignment in John D. Dingell VA medical center in Detroit, Michigan. First we describe the patient

data used and how parameters are estimated from historical records to generate problem instances for the optimization problem. Next, we present results from numerical experiments illustrating the quality of solutions and value of stochastic solution (VSS). Finally, we report results from a series of sensitivity analyses done regarding to scenario grouping and input parameters such as the penalty parameter ρ and the number of scenarios considered.

The proposed algorithm was encoded in Pomo algebraic modeling language and used CPLEX 12.6 callable library in Python 2.7. We set absolute (cplex) mipgap tolerance to 1e-4 and we applied the default settings for the clique and feasibility pump switches in the solver. Moreover, we set the MIP search emphasis to "moving best bound" (option 3) and maximum cplex run time to 6 hours in scenario subproblems. Numerical experiments were run on a 7-core Dell 2.00 GHz machine with 16 GB RAM. Patient healthcare demands on different professional lines were generated using multivariate prediction models in R language [60].

4.2.3.1 Patient Data and Problem Instances

We used a set of 1,000 ($= |I|$) randomly selected patients visited the VA facility during FY 2012-13. The Decision Support System (DSS) and National Patient Care Database (NPCD) files of the VA Corporate Data Warehouse (CDW) were employed to extract patient-level factors such as demographic and socioeconomic variables. All patient visits to primary care were assembled for a total capture period of one year. Visits from other primary care related clinics, such as internal medicine or geriatric primary care, were dropped because health services requested by such visits are generally not rendered

through medical homes. The healthcare demands generated on each professional line were measured in Relative Value Unit (RVU; Dummit, 2009) on a yearly basis. The RVU schema has been widely used for reimbursement and each value was assigned to a particular service (as defined by a coding system called Current Procedural Terminology or CPT) rendered by a provider. The values are adjusted by geographic regions so that, for example, a 99213 CPT code (refers to office/other outpatient services) performed in Manhattan is worth more than when performed in Dallas. One advantage of using RVUs in our approach as opposed to simple face-to-face visit counts lies in its ability to further accommodate workloads that are generated by telephone encounters.

We designed 5 problem instances in this study. According to the VA administrative records, for the sample taken, there were 3 ($= |J|$) PCMH team arranged in Detroit facility and most of the time four professional lines were working in each team: NP (or PCP or Physician's Assistant), RN, LPN (or equivalent), and medical clerk. However there were some cases in which a patient in the sample called for a nutritionist or pharmacist during his/her visit. Thus, to be on the safe side, we set the number of professions ($= |R|$) to 5. To generate data for each problem instance we employed multivariate predictions that can take into account the interdependencies among demands on professional lines [21]. On our sample such a prediction model produced 8% mean absolute percentage error (MAPE) averaged over all five professions. Thus we used those predictions to generate healthcare demand scenarios as follows. We used empirical prediction errors to populate the scenarios generated. In particular, for each instance we

first took a *bootstrap* sample (with replacement) of 1,000 patients and computed the prediction and empirical error vectors (5-dimension) for each record. We repeated this process 10 times (= $|\mathcal{S}|$) and we added the prediction error vectors to the point estimate vectors to obtain a demand scenario. Note that the probability of each scenario is equally likely, that is, $p_s = \frac{1}{|\mathcal{S}|}, \forall\, s \in \mathcal{S}$.

Now we explain how available service times are calculated. It is noted that, in most VA PCMH practices, a "basic" member is assigned to a single team and 100% of its time (1 Full-Time Equivalent) is devoted to that team alone, that is, they are not shared among other teams. These basic positions are NP (or PCP), RN, LPN, and medical clerk. However, there are shared positions like nutritionist and pharmacists that are consulted on a referral basis as needed by the patient. Simply put, a nutritionist may have patients from all of the PCMH teams or from just one or two depending on the needs of the patients, not necessarily depending on the teams themselves. Thus, to calculate the available annual service time granted in each team, we added hours provided by the shared positions to those given by the basic members (= b_{jr}). Since we have three teams in the studied facility, we assumed that a nutritionist (or a pharmacist) grants one-third of his/her service times for each team. This assumption is reasonable and in line with the VA policy, though it can be modified when applied in other settings. As a result, the RHS of (6e) is changed as $b_{jr} \longleftarrow b_{jr} + \frac{1}{3} b_{j,Nut} + \frac{1}{3} b_{j,Pharm}$. For full-time federal employees there are approximately 260 working days with approximately 23 days granted for vacation, which gives a total of 1890 hours per year. The remaining is related to the

maximum RVU values per hour for each profession, which was obtained by the CPT codes performed and the Resource-Based Relative Value Unit schema. For example, the maximum RVU/hour for a physician was 12 in the Detroit facility, so he/she could deliver about 1890×12 RVUs during a year.

We assumed the cost of initial patient assignment ($=\tilde{c}_{ij}, \forall\ (i,j) \in (I\times J)$) is zero or equivalently this is the same for all patients considered so that the total first-stage cost becomes constant. Nonetheless, all results and analyses can be applied to non-zero first-stage costs as well. We set the cost parameters as relative weights. Based on the VA primary care policy, the assignment cost is similar for all PCMH teams but can be different for different patient types. There were four distinct types of patients based on primary Care Assessment Need (or CAN) score in the sample. The CAN score is a general illness severity score that reflects the likelihood of admission or death within a specified time period, and it works somewhat similar to diagnostic cost group (DxCG) risk score. The score is commonly expressed as a percentile ranging from 0 (lowest risk) to 99 (highest risk) and it shows how a VA patient is compared with others pertaining to the chances of hospitalization or death. Thus, we set the assignment costs ($= c_{ij}$) as 1, 1.5, 2, and 2.5. The overtime penalty cost ($= \beta_r$) for PCP, RN, LPN, Nutrition or Pharmacist, and medical clerk was set to 9, 7, 5, 3, and 1, respectively. The reassignment costs were set sufficiently large ($\alpha_i = 20$) to trigger "continuity of care" in the solutions implying that, unless necessary, a previously established patient assignment should not be changed.

4.2.3.2 Value of the Stochastic Solution

In this section we examine the value of information and benefits of applying our stochastic model to the problem instances. The value of stochastic solution (VSS) shows the expected loss of ignoring uncertainty when, instead of stochastic model, we solve the mean value problem in which all random variables are replaced by their means. Table 9 presents the solutions for deterministic model, stochastic program TSSPA2 (with 10 scenarios per each instance), and the VSS for each problem instance. For this, we set the algorithmic parameters as $\rho^0 = 0.5$, $\partial = 1.04$, and $\varepsilon = 1e - 2$. Note that "Asg. cost" and "ReAsg. Cost" denote assignment cost and reassignment cost. As appeared, the use of stochastic model saves the cost of 271 units on average. Of interest, the reassignment cost and overtime cost move in opposite direction. This happens because when healthcare supplies are insufficient to fulfill the demands, we have to either move patients to other PCMH teams or ask current staffs to do overtime shifts. The deterministic solutions have more assignment costs than stochastic solutions.

4.2.3.5 Computational experiments

In this section we evaluate the performance of the proposed algorithm as a function of the number of scenarios considered. We keep the algorithmic parameters fixed but we performed warm-starting the individual scenario subproblem for iteration $\nu \geq 1$ using solutions from the previous iterations. A maximum runtime of 2 hours was allowed for the algorithm and the cplex branch and bound was terminated at 6 hours. For each case, we record the incumbent objective value at the termination, MIP lower bound, and optimality gap relative to the objective value. The results displayed in Table 8 point to the

difficulty of solving the extensive form of the TSSPA2 problem. In no case was an optimality gap less than 0.54% examined. As the number of scenarios increases, the problem becomes harder for the algorithm, which cannot solve the problem for any but the first smallest instance within the allocated time.

Table 8 Solution quality statistics for the proposed algorithm

# Scenarios	Best Objective Value	MIP Lower Bound	% Gap
5	11691.3	11628.7	0.54
10	11335.5	11232.4	0.91
25	11305	11123.2	1.61
50	11414.1	11037.3	3.30
100	11977.7	10946.1	8.61

Table 9 Comparison of deterministic solutions and stochastic solution, and VSS

	Deterministic solutions				Stochastic solutions				
Instance	Asg. Cost	ReAsg. Cost	Overtime Cost	Total Cost	Asg. Cost	ReAsg. Cost	Overtime Cost	Total Cost	VSS
1	1619	874	8872.6	11365.6	1583.6	760	8739.1	11089.2	276.4
2	1624.5	988	8790	11402.5	1610	874	8654.3	11138.3	264.2
3	1627	912	8934	11473	1612.5	722	8857.4	11191.9	281.1
4	1638	760	9110.3	11508.3	1613.5	608	9016.5	11238	270.6
5	1625.5	836	8894.7	11356.2	1617.5	798	8678.2	11093.7	262.5
Mean	1626.8	874	8920.3	11421.1	1607.4	752.7	8789.1	11150.2	271
StDev	6.9	85	118.5	67	13.6	98.4	149.4	64.2	8

Chapter V: Conclusion

A key factor in the success of medical homes in delivering quality and coordinated care lies in their teams' ability to handle uncertainties that can be caused by different sources such as patient/physician appointment scheduling, care logistics, and more importantly patients' health demands. This paper addresses the problem of clinical demand prediction in the presence of nested sources of variation at different operational levels. We collected outpatient visit data from a large sample of Veterans Affairs hospitals and investigated the relationship between risk factors at three operational levels and total care demands on a yearly basis. We propose a multivariate multilevel generalized linear model in a Bayesian framework to predict the care demand portfolio in medical home practices. The proposal can fit heteroscedastic variances and unstructured covariance matrices for nested random effects and residuals as well as their interactions with categorical and continuous covariates simultaneously.

We find that utilizing a multilevel analysis with nested random components can greatly contribute to model fit in hierarchical healthcare systems. Further, we show that risk-adjustment for patient disease conditions and their comorbidities extensively enhance the prediction power of our model. Our results confirm that using a multivariate as opposed to a univariate approach can significantly shrink the correct credible intervals for workload predictions thus allowing for a more precise estimation of either outcome. The approach used in this paper has a general application and could also be employed for analysis of multiple health outcomes in a variety of health analytics contexts.

Turning to specific results from recent VA data, we see that overall, the primary care is positively associated with the non-primary care after accounting for all studied sources of variability. We find the association between patient-level predictors such as age and the care workloads varies considerably among PCMH teams within a hospital. Further, the effect of patient non-adherence on care demands is subject to change from one hospital to another. Moreover, it is found that patient oldness can contribute to the increased care demands required for heart, nutritional, and gastrointestinal diseases.

There are some limitations to this research that need to be mentioned. First, the data in our study are collected solely from a veteran population (with fewer female and more senior patients) who receives support from government budgets. Thus the results from our study may not fully generalize to other health care systems. Second the data used is administrative and not real time, so some issues such as model tuning and calibration should be taken into account when dealing with online prediction efforts.

Our work can further be extended in some fronts. One challenging direction would be to modify the proposed approach to handle longitudinal observations from past history of care demands for a specific patient profile. This may be done by expanding the multivariate distribution of outcomes to include a temporal dimension which requires great care in model specification and implementations thanks to various inter-correlations. Alternatively, one can combine some autoregressive terms to the variance structure introduced in this work. Another issue worth exploring is related to the way that one can adjust for patient risk or comorbidities. Although several algorithms such as Clinical Risk Group (CRG), *veriskhealth* DxCG [82], and CMS's HCC software

have been used in the literature, no scientific study is available to systematically evaluate the impacts of each algorithm on prediction modeling of care demands.

Moreover, we propose a Bayesian function selection approach based on spike and slab priors for the hierarchical structured additive models with a multivariate response. The prior setting adopted in our work is a Bayesian hierarchical structure with a bimodal density on the hyper-variance of the coefficient blocks with one part being a narrow spike around the origin and the other part being a wide slab. We demonstrate how one can parameterize a special class of multi-response hierarchical structured additive model, that is, a multivariate linear multilevel spline model, within a standard structural equation modeling framework, and thus bridge the connection between multivariate multilevel STAR models and generalized latent variable models. We then apply our methods to patient centered medical home data obtained from a large number of VA medical facilities during fiscal years 2011–12. Our work is the first attempt to develop a portfolio based demand prediction model for patient centered medical home within the OR/MS or IE community. We aggregate three levels of hierarchical data including information from outpatients, the medical team responsible to render the care to the patients, and the VA facilities. We find that the sets of chosen predictors introduced by the model are different for the primary care and the non-primary workloads. Our findings also confirm that taking hierarchical heterogeneity into account is associated with better prediction accuracy, especially when the data has more than two levels. Moreover, in this research we prposed we presented a balanced patient assignment model under healthcare demand uncertainty with application to patient centered medical home. The model was formulated

as a two stage stochastic integer program with mixed 0-1 recourse. An assignment of patients to medical home teams is decided a priori, and once the actual demand is revealed, reassignments can be performed if there are overloaded team members. Different penalties were considered for reassigning patients with positive demands and calling for personnel to do overtime services. The objective is to minimize the total expected costs. We proposed an efficient scenario decomposition strategy inspired by the Rockafellar and Wets progressive hedging approach to address the problem. We also presented a lower bound for mixed integer case that can be found in every iteration of the algorithm. The algorithm outperforms the commercial solver when directly applied to the multiscenario formulations in both solution quality and computational time. We applied our methods to an empirical study for outpatient assignment in patient centered medical home at John D. Dingell VA medical center. Problem instances were generated using a multivariate prediction model that estimated correlated demands with an acceptable error rate. Our findings indicate that solving the stochastic problem, as compared to the mean value problem, would save the cost of 271 units on average. We conducted numerical tests to evaluate the effect of number of scenarios and quality of lower bounds on the performance of the proposed algorithm. We found that significant amount of computing is pertained to solving the scenario subproblem which can be saved by parallel computation of the subproblem.

REFERENCES

[1] Green LV, Savin S, Murray M (2007) Providing timely access to care: what is the right patient panel size? Joint commission journal on quality and patient safety 33 (4):211-218

[2] Murray M, Davies M, Boushon B (2007) Panel size: answers to physicians' frequently asked questions. Family practice management 14 (10):29-32

[3] Ostbye T, Yarnall KS, Krause KM, Pollak KI, Gradison M, Michener JL (2005) Is there time for management of patients with chronic diseases in primary care? Annals of family medicine 3 (3):209-214

[4] Naessens JM, Stroebel RJ, Finnie DM, Shah ND, Wagie AE, Litchy WJ, Killinger PJ, O'Byrne TJ, Wood DL, Nesse RE (2011) Effect of multiple chronic conditions among working-age adults. The American journal of managed care 17 (2):118-122

[5] Potts B, Adams R, Spadin M (2011) Sustaining primary care practice: a model to calculate disease burden and adjust panel size The Permanente journal 15 (1):53-56

[6] Balasubramanian H, Banerjee R, Denton B, Naessens J, Stahl J (2010) Improving clinical access and continuity through physician panel redesign. Journal of general internal medicine 25 (10):1109-1115

[7] Rittenhouse DR, Shortell SM, Fisher ES (2009) Primary care and accountable care--two essential elements of delivery-system reform. The New England journal of medicine 361 (24):2301-2303

[8] Patient-Centered Primary Care Collaborative (PCPCC). (2009). Available at http://www.pcpcc.org/about/medical-home. Accessed August 14, 2013

[9] Backer LA (2009) Building the case for the patient-centered medical home. Family practice management 16 (1):14-18

[10] Fisher ES (2008) Building a medical neighborhood for the medical home. The New England journal of medicine 359 (12):1202-1205

[11] Beal AC, Fund C (2007) Closing the divide: how medical homes promotes equity in health care. Commonwealth Fund

[12] Rittenhouse DR, Shortell SM (2009) The patient-centered medical home: will it stand the test of health reform? Journal of the American Medical Association 301 (19):2038-2040

[13] Reid RJ, Coleman K, Johnson EA, Fishman PA, Hsu C, Soman MP, Trescott CE, Erikson M, Larson EB (2010) The group health medical home at year two: cost savings, higher patient satisfaction, and less burnout for providers. Health affairs 29 (5):835-843

[14] Bates DW, Bitton A (2010) The future of health information technology in the patient-centered medical home. Health affairs 29 (4):614-621

[15] Bitton A, Martin C, Landon BE (2010) A nationwide survey of patient centered medical home demonstration projects. Journal of general internal medicine 25 (6):584-592

[16] Klein S (2011) The Veterans Health Administration: implementing patient-centered medical homes in the nation's largest integrated delivery system. Commonwealth Fund

[17] P. A. Nutting, W. L. Miller, B. F. Crabtree, C. R. Jaen, E. E. Stewart, and K. C. Stange. Initial Les- sions From the First National Demonstration Project on Practice Transformation to a Patient-Centered Medical Home. Annals of Family Medicine, 7(3):254 – 260, 2009.

[18] C. R. Jaen, R. L. Ferrer, W. L. Miller, R. F. Palmer, R. Wood, M. Davila, E. E. Stewart, B. F. Crabtree, P. A. Nutting, and K. C. Stange. Patient Outcome at 26 Month in the Patient-Centered Medical Home National Demonstration Project. Annals of Family Medicine, 8(1):557 – 567, 2010.

[19] B. F. Crabtree, Nutting P. A., Miller W. L., Stange K. C., Stewart E. E., Palmer K. C., and Jaen C. R. Summary of the National Demonstration Project and Recommendations for the Patient-Centered Medical Home. Annals of Family Medicine, 8(1):580 – 590, 2010.

[20] Ajorlou, Saeede, Issac Shams (2014), and Kai Yang. "An analytics approach to designing patient centered medical homes." Healthcare Management Science, Ahead of print, DOI: 10.1007/s10729-014-9287-x.

[21] Shams, I., Ajorlou, S., Yang, K. (2014). A multivariate hierarchical Bayesian framework for healthcare predictions with application to medical home study in the Department of Veteran Affairs. arXiv preprint arXiv:1403.0674.

[22] Shams, I., Ajorlou, S., Yang, K. (2014). A Two Stage Stochastic Mixed 0-1 Programming Approach for balanced Patient Assignment in Patient Centered Medical Home. Submitted to IIE Transactions on Healthcare Systems Engineering.

[23] Goldstein, H. (2011). Multilevel statistical models: John Wiley & Sons. Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 55(4), 757-796.

[24] Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. & Kneib, T. (2013). Multilevel structured additive regression. Statistics and Computing, 24(2), 223-238.

[25] Kammann, E. & Wand, M.P. (2003). Geoadditive models. Journal of the Royal Statistical Society : Series C (Applied Statistics), 52(1), 1-18.

[26] Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 55(4), 757-796.

[27] Claeskens, G. & Hjort, N.L. (2003). The focused information criterion. Journal of the American Statistical Association, 98(464), 900-916.

[28] Breiman, L. (1995). Better subset regression using the nonnegative garrote. Technometrics, 37(4), 373-384.

[29] Pu, W. & Niu, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. Journal of Multivariate Analysis, 97(3), 733-758.

[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 58(1), 267-288.

[31] Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348-1360.

[32] Chen, Y., Du, P. & Yuedong, W. (2013). Variable selection in linear models. Wiley Interdisciplinary Reviews: Computational Statistics, 6(1), 1-9.

[33] Fan, J. & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20(1), 101-148.

[34] Huang, J., Horowitz, J.L. & Wei, F. (2010). Variable selection in nonparametric additive models. Annals of statistics, 38(4), 2282-2313.

[35] Kundu, S. & Dunson, D.B. (2013). Bayes variable selection in semiparametric linear models. Journal of the American Statistical Association(just-accepted).

[36] Brown, P.J., Vannucci, M. & Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 60(3), 627-641.

[37] Cai, J., Fan, J., Li, R. & Zhou, H. (2005). Variable selection for multivariate failure time data. Biometrika, 92(2), 303-316.

[38] George, E.I. & McCulloch, R.E. (1993). Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88(423), 881-889.

[39] Liang, F., Paulo, R., Molina, G., Clyde, M.A. & Berger, J.O. (2008). Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association, 103(481), 410-423.

[40] O'Hara, R.B. & Sillanpää, M.J. (2009). A review of Bayesian variable selection methods: what, how and which. Bayesian analysis, 4(1), 85-117.

[41] Spiegelhalter, D.J., Best, N.G., Carlin, B.P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society : Series B (Statistical Methodology), 64(4), 583-639.

[42] Müller, S., Scealy, J. & Welsh, A. (2013). Model selection in linear mixed models. Statistical Science, 28(2), 135-167.

[43] Antoniadis, A. & Fan, J. (2001). Regularization of wavelet approximations. Journal of the American Statistical Association, 96(455), 939-967.

[44] Lin, Y. & Zhang, H.H. (2006). Component selection and smoothing in multivariate nonparametric regression. Annals of statistics, 34(5), 2272-2297.

[45] Marra, G. & Wood, S.N. (2011). Practical variable selection for generalized additive models. Computational Statistics & Data Analysis, 55(7), 2372-2387.

[46] Scheipl, F., Fahrmeir, L. & Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. Journal of the American Statistical Association, 107(500), 1518-1532.

[47] Lang, S. & Brezger, A. (2004). Bayesian P-splines. Journal of computational and graphical statistics, 13(1), 183-212.

[48] Kline, R.B. (2011). Principles and practice of structural equation modeling: Guilford press.

[49] Rabe-Hesketh, S., Skrondal, A. & Pickles, A. (2004). Generalized multilevel structural equation modeling. Psychometrika, 69(2), 167-190.

[50] Ishwaran, H. & Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. Annals of statistics, 730-773.

[51] Moerbeek M (2004) The consequence of ignoring a level of nesting in multilevel analysis. Multivariate behavioral research 39 (1):129-149.

[52] Goldstein H (2011) Multilevel statistical models. Wiley series in probability and statistics, 4 edn. John Wiley & Sons.

[53] Srivastava MS, von Rosen T, von Rosen D (2008) Models with a Kronecker product covariance structure: estimation and testing. Mathematical methods of statistics 17 (4):357-370

[54] McCulloch CE, Searle SR, Neuhaus JM (2008) Generalized, linear, and mixed models. Wiley series in probability and statistics, 2 edn. John Wiley & Sons .

[55] Pinheiro JC, Bates DM (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of computational and graphical statistics 4 (1):12-35.

[55] Gamerman D, Lopes HF (2006) Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press .

[56] Damlen P, Wakefield J, Walker S (1999) Gibbs sampling for Bayesian non□ conjugate and hierarchical models by using auxiliary variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61 (2):331-344.

[57] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4):583-639.

[58] Needleman, J., Buerhaus, P., Mattke, S., Stewart, M. & Zelevinsky, K. (2002). Nurse-staffing levels and the quality of care in hospitals. New England Journal of Medicine, 346(22), 1715-1722.

[59] Dummit, L. (2009). Relative Value Units (RVUs). In: National Health Policy Forum Publications. Available at http://www.nhpf.org/library/the-basics/Basics_RVUs_02-12-09.pdf, Accessed on March 18, 2014.

[60] Ajorlou, S., Shams, I. & Yang, K. (2014). An analytics approach to designing patient centered medical homes. Health Care Managemet Science. Ahead of print. doi: 10.1007/s10729-014-9287-x.

[61] Toktas, B., Yen, J.W. & Zabinsky, Z.B. (2006). Addressing capacity uncertainty in resource-constrained assignment problems. Computers & Operations Research, 33(3), 724-745.

[62] Birge, J.R. & Louveaux, F. (2011). Introduction to stochastic programming: Springer.

[63] Alonso-Ayuso, A., Escudero, L.F., Garìn, A., Ortuño, M.T. & Pérez, G. (2003). An approach for strategic supply chain planning under uncertainty based on stochastic 0-1 programming. Journal of Global Optimization, 26(1), 97-124.

[64] Mehrotra, S. & Papp, D. (2013). Generating moment matching scenarios using optimization techniques. SIAM Journal on Optimization, 23(2), 963-999.

[65] Gassmann, H.I. & Ziemba, W.T. (2013). Stochastic Programming: Applications in Finance, Energy, Planning and Logistics: World Scientific.

[66] Ryan, S.M., Wets, R.J.-B., Woodruff, D.L., Silva-Monroy, C. & Watson, J.-P. (2013). Toward scalable, parallel progressive hedging for stochastic unit commitment. In: Power and Energy Society General Meeting (PES), 2013 IEEE (pp. 1-5): IEEE.

[67] Van Slyke, R.M. & Wets, R. (1969). L-shaped linear programs with applications to optimal control and stochastic programming. SIAM Journal on Applied Mathematics, 17(4), 638-663.

[68] Rockafellar, R.T. & Wets, R.J.-B. (1991). Scenarios and policy aggregation in optimization under uncertainty. Mathematics of Operations Research, 16(1), 119-147.

[69] Hvattum, L.M. & Løkketangen, A. (2009). Using scenario trees and progressive hedging for stochastic inventory routing problems. Journal of Heuristics, 15(6), 527-557.

[70] Ryan, S.M., Wets, R.J.-B., Woodruff, D.L., Silva-Monroy, C. & Watson, J.-P. (2013). Toward scalable, parallel progressive hedging for stochastic unit commitment. In: Power and Energy Society General Meeting (PES), 2013 IEEE (pp. 1-5): IEEE.

[71] Veliz, F.B., Watson, J.-P., Weintraub, A., Wets, R.J.-B. & Woodruff, D.L. (2014). Stochastic optimization models in forest planning: a progressive hedging solution approach. Annals of Operations Research, 1-16.

[72] Gade, D., Hackebeil, G., Ryan, S., Watson, J., Wets, R. & Woodruff, D. (2013). Obtaining lower bounds from the progressive hedging algorithm for stochastic mixed-integer programs. Under Review.

[73] Meyer, R.R. (1975). Integer and mixed-integer programming models: General properties. Journal of Optimization Theory and Applications, 16(3-4), 191-206.

[74] Lulli, G. & Sen, S. (2004). A branch-and-price algorithm for multistage stochastic integer programming with application to stochastic batch-sizing problems. Management Science, 50(6), 786-796.

[75] Liu CF, Sales AE, Sharp ND, Fishman P, Sloan KL, Todd-Stenberg J, Nichol WP, Rosen AK, Loveland S (2003) Case-mix adjusting performance measures in a veteran population: pharmacy- and diagnosis-based approaches. Health services research 38 (5):1319-1337

[76] Hosmer Jr DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression. John Wiley & Sons

[77] Petersen LA, Byrne MM, Daw CN, Hasche J, Reis B, Pietz K (2010) Relationship between clinical conditions and use of Veterans Affairs health care among Medicare-enrolled Veterans. Health services research 45 (3):762-791

[78] RDC Team (2005) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria

[79] Gueorguieva R (2001) A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. Statistical modelling 1 (3):177-193

[80] Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Statistical science:457-472

[81] Copas JB (1983) Regression, prediction and shrinkage. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 311-354

[82] VERISK Heath Inc (2011) Verisk Health DxCG medical classification system – Version 7 structural summary Available at http://www.veriskhealth.com/verisk-advantage/DxCG-Medical-Classification-V7-Structural-Summary.pdf. Accessed March 18, 2014.

ABSTRACT

AN ANALYTICS APPROACH TO DESIGNING PATIENT CENTERED MEDICAL HOME

by

SAEEDE AJORLOU

August 2014

Advisor: Dr. Kai Yang

Major: Industrial and Systems Engineering

Degree: Doctor of Philosophy

Recently the patient-centered medical home (PCMH) model has become a popular team-based approach focused on delivering more streamlined care to patients. In current practices of medical homes, a clinical-based prediction frame is recommended because it can help match the portfolio capacity of PCMH teams with the actual load generated by a set of patients. Without such balances in clinical supply and demand, issues such as excessive under and over utilization of physicians, long waiting time for receiving the appropriate treatment, and non-continuity of care will eliminate many advantages of the medical home strategy. In this research, we formulate the problem into two phases.

At the first phase we proposed a multivariate version of multilevel structured additive regression (STAR) models which involves a set of health care responses defined at the lowest level of the hierarchy, a set of patient factors to account for individual heterogeneity, and a set of higher level effects to capture heterogeneity and dependence

between patients within the same medical home team and facility. We show how a special class of such models can equivalently be represented and estimated in a structural equation-modeling framework. A Bayesian variable selection with spike and slab prior structure is then developed that allows including or dropping single effects as well as grouped coefficients representing particular model terms. We use a simple parameter expansion to improve mixing and convergence properties of Markov chain Monte Carlo simulation. A detailed analysis of the VHA medical home data is presented to demonstrate the performance and applicability of our method. In addition, by extending the hierarchical generalized linear model to include multivariate responses, we develop a clinical workload prediction model for care portfolio demands in a Bayesian framework. The model allows for heterogeneous variances and unstructured covariance matrices for nested random effects that arise through complex hierarchical care systems. We show that using a multivariate approach substantially enhances the precision of workload predictions at both primary and non-primary care levels. We also demonstrate that care demands depend not only on patient demographics but also on other utilization factors, such as length of stay. Our analyses of a recent data from Veteran Health Administration further indicate that risk adjustment for patient health conditions can considerably improve the prediction power of the model.

For the second phase, with the help of the model developed in first phase, we are able to estimate the annual workload demand portfolio for each patient with given attributes. Together with the healthcare service supply data, and based on the principles of balancing supply and demand, we developed stochastic optimization models to allocate patients to

all PCMH teams in order to make balance between supply and demand in healthcare system. We proposed different stochastic models and two solution approaches such as Progressive Hedging and L shaped Benders Decomposition. We described the application of the two mentioned algorithms and finally we compared the performance of the two methods.

AUTOBIOGRAPHICAL STATEMENT

Saeede Ajorlou was born in Tehran, Iran on August 25, 1984. She received the Bachelor of Science in Computer Software Engineering in 2007 and Master of Science degree in Industrial Engineering in 2010 from Iran University of Science and Technology. In 2011, she got admitted for Ph.D. program in the Industrial and Systems Engineering at the Wayne State University, Detroit, Michigan / USA. When she was Ph.D. student, she worked as a research assistant in Healthcare Systems Engineering Group under supervision of Dr. Kai Yang. After her graduation, she plans to continue her research as a postdoctoral fellow in department of Industrial and Operations Engineering at University of Michigan.

During her studies at Wayne State University and Iran University of Science and Technology, she made a number of technical presentations at INFORMS, IIE and several conferences and she nominated three times for best paper award in WCE and IAENG conferences. Her articles have been published and/or are under review in journals like IIE Transactions, IIE Transactions on Healthcare Systems Engineering, Healthcare Management Science, Computer and Industrial Engineering, Intelligent Manufacturing Systems and Health Service Research. She is a member of INFORMS and IIE.