# Statistical Methods for the Analysis of Observational Data With Multiple Correlated Outcomes

## Permanent link

## Terms of Use

# Share Your Story

# Statistical Methods for the Analysis of Observational Data with Multiple Correlated Outcomes

A DISSERTATION PRESENTED
BY
TIANYI CAI
TO
THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
BIOSTATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
AUGUST 2017

# Statistical Methods for the Analysis of Observational Data with Multiple Correlated Outcomes

## Abstract

In this work, we consider three problems in applied statistics motivated by complex datasets, with approaches from both Frequentist and Bayesian paradigms. Chapter 2 is motivated by case-control data collected for the Army Study to Assess Risk and Resilience in Servicemembers. We derive estimation and testing methods for data sampled by a composite indicator matched on covariates, with an added complexity of misclassified outcomes. Chapter 3 is motivated by multilevel data collected from the Consumer Assessment of Healthcare Providers and Systems surveys. We develop a spatial-temporal Bayesian random effects model with a flexible parameterization, and formulate a Bayesian hat matrix to transparently assess how information is being used in construction of the model estimates. Finally, a cross-validation approach is implemented to evaluate models. Chapter 4 is motivated by observational data from a large administrative database of Medicare beneficiaries, containing patients clustered by hospital providers. We propose a Bayesian hierarchical model to assess associations at the hospital level of the model. A case-mix adjustment is provided at the patient level, with adjustment for hospital-level confounders at the second level. A skew-$t$ distribution is used for the random effects to allow greater flexibility and to compare model adequacy.

# Table of Contents

# List of Tables

# List of Figures

To Drs. Jun Cai *&* Sufeng Xu.

# Acknowledgments

This monumental undertaking is the culmination of years of collaborative work from many gifted individuals. To my advisor Professor Tianxi Cai: thank you guiding me from day one with your enduring patience, which was indispensable in turning my first year summer project into a chapter in my thesis. When I struggled, your encyclopedic knowledge was always there to push me in the right direction. To my advisor Professor Francesca Dominici: thank you for mentoring and developing me over the last five years with your tireless dedication to excellence and benevolence. Most personally, thank you for lifting my spirits when I felt helpless.

To Professor Alan Zaslavsky: thank you for acting as a third advisor to me with your guidance on our paper. Your timely and detailed emails, boundless wisdom, and amiable demeanor made it a pleasure to learn from you. To Professor Sherri Rose, thank you for your career advice, contributions to our paper, and friendly discussions. To Professor Brent Coull, thank you for advancing my research with your insightful questions during committee meetings, and for your flexibility in accommodating my data science pursuits this summer.

I would like to thank the students, faculty, and staff of the Department of Biostatistics for fostering a cooperative and thought-provoking culture. A few fellow students deserve special mention for being part of my support group over the years. These include Yeting Du, Ryan

Sun, Sixing Chen, Kathy Evans, and Jessica Gronsbell for being there from the start, as well as Caleb Miles, Sam Tracy, and Caleb 'v' Lareau for motivating me to go to the gym.

Special thanks go out to my friends back at home for keeping me grounded and level-headed. I would like to especially thank Curtis Watson, Tomson Hecky, Zak Nitsch, Justin Ziggy Garzon, and Hish Weeraratne for devoting time to me whenever I was back in town.

My most important relationship formed during graduate school is with my dearest Sheila Gaynor. Thank you, Sheila, for your wonderful altruism, the lovely dates, proofreading my writing, and being a perfect companion to me.

Lastly, I thank my parents Jun Cai and Sufeng Xu for raising me from a little boy into the man that I am today. Their work ethic and remarkable accomplishments continue to inspire me to this day, and I would not have gone this far without their guidance and support.

*"What makes the desert beautiful," said the little prince, "is that somewhere it hides a well."*

– Antoine de Saint Exupéry

# INTRODUCTION

Modern datasets are frequently characterized by complications that do not allow analyses via standard methods. To provide accurate inference or prediction, statistical methods need to account for the complexities of the data, be it an irregular sampling scheme, correlated outcomes, an intricate hierarchical structure, or otherwise.

In Chapter 2, we study case-control data with multiple correlated binary outcomes, an unconventional sampling scheme, and misclassified outcomes. Correlated binary outcomes, such as multiple disease indicators for one patient, arise frequently in public health research. When an individual outcome of interest is rare, sampling the data via a case-control design is preferred for its higher power, but when analyzing multiple outcomes, it is more feasible to sample based on a composite outcome. Ignoring this composite sampling in the analysis produces biased estimates for the odds ratios. An additional layer of complication arises when gold standard outcome information is not available due to practical difficulties in ascertaining the outcomes. Under such settings, the outcome status is often estimated based on predicted probabilities derived from fitting a risk prediction model in a validation set using electronic medical records. Traditionally, such estimated probabilities are thresholded to classify the true outcome status, resulting in potentially misclassified outcomes. In this chapter, we dis-

cuss estimation and testing procedures to improve power by directly modeling the predicted probabilities, along with inverse probability weighting to account for the composite sampling. We show via simulation results that the proposed methods perform well in finite samples. Estimation results show a reduced bias when accounting for the composite sampling, while testing results show increased power when using the predicted probabilities. The methodology is illustrated on data from the Army Study to Assess Risk and Resilience in Servicemembers New Soldier Study.

In Chapter 3, we study sufficient statistics for survey data summarized by areas nested within states and across multiple years. Each year, surveys are conducted to assess the quality of care for Medicare beneficiaries, using instruments from the Consumer Assessment of Healthcare Providers and Systems program. In each state, depending on the heterogeneity of survey measures for Fee-for-Service beneficiaries, their results are currently presented pooled at the state level or unpooled for smaller substate areas nested within the state. We fit spatial-temporal Bayesian random effects models using a flexible parameterization to estimate mean scores for each of the domains formed by 94 areas in 32 states measured over 5 years. A Bayesian hat matrix provides a heuristic interpretation of the way the model combines information for estimates in these domains. The model can be used to choose between reporting of state- or substate-level direct estimates in each state, or as a source of alternative small area estimates superior to either direct estimate. We compare several candidate models using log pseudo-marginal likelihood and posterior predictive checks. Results from the best-performing model for 8 measures surveyed from 2012 to 2016 show substantial reductions in mean squared

error over direct estimates.

In Chapter 4, we study patient data clustered by hospitals in search of an average causal effect at the hospital level, while controlling for patient-level characteristics. Hospital profiling is a method of evaluating medical providers based on a set outcome, typically performed using mortality rates. In this paper, we are interested in extending the models used in hospital profiling to perform a hospital-level analysis to determine the relationship between palliative care received and the aggressiveness of end-of-life (EOL) treatments (e.g. chemotherapy, radiotherapy, re-admissions) for patients with advanced cancer. This analysis will also allow us to identify hospital-specific characteristics that explain the variation in EOL outcomes across hospitals. Towards these goals, we develop Bayesian hierarchical models for our dataset of 408 hospitals, including 20,400 Medicare patients with advanced lung, pancreas, colorectal, or brain cancer. At the first stage of the model, we adjust for case-mix bias using patient-level data. At the second stage, we estimate hospital-level risk of EOL outcomes, and adjust for hospital-level covariates to assess the association between hospital-level receipt of palliative care on EOL outcomes in hospitals. Characterizing between-hospital variability and determining whether receiving palliative care reduces treatment aggression is important due to the important health policy implications in the utilization of EOL care.

# Global Tests for Misclassified Outcomes in Case-Control Data with Composite Sampling

Tianyi Cai
Department of Biostatistics
Harvard University

Chia-Yen Chen
Psychiatric and Neurodevelopmental Genetics Unit
Massachusetts General Hospital

Ronald Kessler
Department of Health Care Policy
Harvard Medical School

Jordan Smoller
Psychiatric and Neurodevelopmental Genetics Unit
Massachusetts General Hospital

Murray Stein
Department of Psychiatry
University of California, San Diego

Tianxi Cai
Department of Biostatistics
Harvard University

## 2.1 Introduction

For a single binary outcome, a case-control study provides valid estimators of odds ratios, provided the outcome of interest is the variable that defines the case-control sampling (Prentice & Pyke, 1979). However, when there are multiple correlated binary outcomes, it is often more pragmatic to sample based on a composite outcome, where cases are subjects with any of the binary outcomes, and controls are subjects with none of the binary outcomes. This sampling can be matched based on some covariates. Using a composite outcome for sampling optimizes the amount of information used, which is particularly important when the cost of attaining accurate data is high. If this atypical sampling is ignored in analyzing the data, the resulting odds ratios for the multiple outcomes will be biased, because the variables being analyzed are not the same as the variables defining the case-control sampling. Problems of similar nature have been discussed in the literature. One example is the analysis of secondary phenotypes in case-control genome-wide association studies, where subjects are sampled based on a different primary outcome. Proposed solutions include inverse probability weighting (Richardson et al., 2007; Monsees et al., 2009; Schifano et al., 2013), and retrospective likelihood (Lin & Zeng, 2009).

In addition to this composite sampling issue, binary outcomes can be potentially misclassified in practice, as gold standard information is not always available due to difficulties in its ascertainment. Methods to account for outcome misclassification have been studied exten-

sively in the literature, and typically model the misclassification rates using information from a smaller validation set where the true outcomes are available (Pepe, 1992; Bollinger & David, 1997; Lyles et al., 2011; Edwards et al., 2013). Outcome misclassification in case-control studies have been covered (Jurek et al., 2013; Gilbert et al., 2014), but not with multiple outcomes and composite sampling. Among the literature on binary outcome misclassification, there have been few applications analyzing datasets that contain the probability of the outcome in addition to the misclassified binary indicator, which is a distinct advantage of electronic medical record (EMR) phenotyping (Liao et al., 2015). These probabilities are typically generated from risk prediction models in validation sets. While the probabilities can be thresholded to yield potentially misclassified outcomes, Sinnott et al. (2014) showed that directly modeling these probabilities instead of the thresholded outcomes improve effect estimation and testing power.

In this chapter, we extend the work done by Sinnott et al. (2014) to account for multiple outcomes with a composite sampling scheme. Specifically, we jointly model the probability of the multiple outcomes, use inverse probability weighting (IPW) to address the composite sampling issue, and derive marginal and global score tests to assess the association between single-nucleotide polymorphism (SNP) sets and a group of correlated outcomes. These methods yield unbiased estimates of odds ratios and improved testing efficiency over methods involving the binary misclassified outcomes. We demonstrate the efficacy of our methods in finite samples via simulation results, and apply the methods to the New Soldier Study dataset from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS).

The remainder of this chapter is organized as follows. In Section 2.2, we describe the data

structure, composite sampling, and the models used for estimation and testing. Section 2.3

provides simulation results suggesting the use of IPW to adjust for estimation bias arising from

composite sampling, and the use of predicted probabilities to improve power in testing. The

proposed procedures are applied to the New Soldier Study dataset in Section 2.4. A discussion

and concluding remarks are presented in Section 2.5.


## 2.2   Case-Control Model

Genome-wide association studies seek associations between SNPs and traits of interest such

as diseases. For large populations of interest, it may be too costly or impractical to genotype

every individual to acquire their SNPs, so subjects are typically sampled into a case-control

study and subsequently genotyped. That is, complete data including SNPs are available for the

case-control population but not the target population. Additionally, in the presence of misclas-

sified outcomes, a small validation set containing the true outcomes is needed to establish the

misclassification rates and to generate the predicted probabilities for each outcome.


### 2.2.1   Data Structure

Suppose our observable data come from a case-control design of size $n$ sampled from a target

population of size $N$, and consist of independent and identically distributed random variables

$\left[ p_{ki}, \tilde{d}_{ki}, w_i, \mathbf{s}_i^\mathsf{T}, \mathbf{x}_i^\mathsf{T} \right]^\mathsf{T}$, $k = 1, \ldots K$, $i = 1, \ldots, n$, where $K$ is the number of outcomes, $p_{ki}$ is the

predicted probability of the outcome, $\tilde{d}_{ki}$ is the binary outcome thresholded from $p_{ki}$, $w_i$ is the

7

inverse probability of being sampled into the case-control data, $\mathbf{s}_i$ is a vector of the number of risk alleles at various SNPs, and $\mathbf{x}_i$ is a vector of covariates (including intercept) for which we wish to adjust, such as age, gender, and principal components (Price et al., 2006). Let $d_{ki}$ represent the true gold standard outcome, which is not observed. For each true outcome, we assume that a standard logistic regression holds, with

$$\mathbb{P}\left(d_{ki} = 1 | \mathbf{x}_i, \mathbf{s}_i\right) = g\left(\boldsymbol{\alpha}_{0k}^\mathsf{T} \mathbf{x}_i + \boldsymbol{\beta}_{0k}^\mathsf{T} \mathbf{s}_i\right)$$

where $g\left(x\right) = e^x / \left(1 + e^x\right)$. Our goal is to make inferences on the true odds ratios $e^{\boldsymbol{\beta}_{0k}}$. Naively modeling $\mathbb{P}\left(\tilde{d}_{ki} = 1 | \mathbf{x}_i, \mathbf{s}_i\right)$ will generally yield incorrect inferences, as discussed in Section 2.2.3.

## 2.2.2 Description of Composite Sampling

In a standard case-control study, cases $\left(\tilde{d}_{ki} = 1\right)$ and controls $\left(\tilde{d}_{ki} = 0\right)$ are sampled for each outcome $k$. However, doing a separate case-control study for each outcome would lead to subjects that were genotyped for the analysis of one outcome but not necessarily used in the analysis of other outcomes. Due to the cost of genotyping, we would like for such subjects to be used as efficiently as possible. As a result, the sampling is done on a composite outcome $\bar{d}_i = \mathbb{I}\left(\sum_{k=1}^K \tilde{d}_k > 0\right) = 1 - \prod_{k=1}^K \left(1 - \tilde{d}_{ki}\right)$, where subjects with any outcome are potential cases, while subjects with no outcomes are potential controls. In this setting, all genotyped subjects are used in the analysis for all outcomes. Cases $\left(\bar{d}_i = 1\right)$ are sampled with probability

$\lambda \in (0, 1]$, and controls ($\bar{d}_i = 1$) are sampled to equal the number of cases, possibly matched on one or more covariates. The sampling indicator is $v_i = \mathbb{I}$ (subject $i$ is sampled) $\sim \text{Bern}(\pi_i)$, and we define the inverse probability weight of being sampled as $w_i = \frac{v_i}{\pi_i}$. In our setting, $w_i$ is known ahead of time, although it can be shown that estimating $w_i$ is counterintuitively more efficient (Wooldridge, 2007).

The composite sampling is one of many atypical sampling procedures that necessitate the use of IPW. Monsees et al. (2009) showed that the IPW method has no biases, valid Type I error rates, but generally larger variances of parameter estimates across all scenarios. They provided some qualitative intuition behind how IPW address the sampling bias. The theoretical reasoning behind using IPW to correct this sampling problem has been explained by viewing the case-control study as a two-stage design (Reilly & Pepe, 1995; Siegmund et al., 1999), or a missing data problem (Robins et al., 1995; Robins & Rotnitzky, 1995; Wacholder, 1996).

### 2.2.3 Justification for Using Predicted Probabilities

Suppose the true outcome status $d_k$ is known for a small validation set, from which we are able to construct an algorithm that uses electronic medical records to predict the probability $p_{ki} = \mathbb{P}(d_{ki} = 1|\mathbf{u}_{ki})$ for each subject, for some variables $\mathbf{u}_{ki}$ determined by the algorithm. The $p_{ki}$ are assumed to be generated such that each $p_{ki}$ is conditionally independent of covariates

$\mathbf{x}_i$ and the SNPs $\mathbf{z}_i$, given the true binary outcome $d_{ki}$. That is, we assume

$$p_{ki} \perp \mathbf{x}_i, \mathbf{s}_i | d_{ki} \tag{$*$}$$

The $\tilde{d}_{ki}$ are then obtained by thresholding the $p_{ki}$ at a cutoff $c_k$ satisfying certain sensitivity $\text{se}_k = \mathbb{P}\left(\tilde{d}_{ki} = 1 | d_{ki} = 1\right)$ or specificity $\text{sp}_k = \mathbb{P}\left(\tilde{d}_{ki} = 0 | d_{ki} = 0\right)$ conditions. That is, $\tilde{d}_{ki} = \mathbb{I}\left(p_{ki} > c_k\right)$, where $c_k$ is determined in the validation set. To prioritize a high positive predictive value (PPV), we set a high specificity, as $\text{PPV}_{ki} = \frac{\text{se}_k}{\text{se}_k + (1 - \text{sp}_k)\frac{\mathbb{P}(d_{ki}=0)}{\mathbb{P}(d_{ki}=1)}}$. We see that PPV increases with specificity, even if sensitivity is low.

Ultimately, we are interested in the association between SNPs indicated by $\mathbf{s}_i$ and the true outcome statuses $d_k$, which are not observed. One method (Kurreeman et al., 2011) would be to fit a standard logistic regression model using $\tilde{d}_k$, or

$$\mathbb{P}\left(\tilde{d}_{ki} = 1 | \mathbf{x}_i, \mathbf{s}_i\right) = g\left(\boldsymbol{\alpha}_{0k}^{*\mathsf{T}}\mathbf{x}_i + \boldsymbol{\beta}_{0k}^{*\mathsf{T}}\mathbf{s}_i\right)$$

However, Magder & Hughes (1997) showed that under assumption ($*$), $\left(\boldsymbol{\alpha}_{0k}^*, \boldsymbol{\beta}_{0k}^*\right)$ do not generally equal the true $(\boldsymbol{\alpha}_{0k}, \boldsymbol{\beta}_{0k})$. Neuhaus (1999) provides an adjustment for this by noting that

$$\mathbb{E}\left(\frac{\tilde{d}_{ki} - 1 + \text{sp}_k}{\text{se}_k - 1 + \text{sp}_k}\bigg| \mathbf{x}_i, \mathbf{s}_i\right) = g\left(\boldsymbol{\alpha}_{0k}^{\mathsf{T}}\mathbf{x}_i + \boldsymbol{\beta}_{0k}^{\mathsf{T}}\mathbf{s}_i\right) = \mathbb{P}\left(d_{ki} = 1 | \mathbf{x}_i, \mathbf{s}_i\right) \tag{2.1}$$

That is, for estimation, we are able to use the outcome $d_k^* = \frac{\tilde{d}_{ki} - 1 + \text{sp}_k}{\text{se}_k - 1 + \text{sp}_k}$ in place of $d_k$ and get unbiased estimators for $\boldsymbol{\alpha}_{0k}$ and $\boldsymbol{\beta}_{0k}$.

Similarly, by modeling $p_k$ directly, we can show that, under assumption ($*$),

$$\mathbb{E}\left(\left.\frac{p_{ki} - \mu_{kp0}}{\mu_{kp1} - \mu_{kp0}}\right| \mathbf{x}_i, \mathbf{s}_i\right) = \mathbb{P}\left(d_{ki} = 1 | \mathbf{x}_i, \mathbf{s}_i\right)$$

Details are provided in Appendix A. Here, $\mu_{kp1} = \mathbb{E}\left(p_k | d_k = 1\right)$ and $\mu_{kp0} = \mathbb{E}\left(p_k | d_k = 0\right)$ are derived from the validation set. Now let

$$y_{ki} = \frac{p_{ki} - \mu_{kp0}}{\mu_{kp1} - \mu_{kp0}}, \, \mathbb{E}\left(y_{ki} | \mathbf{x}_i, \mathbf{s}_i\right) = \mathbb{P}\left(d_k = 1 | \mathbf{x}_i, \mathbf{s}_i\right) \tag{2.2}$$

We are interested in estimation and testing using the transformed probabilities $y_{ki}$ instead of the misclassified $\tilde{d}_{ki}$ or adjusted $d_{ki}^*$. The predicted probabilities $p_{ki}$ provide more information about the accuracy of the outcome classification than $\tilde{d}_{ki}$, as a $p_{ki}$ high above the threshold has more certain outcome status than a $p_{ki}$ barely above the threshold.

## 2.2.4 Inverse Probability Weighted Estimator

Throughout this section, our outcome will be $y_{ki}$ as defined in (2.2). For simplicity, let $\theta_k = (\alpha_k, \beta_k)$ denote the full set of coefficients and $\mathbf{z}_i = (1, \mathbf{x}_i^\top, \mathbf{s}_i^\top)^\top$ denote the full set of regressors so that $\mathbb{E}\left(y_{ki} = 1 | \mathbf{z}_i\right) = g\left(\theta_{0k}^\top \mathbf{z}_i\right)$. A weighted logistic regression for each outcome $k$ corresponds to solving the set of estimating equations

$$\Psi_n\left(\theta_k\right) = n^{-1}\sum_{i=1}^{n} w_i \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\top \mathbf{z}_i\right)\right] = 0 \tag{2.3}$$

Let $\hat{\boldsymbol{\theta}}_k$ be the solution to $\boldsymbol{\Psi}_n(\boldsymbol{\theta}_k) = 0$, where $\hat{\boldsymbol{\theta}}_k \overset{p}{\to} \boldsymbol{\theta}_{0k}$ (Van der Vaart, 2000). Let $\boldsymbol{\psi}_{\boldsymbol{\theta}_{0k}} = w_i \mathbf{z}_i [y_{ki} - g(\boldsymbol{\theta}_{0k}^\mathsf{T} \mathbf{z}_i)]$ so that $\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}} = w_i \mathbf{z}_i \mathbf{z}_i^\mathsf{T} g(\boldsymbol{\theta}_{0k}^\mathsf{T} \mathbf{z}_i) [1 - g(\boldsymbol{\theta}_{0k}^\mathsf{T} \mathbf{z}_i)]$. The asymptotic distribution of $\hat{\boldsymbol{\theta}}_k$ is then (Van der Vaart, 2000)

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k\right) \overset{d}{\to} \mathcal{N}\left(0, \left[\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)\right]^{-1} \mathbb{E}\left(\boldsymbol{\psi}_{\boldsymbol{\theta}_{0k}} \boldsymbol{\psi}_{\boldsymbol{\theta}_{0k}}^\mathsf{T}\right) \left[\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)\right]^{-1}\right)$$

Details are provided in Appendix B. Estimation of $\boldsymbol{\theta}_{0k}$ involves directly solving the estimating equation (2.3) with a Newton-Raphson algorithm, as existing software for logistic regression require the outcome to be between 0 and 1, which is not satisfied when transforming the $p_k$ into $y_k$.

For comparison, we would also expect $\hat{\boldsymbol{\theta}}_k \overset{p}{\to} \boldsymbol{\theta}_{0k}$ if we replace $y_{ki}$ in (2.3) with $d_{ki}^*$, as seen in (2.1). However, estimation bias occurs when using $\tilde{d}_{ki}$ as the outcome, since $\boldsymbol{\theta}_{0k}$ in general is not the solution to $\mathbb{E}\left(w_i \mathbf{z}_i \left[\tilde{d}_{ki} - g(\boldsymbol{\theta}_k^\mathsf{T} \mathbf{z}_i)\right]\right) = 0$.

### 2.2.5   Score Test Statistics

For a given SNP set $\mathbf{s}$, we are interested in the association between $\mathbf{s}$ and a particular outcome, or $H_0 : \boldsymbol{\beta}_k = 0$. In the presence of multiple correlated outcomes, we may be interested in the association between $\mathbf{s}$ and all the outcomes, in which case a marginal test will not suffice. In that case, we would like the global test $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_K = 0$.

We first develop a marginal score test $H_0 : \boldsymbol{\beta}_k = 0$ for each outcome, based on the inverse

12

probability weighted estimating equation (2.3). Let

$$\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) = n^{-1}\sum_{i=1}^{n} w_i\mathbf{s}_i\left[y_{ki} - g\left(\tilde{\boldsymbol{\alpha}}_k^\mathsf{T}\mathbf{x}_i\right)\right]$$

where $\tilde{\boldsymbol{\alpha}}_k$ is an estimate for $\boldsymbol{\alpha}_k$ under $H_0$. Then (Appendix C)

$$\sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \overset{d}{\to} \mathcal{N}\left(0, \Sigma_k = \mathbb{E}\left[\boldsymbol{\xi}_{ki}\boldsymbol{\xi}_{ki}^\mathsf{T}\right]\right)$$

where

$$\boldsymbol{\xi}_{ki} = w_i\left(\mathbf{s}_i - \mathbb{E}\left[w_i\mathbf{s}_i\mathbf{x}_i^\mathsf{T}g'\left(\boldsymbol{\alpha}_k^\mathsf{T}\mathbf{x}_i\right)\right]\left(\mathbb{E}\left[w_i\mathbf{x}_i\mathbf{x}_i^\mathsf{T}g'\left(\boldsymbol{\alpha}_k^\mathsf{T}\mathbf{x}_i\right)\right]\right)^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_k^\mathsf{T}\mathbf{x}_i\right)\right]$$

Two test statistics to consider are

$$\begin{aligned}
\mathbf{M}_1 &= n\hat{\mathbf{S}}_k^\mathsf{T}\left(\tilde{\boldsymbol{\alpha}}_k\right)\Sigma_k^{-1}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \sim \chi_m^2 \\
\mathbf{M}_2 &= n\hat{\mathbf{S}}_k^\mathsf{T}\left(\tilde{\boldsymbol{\alpha}}_k\right)\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \sim \sum_{i=1}^{m}\lambda_i\left(\Sigma_k\right)\chi_1^2
\end{aligned}$$

where $m = \dim(\mathbf{s})$ and $\lambda_i(\Sigma_k)$ are the non-zero eigenvalues of $\Sigma_k$. $\mathbf{M}_1$ is a standard score test and has a simpler distribution, but in the presence of highly correlated SNPs, $\Sigma_k$ may be singular, in which case $\mathbf{M}_2$ is preferred. Simulations show that $\mathbf{M}_2$ generally has higher power than $\mathbf{M}_1$, so we report results for $\mathbf{M}_2$ only.

13

For global tests $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_K = 0$, we have

$$\sqrt{n}\vec{S}\left(\tilde{\boldsymbol{\alpha}}\right) := \sqrt{n}\left[\hat{S}_1^{\mathsf{T}}\left(\tilde{\boldsymbol{\alpha}}_1\right), \ldots, \hat{S}_K^{\mathsf{T}}\left(\tilde{\boldsymbol{\alpha}}_K\right)\right]^{\mathsf{T}}$$

where $\tilde{\boldsymbol{\alpha}}_k$ are estimates for $\boldsymbol{\alpha}_k$ under $H_0$. Then (Appendix D)

$$\sqrt{n}\vec{S}\left(\tilde{\boldsymbol{\alpha}}\right) \xrightarrow{d} \mathcal{N}\left(0, \vec{\Sigma} = \mathbb{E}\left(\vec{\xi}_i \vec{\xi}_i^{\mathsf{T}}\right)\right)$$

where $\vec{\xi}_i = [\xi_{1i}^{\mathsf{T}}, \ldots, \xi_{Ki}^{\mathsf{T}}]^{\mathsf{T}}$. We once again consider two test statistics

$$
\begin{aligned}
\mathbf{G}_1 &= n\vec{S}^{\mathsf{T}}\left(\tilde{\boldsymbol{\alpha}}\right)\vec{\Sigma}^{-1}\vec{S}\left(\tilde{\boldsymbol{\alpha}}\right) \sim \chi^2_{mK} \\
\mathbf{G}_2 &= n\vec{S}^{\mathsf{T}}\left(\tilde{\boldsymbol{\alpha}}\right)\vec{S}\left(\tilde{\boldsymbol{\alpha}}\right) \sim \sum_{i=1}^{mK} \lambda_i\left(\vec{\Sigma}\right)\chi^2_1
\end{aligned}
$$

where $\lambda_i\left(\vec{\Sigma}\right)$ are the non-zero eigenvalues of $\vec{\Sigma}$. Once again, we report results for $\mathbf{G}_2$ only, due to its higher power in simulations (results not shown).

## 2.3 Simulation Study

Simulation studies are performed to evaluate the performance of the proposed methods in finite samples. We consider two main tasks: estimating $\boldsymbol{\beta}_k$, and performing the marginal test $H_0 : \boldsymbol{\beta}_k = 0$ or global test $H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K = 0$. Comparisons are made based on the presence or absence of IPW, and across outcomes $d_k$ (gold standard), $\tilde{d}_k$ (binary thresholded

14

probabilities), $d_k^*$ (adjusted $\tilde{d}_k$), and $y_k$ (transformed probabilities).

Using R, 2000 datasets are generated for each empirical calculation. For each simulated dataset, we generate genetics, covariates, and outcomes for $N = 40,000$ subjects. SNPs from the ASAH1 gene on chromosome 8 are generated using HAPGEN2 software (Su et al., 2011), and a single covariate indicating the quartile of childhood adversity is generated. Composite sampling matched on childhood adversity is performed on the $40,000$ subjects to generate the case-control data. Four disorders are generated, with prevalences of 0.20, 0.15, 0.10, and 0.05, respectively. For each disorder, we generate risk scores $u_k$ from a $\mathcal{N}\left(\mu_0 + \mu_1 d_k, \sigma_k^2\right)$ distribution, where $\mu_1$ is set to achieve a desired Area Under the Receiver Operating Characteristic Curve (AUC) when using $p_k$ to predict $d_k$ in the validation set of size 400.

Due to the small size of the validation set, the performance of $d_k^*$ and $y_k$ depend heavily on the reliability of information gathered from the validation set. In our referenced simulation results, a new validation set is generated for each of the 2000 simulated datasets, which allows the variability of the validation set to have minimal effect on the results we observe. By contrast, when we generate one validation set for all 2000 simulated datasets, the results are more variable, although the major takeaways are still evident.

For estimation, we expect IPW estimating equations with outcome $d_k^*$ and $y_k$ to provide unbiased estimates of $\boldsymbol{\beta}$. Bias is expected when using $\tilde{d}_k$ for estimation, regardless of the use of IPW. For testing, it is of interest to compare the power of our score tests based on the IPW score equations across outcomes $\tilde{d}_k$, $d_k^*$, and $y_k$. For the marginal tests, comparisons are also made against the SNP-set Sequence Kernel Association Test (SKAT) (Wu et al., 2011) without

15

any IPW.

## 2.3.1 Estimation

To demonstrate the effect of the IPW with minimal complexity, estimation is performed using

1 simulated SNP (rs17515264 from ASAH1 on chromosome 8) with a minor allele frequency

of 0.15. For each of the four disorders, the true $\beta_k = 0.4$, corresponding to an odds ratio

of about 1.5. For $d_k$ and $\tilde{d}_k$, standard logistic regression is used, while the coefficients for $y_k$

and $d_k^*$ are solved using a Newton-Raphson algorithm. For the validation set, we consider

specificity levels of 0.95 and 0.97, and AUCs of 0.90 and 0.94. Results of the estimation for

these configurations are presented in Figure 3.1.
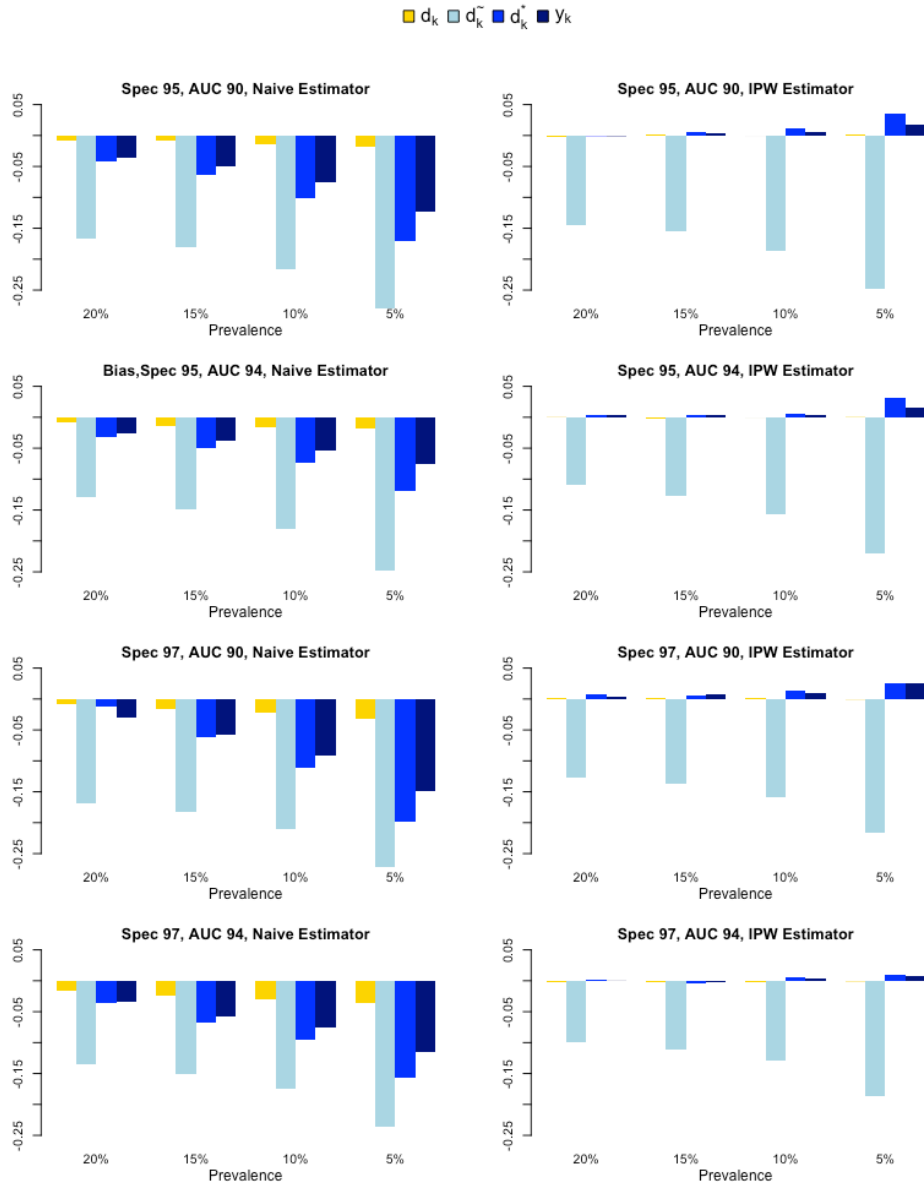
Within each plot, each group of four bars represents a separate disorder with prevalence indicated below. The four bars represent the four outcomes used in estimation of the bias. $d_k$ is the gold standard true outcome, $\tilde{d}_k$ is the thresholded probability, $d_k^*$ is the adjusted $\tilde{d}_k$, and $y_k$ is the transformed probability.

17

Generally, the gold standard $d_k$ has the lowest bias across all scenarios, while the misclassified $\tilde{d}_k$ has the highest bias. Using the transformed probabilities $y_k$ or the adjusted $d_k^*$ instead of $\tilde{d}_k$ reduced some of the bias, with $y_k$ slightly outperforming $d_k^*$. When using the naive estimator, we observe attenuation of effect estimates as indicated by a general negative bias that increases as the disorder gets rarer. For $d_k$, $d_k^*$, and $y_k$, this bias is removed through the IPW estimator, with the exception of some noise in the rarest (5% prevalence) disorder. For the misclassified $\tilde{d}_k$, this bias is slightly reduced through the IPW estimator.

As expected, changes in AUC do not affect the gold standard $d_k$, but a higher AUC improves the estimation accuracy for $\tilde{d}_k$, $d_k^*$, and $y_k$ for both naive and IPW estimators. We see virtually no bias in the IPW estimators for $d_k^*$ and $y_k$ when the AUC is 0.94 and the disorder prevalence is 10% or greater. Additionally, a higher specificity (0.97 instead of 0.95) appears to increase the bias in many scenarios. However, this effect is not significant when using the IPW estimator with a high AUC.

### 2.3.2    Testing

All tests are verified to have valid size (simulations not shown). Using 9 tagged SNPs from the ASAH1 gene on chromosome 8, testing is performed on four disorders of interest with prevalences of 0.20, 0.15, 0.10, and 0.05, respectively. Once again, for the validation set, we consider specificity levels of 0.95 and 0.97, and AUCs of 0.90 and 0.94. Comparisons are made using the score tests with $d_k$, $\tilde{d}_k$, $d_k^*$, and $y_k$ as the outcome. For an additional comparison with

18

the marginal tests, we compute the power from SKAT on the binary outcomes $d_k$ and $\tilde{d}_k$. At the time of this analysis, there are no SKAT packages incorporating IPW or multiple outcomes.

### 2.3.3   Marginal Tests

Due to the high correlation of the SNPs, all tests are done based on the test statistic

$$\mathbf{M}_2 = n\hat{\mathbf{S}}_k^\intercal\left(\tilde{\boldsymbol{\alpha}}_k\right)\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \sim \sum_{i=1}^{m} \lambda_i \chi_1^2$$

In order to illustrate the differences between the various methods, we did not want the strength of association $\boldsymbol{\beta}_k$ to be overwhelmingly large or small. As a result, two causal SNPs are selected, with log odds ratio 0.1 or 0.2 for each disorder, as shown in Figure 3.2. In all cases, when using the score test, the gold standard $d_k$ has the highest power, while the misclassified $\tilde{d}_k$ has the lowest power. Using the transformed probabilities $y_k$ or the adjusted $d_k^*$ increases power, with $y_k$ outperforming $d_k^*$. While $d_k^*$ reduced estimation bias compared to $\tilde{d}_k$, we see that both outcomes have similar power when used in testing. SKAT does not incorporate the composite sampling, so its power is much lower power in all scenarios, with SKAT using $d_k$ having higher power than SKAT using $\tilde{d}_k$. As the disorders get rarer, we see a general decrease in power, but no change in the rankings of the outcomes.

19

**Figure 2.2:** Power from Marginal Tests
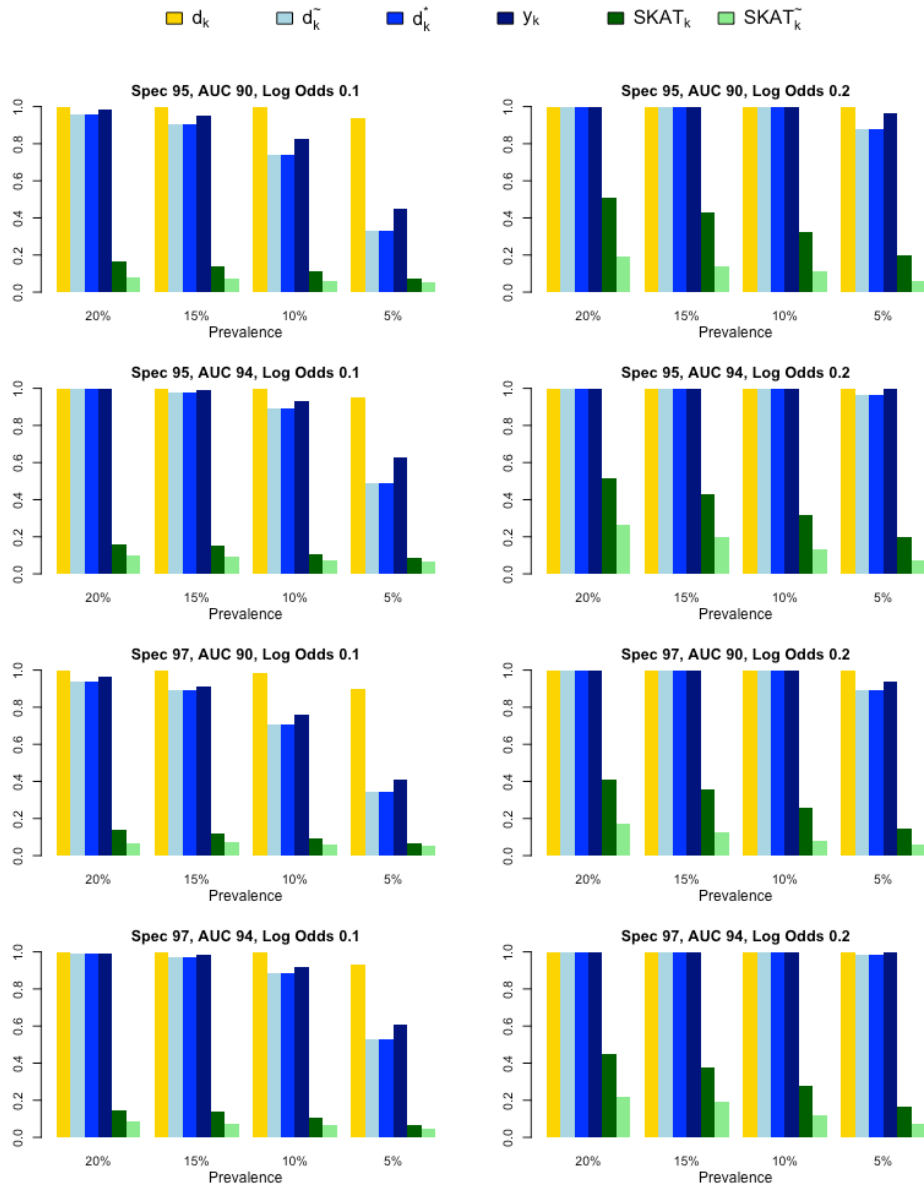


Within each plot, each group of four bars represents a separate disorder with prevalence indicated below. The first four of the six bars represent the four outcomes used in the power calculations. $d_k$ is the gold standard true outcome, $\tilde{d}_k$ is the thresholded probability, $d_k^*$ is the adjusted $\tilde{d}_k$, and $y_k$ is the transformed probability. The last two of the six bars represent the power from SKAT using $d_k$ and $\tilde{d}_k$.

20

Changes in AUC do not affect the gold standard $d_k$, but a higher AUC improves the testing

power for $\tilde{d}_k$, $d_k^*$, and $y_k$. There does not appear to be significant changes in power when using

a specificity of 0.97 compared to a specificity of 0.95.

## 2.3.4 Global Tests

Due to the high correlation of the SNPs, all tests are done based on the test statistic

$$
\mathbf{G}_2 = n\vec{\mathbf{S}}^{\mathsf{T}}(\tilde{\boldsymbol{\alpha}})\,\vec{\mathbf{S}}(\tilde{\boldsymbol{\alpha}}) \sim \sum_{i=1}^{mK} \lambda_i \chi_1^2
$$

Once again, two causal SNPs are selected with log odds ratio 0.2, as shown in Figure 3.3. Strong,

mild, and weak associations are determined based on the number of outcomes associated with

the causal SNPs, as well as the strength of the association. We see a similar pattern in the

performance of the outcomes as in the marginal tests. In all cases, the gold standard $d_k$ has

higher power than $y_k$, which has higher power than $d_k^*$ and $\tilde{d}_k$. While $d_k^*$ reduced estimation

bias compared to $\tilde{d}_k$, we see that in this case $\tilde{d}_k$ has higher power than $d_k^*$ across all scenarios.

The changes in AUC and specificity have similar effects in the global tests as in the marginal

tests.

**Figure 2.3:** Power from Global Tests
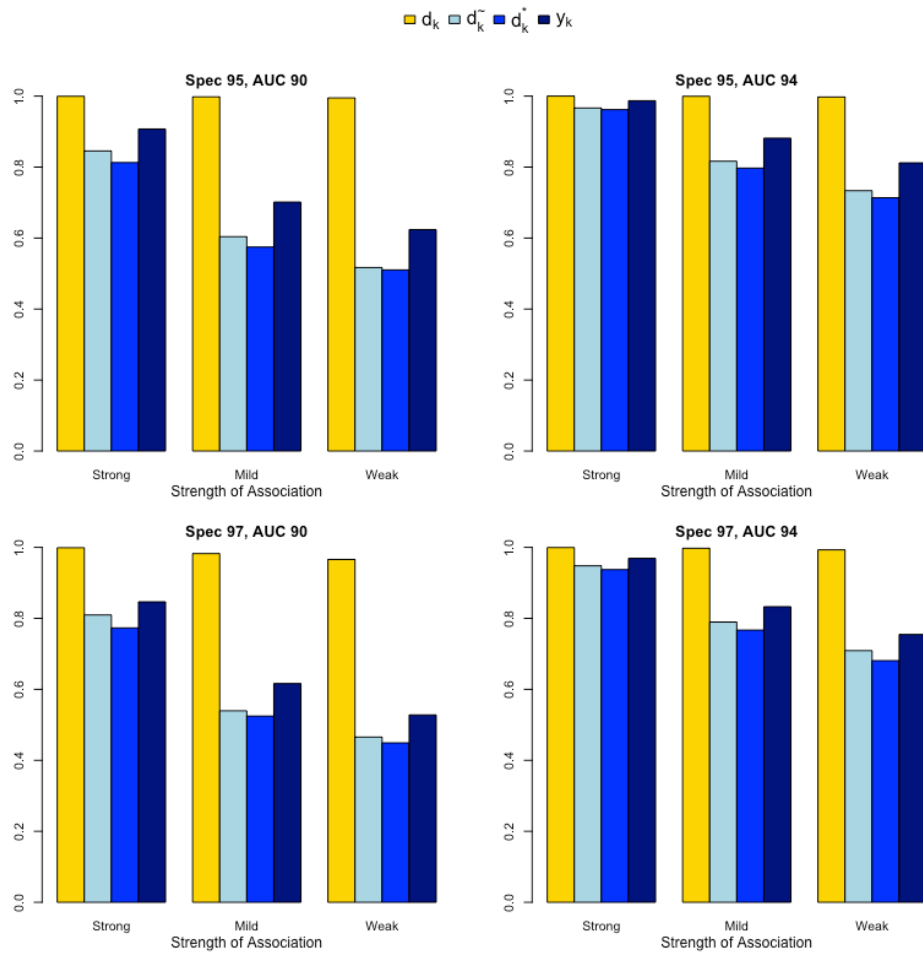


Within each plot, each group of four bars represents a separate disorder with prevalence indicated below. The four bars represent the four outcomes used in the power calculations. $d_k$ is the gold standard true outcome, $\tilde{d}_k$ is the thresholded probability, $d_k^*$ is the adjusted $\tilde{d}_k$, and $y_k$ is the transformed probability.

## 2.4 Data Example: Army STARRS New Soldier Study

Both composite sampling and outcome misclassification are issues in the data from the Army Study To Assess Risk and Resilience in Servicemembers (Army STARRS), which is the largest study of mental health ever conducted among military personnel. The multi-component study hopes to address many issues centered around suicide prevention. Formed in 2009 by the US Army in partnership with the National Institute of Mental Health, Army STARRS was motivated by the increasing suicide rate among soldiers, which surpassed the suicide rate among civilians with similar demographics in the late 2000s (Kuehn, 2009). Since the study's inception, the research team has worked at 75 locations worldwide, collected data from over 100,000 soldiers, and published a variety of papers on topics such as suicide risk factors (Schoenbaum et al., 2014), prevalence of mental illnesses (Nock et al., 2014), and the study design (Kessler et al., 2013).

Within Army STARRS is a genome-wide association study for multiple correlated psychiatric disorders, whose goal is to assess how each of the disorders are related to genetic variants denoted as SNPs. Towards these goals, soldiers were subsampled into the New Soldier Study (NSS) for genotyping based on composite sampling of having any DSM-IV disorders of interest, including major depressive disorder, generalized anxiety disorder, panic disorder, PTSD, suicide attempt, and other deliberate self-harm. The binary DSM-IV disorders are based on thresholded probabilities using electronic medical records. We are also able to directly model

these probabilities by making the transformation discussed in Section 2.2.3. At the time of this analysis, the predicted probability is available only for PTSD, so we are unable to apply our methods to multiple outcomes. The New Soldier Study is broken down into NSS1 and NSS2, based on genotyping method and the timeline of data collection.

We perform joint analyses to test the association of a SNP set and PTSD, based on the top 20 variants in the European-American meta-analysis by Duncan et al. (2017). We also perform a univariate test using the genetic risk score (Dudbridge, 2013) calculated from the aforementioned SNP set. As a basis for comparison, unweighted analysis using binary thresholded PTSD outcomes is performed. Rare variants with minor allele frequency less than 0.1 are removed, along with SNPs in high LD, yielding 5 SNPs from Duncan et al. (2017). The results are summarized in Tables 2.1 and 2.2.

**Table 2.1:** Estimation and Testing Results from the NSS

| SNP | chr | Unweighted, Binary PTSD | | | Weighted, Probability PTSD | | |
|---|---|---|---|---|---|---|---|
| | | NSS1 OR (SE) | NSS2 OR (SE) | MA-$p$ | NSS1 OR (SE) | NSS2 OR (SE) | MA-$p$ |
| chr8_125827954_I | 8 | 1.13 (0.067) | 0.94 (0.107) | 0.042 | 1.10 (0.104) | 0.93 (0.224) | 0.266 |
| rs7400289 | 13 | 1.05 (0.070) | 1.16 (0.120) | 0.091 | 1.17 (0.105) | 1.52 (0.229) | 0.008 |
| chr4_154022605_I | 4 | 1.10 (0.057) | 1.05 (0.091) | 0.053 | 1.12 (0.086) | 0.86 (0.186) | 0.077 |
| chr4_63080381_D | 4 | 0.95 (0.057) | 1.07 (0.089) | 0.164 | 0.95 (0.087) | 0.90 (0.166) | 0.270 |
| rs577266 | 9 | 0.97 (0.053) | 0.92 (0.082) | 0.187 | 0.91 (0.081) | 0.94 (0.167) | 0.168 |
| Global $p$-value | | 0.198 | 0.601 | 0.119 | 0.273 | 0.591 | 0.162 |
| GRS | | 0.278 | 0.878 | 0.244 | 0.188 | 0.080 | 0.015 |

Estimated odds ratios (OR) and log odds ratio standard errors (SE) for the 5 SNPs from Duncan et al. (2017). The global $p$-value is for the joint significance of the entire SNP set shown, while the GRS $p$-value is for the significance of the genetic risk score calculated from the entire SNP set.

**Table 2.2:** Estimates and Standard Errors of Principal Components

| PC | Unweighted, Binary PTSD | | Weighted, Probability PTSD | |
| --- | --- | --- | --- | --- |
| | NSS1 log OR (SE) | NSS2 log OR (SE) | NSS1 log OR (SE) | NSS2 log OR (SE) |
| 1 | 1.99 (3.87) | -6.26 (6.01) | 1.51 (5.75) | -6.99 (11.45) |
| 2 | 6.14 (3.82) | -2.63 (5.73) | 2.13 (6.00) | -4.86 (11.39) |
| 3 | 0.51 (3.82) | -4.75 (5.94) | -2.67 (5.91) | 1.55 (12.07) |
| 4 | -4.21 (3.83) | 1.16 (6.27) | -8.15 (5.86) | -1.85 (12.42) |
| 5 | -2.48 (3.86) | 7.64 (6.14) | -4.32 (5.85) | 10.40 (12.51) |
| 6 | -4.86 (3.78) | -1.05 (6.02) | -6.97 (5.74) | -15.64 (12.20) |
| 7 | 1.99 (3.86) | 6.63 (5.84) | -2.15 (6.11) | -2.24 (11.77) |
| 8 | -1.71 (3.84) | 1.62 (5.91) | -3.20 (5.80) | 8.66 (11.68) |
| 9 | 8.87 (3.85) | 9.67 (6.15) | 13.26 (5.91) | 28.33 (12.67) |
| 10 | -3.75 (3.84) | 1.49 (5.78) | -1.36 (6.00) | -4.66 (11.39) |

Estimated log odds ratios (OR) and log odds ratio standard errors (SE) for the first 10 principal components.

The joint analysis and genetic risk score analysis show no significant global association between these SNPs and PTSD as a binary or probability outcome. This is not surprising, as Duncan et al. (2017) performed meta-analyses of their data with the NSS data (Stein et al., 2016), only to find no genome-wide significant associations. For estimation, the weighted analysis

did not appear to impact the odds ratio estimates, indicating weights that are potentially inconsistent with the sampling design. As expected, the standard errors are higher in the weighted probability estimator.

## 2.5 Discussion

In this chapter, we propose estimation and testing procedures for data with composite sampling and misclassified outcomes. In particular, we incorporate inverse probability weighting to adjust for the atypical sampling, and directly model the probability of outcome to improve estimation accuracy and testing power. These methods are applied to the Army STARRS New Soldier Study, globally testing the SNP set comprised of SNPs proposed by Duncan et al. (2017). The lack of global significance across our tests generally corroborate the most recent literature by Duncan et al. (2017) of no genome-wide significant single variants.

Numerical studies suggest that our methods contain the least bias and highest power when compared to naive or existing methods. Estimation using the predicted probabilities produces lower bias than logistic regression with the misclassified outcomes, and estimation using the adjusted binary outcomes. Across all outcomes of interest, our predicted probabilities demonstrated the highest power in testing. Our methods also demonstrate higher power and flexibility than SKAT, as the latter incorporates neither IPW nor multiple outcomes. In practical applications, the performance of our methods depend on the reliability of estimates produced from the validation set. A validation set with a low AUC or too high or low specificity will

induce bias and reduce testing power of our proposed methods, but this is true for all methods wishing to account for misclassified outcomes.

We can observe further gains in power by augmenting our existing estimators and score test statistics with some $\hat{\boldsymbol{\epsilon}} = n^{-1} \sum_{i=1}^{n} (1 - w_i)\, \varphi_i(\mathbf{z}_i) \xrightarrow{p} 0$, where $\hat{\boldsymbol{\epsilon}}$ is a function of auxiliary covariates $\mathbf{z}_i$ related to the SNPs. Improving the efficiency of our estimators and test statistics in this case amounts to selecting an optimal $\varphi_i$. While simulations have shown that this augmentation improves efficiency, there is typically no set way in practice to find the best auxiliary variables $\mathbf{z}_i$ for this task. Details for this augmentation step are provided in Appendix E.

## 2.6  Appendix A: Expectation of $p_k$

Under the assumption $p_{ki} \perp \mathbf{x}_i, \mathbf{s}_i | d_{ki}$, we can simplify

$$
\begin{aligned}
\mathbb{E}\left(p_k | \mathbf{x}, \mathbf{s}\right) &= \int_0^1 \mathbb{P}\left(p_k > \gamma_k | \mathbf{x}, \mathbf{s}\right) d\gamma_k \\
&= \int_0^1 \mathbb{P}\left(p_k > \gamma_k | \mathbf{x}, \mathbf{s}, d_k = 1\right) \mathbb{P}\left(d_k = 1 | \mathbf{x}, \mathbf{s}\right) d\gamma_k \\
&\quad + \int_0^1 \mathbb{P}\left(p_k \leq \gamma_k | \mathbf{x}, \mathbf{s}, d_k = 0\right) \mathbb{P}\left(d_k = 0 | \mathbf{x}, \mathbf{s}\right) d\gamma_k \\
&= \int_0^1 \mathbb{P}\left(p_k > \gamma_k | d_k = 1\right) \mathbb{P}\left(d_k = 1 | \mathbf{x}, \mathbf{s}\right) d\gamma_k \\
&\quad + \int_0^1 \mathbb{P}\left(p_k \leq \gamma_k | d_k = 0\right) \mathbb{P}\left(d_k = 0 | \mathbf{x}, \mathbf{s}\right) d\gamma_k \\
&= \mathbb{E}\left(p_k | d_k = 1\right) \frac{\exp\left(\boldsymbol{\alpha}_{k0}^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_{k0}^{\mathsf{T}}\mathbf{s}\right)}{1 + \exp\left(\boldsymbol{\alpha}_{k0}^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_{k0}^{\mathsf{T}}\mathbf{s}\right)} \\
&\quad + \mathbb{E}\left(p_k | d_k = 0\right) \frac{1}{1 + \exp\left(\boldsymbol{\alpha}_{k0}^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_{k0}^{\mathsf{T}}\mathbf{s}\right)}
\end{aligned}
$$

Let $\mu_{kp1} = \mathbb{E}\left(p_k | d_k = 1\right)$ and $\mu_{kp0} = \mathbb{E}\left(p_k | d_k = 0\right)$ so that

$$
\mathbb{E}\left(\left.\frac{p_k - \mu_{kp0}}{\mu_{kp1} - \mu_{kp0}}\right| \mathbf{x}, \mathbf{s}\right) = \frac{\exp\left(\boldsymbol{\alpha}_{k0}^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_{k0}^{\mathsf{T}}\mathbf{s}\right)}{1 + \exp\left(\boldsymbol{\alpha}_{k0}^{\mathsf{T}}\mathbf{x} + \boldsymbol{\beta}_{k0}^{\mathsf{T}}\mathbf{s}\right)} = \mathbb{P}\left(d_k = 1 | \mathbf{x}, \mathbf{s}\right)
$$

## 2.7 Appendix B: Asymptotic Distribution of $\hat{\theta}_k$

In the case-control study, $\hat{\theta}_k$ is the solution to

$$\Psi_n\left(\theta_k\right) = n^{-1}\sum_{i=1}^{n} w_i \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\mathsf{T} \mathbf{z}_i\right)\right] = 0$$

where $w_i = \frac{v_i}{\pi_i}$ and $\hat{\theta}_k \xrightarrow{p} \theta_{k0}$ under regularity conditions. Then a Taylor expansion of $\Psi_n\left(\theta_k\right)$ around the true $\theta_{k0}$ gives

$$0 = \Psi_n\left(\theta_{k0}\right) + \left.\frac{\partial \Psi_n\left(\theta_k\right)}{\partial \theta_k^\mathsf{T}}\right|_{\theta_k = \theta_{k*}}\left(\hat{\theta}_k - \theta_{k0}\right)$$

for some $\theta_{k*}$ satisfying $\left\|\theta_{k*} - \theta_{k0}\right\| \leq \left\|\hat{\theta}_k - \theta_{k0}\right\|$ so that $\theta_{k*} - \theta_{k0} = o_p\left(1\right)$. Let $\psi_{\theta_{0k}} = w_i \mathbf{z}_i \left[y_{ki} - g\left(\theta_{0k}^\mathsf{T} \mathbf{z}_i\right)\right]$ so that $\psi'_{\theta_{0k}} = w_i \mathbf{z}_i \mathbf{z}_i^\mathsf{T} g\left(\theta_{0k}^\mathsf{T} \mathbf{z}_i\right)\left[1 - g\left(\theta_{0k}^\mathsf{T} \mathbf{z}_i\right)\right]$. Then, under regularity conditions,

$$\left.\frac{\partial \Psi_n\left(\theta_k\right)}{\partial \theta_k^\mathsf{T}}\right|_{\theta_k = \theta_{k*}} \xrightarrow{p} -\mathbb{E}\left(\psi'_{\theta_{0k}}\right)$$

and

$$\sqrt{n}\left(\hat{\theta}_k - \theta_{k0}\right) = \mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1}\sqrt{n}n^{-1}\sum_{i=1}^{n} w_i \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\mathsf{T} \mathbf{z}_i\right)\right]$$

$$\xrightarrow{d} \mathcal{N}\left(0, \left[\mathbb{E}\left(\psi'_{\theta_{0k}}\right)\right]^{-1}\mathbb{E}\left(\psi_{\theta_{0k}}\psi_{\theta_{0k}}^\mathsf{T}\right)\left[\mathbb{E}\left(\psi'_{\theta_{0k}}\right)\right]^{-1}\right)$$

30

## 2.8 Appendix C: Marginal Score Tests

For each outcome $k = 1, \ldots, K$, the weighted score is

$$
\begin{aligned}
S_k\left(\theta_k\right) &= n^{-1} \sum_{i=1}^{n} w_i z_i \left[y_{ki} - g\left(\theta_k^{\mathsf{T}} z_i\right)\right] \\
&= \begin{pmatrix} n^{-1} \sum_{i=1}^{n} w_i x_i \left[y_{ki} - g\left(\theta_k^{\mathsf{T}} z_i\right)\right] \\[2mm] n^{-1} \sum_{i=1}^{n} w_i s_i \left[y_{ki} - g\left(\theta_k^{\mathsf{T}} z_i\right)\right] \end{pmatrix}
\end{aligned}
$$

Under $H_0 : \beta_k = 0$, the score is

$$
\Psi_k\left(\alpha_k\right) = n^{-1} \sum_{i=1}^{n} w_i x_i \left[y_{ki} - g\left(\alpha_k^{\mathsf{T}} x_i\right)\right]
$$

Let $\tilde{\alpha}_k$ be the solution to $\Psi_k\left(\alpha_k\right) = 0$ so that $\tilde{\alpha}_k$ is a consistent estimator of $\alpha_{k0}$ under $H_0$.

Now, consider the score test using

$$
\hat{S}_k\left(\tilde{\alpha}_k\right) = n^{-1} \sum_{i=1}^{n} w_i s_i \left[y_{ki} - g\left(\tilde{\alpha}_k^{\mathsf{T}} x_i\right)\right] \xrightarrow{p} 0 \text{ under } H_0
$$

We now need to find the distribution of $\sqrt{n}\hat{S}_k\left(\tilde{\alpha}_k\right)$. A Taylor expansion around $\alpha_{k0}$ gives

$$
\sqrt{n}\hat{S}_k\left(\tilde{\alpha}_k\right) = \sqrt{n}\hat{S}_k\left(\alpha_{k0}\right) - n^{-1} \sum_{i=1}^{n} w_i s_i x_i^{\mathsf{T}} g'\left(\alpha_{k*}^{\mathsf{T}} x_i\right) \sqrt{n}\left(\tilde{\alpha}_k - \alpha_{k0}\right)
$$

for some $\boldsymbol{\alpha}_{k*}$ satisfying $\|\boldsymbol{\alpha}_{k*} - \boldsymbol{\alpha}_{k0}\| \leq \|\tilde{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_{k0}\|$ so that $\boldsymbol{\alpha}_{k*} - \boldsymbol{\alpha}_{k0} = o_p(1)$. Let $\mathbf{A}_k = \mathbb{E}\left[w_i \mathbf{s}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_{k0}^\mathsf{T} \mathbf{x}_i\right)\right]$ and $\mathbf{B}_k = \mathbb{E}\left[w_i \mathbf{x}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i\right)\right]$. Then $n^{-1} \sum_{i=1}^{n} w_i \mathbf{s}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_{k*}^\mathsf{T} \mathbf{x}_i\right) \overset{p}{\to} \mathbf{A}_k$,

and $\sqrt{n}\left(\tilde{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_{k0}\right) = \sqrt{n} n^{-1} \sum_{i=1}^{n} w_i \mathbf{B}_k^{-1} \mathbf{x}_i \left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T} \mathbf{x}_i\right)\right]$ so that

$$
\begin{aligned}
\sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) &= \sqrt{n}\hat{\mathbf{S}}_k\left(\boldsymbol{\alpha}_{k0}\right) - \mathbf{A}_k \sqrt{n} n^{-1} \sum_{i=1}^{n} w_i \mathbf{B}_k^{-1} \mathbf{x}_i \left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T} \mathbf{x}_i\right)\right] \\
&= \sqrt{n} n^{-1} \sum_{i=1}^{n} w_i \left(\mathbf{s}_i - \mathbf{A}_k \mathbf{B}_k^{-1} \mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T} \mathbf{x}_i\right)\right] \\
&\overset{d}{\to} \mathcal{N}\left(0, \Sigma_k = \mathbb{E}\left[\boldsymbol{\xi}_{ki} \boldsymbol{\xi}_{ki}^\mathsf{T}\right]\right)
\end{aligned}
$$

where

$$
\boldsymbol{\xi}_{ki} = w_i \left(\mathbf{s}_i - \mathbb{E}\left[w_i \mathbf{s}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i\right)\right] \left(\mathbb{E}\left[w_i \mathbf{x}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i\right)\right]\right)^{-1} \mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i\right)\right]
$$

Let $\dim(\mathbf{s}) = m$ so that $\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right)$ is an $m \times 1$ vector. We can then use

$$
\mathbf{M}_1 = n \hat{\mathbf{S}}_k^\mathsf{T}\left(\tilde{\boldsymbol{\alpha}}_k\right) \Sigma_k^{-1} \hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \sim \chi_m^2
$$

Alternatively, we can use $\mathbf{M}_2 = n \hat{\mathbf{S}}_k^\mathsf{T}\left(\tilde{\boldsymbol{\alpha}}_k\right) \hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right)$. To derive the distribution of $\mathbf{M}_2$, look at

$$
\mathbf{Z} = \Sigma_k^{-\frac{1}{2}} \sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \sim \mathcal{N}\left(0, \mathbf{I}_{m \times m}\right)
$$

32

so that

$$\mathbf{M}_2 = \left[\sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right)\right]^{\mathsf{T}} \sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) = \left(\Sigma_k^{\frac{1}{2}}\mathbf{Z}\right)^{\mathsf{T}}\left(\Sigma_k^{\frac{1}{2}}\mathbf{Z}\right) = \mathbf{Z}^{\mathsf{T}}\Sigma_k\mathbf{Z}$$

Since $\Sigma_k$ is symmetric, we can use spectral decomposition to write $\Sigma_k = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}$ for an orthogonal matrix $\mathbf{U}$ and diagonal matrix

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}$$

where $\lambda_i$ are the eigenvalues of $\Sigma_k$. By the properties of orthogonal matrices, let $\mathbf{X} = [X_1, \ldots, X_m]^{\mathsf{T}} = \mathbf{U}^{\mathsf{T}}\mathbf{Z} \sim \mathcal{N}\left(0, \mathbf{I}_{m\times m}\right)$. Then,

$$
\begin{aligned}
\mathbf{M}_2 &= \mathbf{Z}^{\mathsf{T}}\Sigma_k\mathbf{Z} \\
&= \mathbf{Z}^{\mathsf{T}}\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}}\mathbf{Z} \\
&= \mathbf{X}^{\mathsf{T}}\boldsymbol{\Lambda}\mathbf{X} \\
&= \sum_{i=1}^{m} \lambda_i\left(\Sigma_k\right) X_i^2
\end{aligned}
$$

By the properties of the multivariate normal distribution, $X_i$ are iid $\mathcal{N}\left(0, 1\right)$ so that $X_i^2 \sim \chi_1^2$.

Then the distribution of $\mathbf{M}_2$ is a linear combination of independent $\chi_1^2$ random variables, where the weights $\lambda_i\left(\Sigma_k\right)$ are the eigenvalues of $\Sigma_k$.

33

## 2.9    Appendix D: Global Score Tests

With the distribution of $\sqrt{n}\hat{S}_k(\tilde{\alpha}_k)$, we can formulate the global test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$ using

$$
\sqrt{n}\vec{S}(\tilde{\alpha}) \quad := \quad \sqrt{n}
\begin{bmatrix}
\hat{S}_1(\tilde{\alpha}_1) \\
\hat{S}_2(\tilde{\alpha}_2) \\
\vdots \\
\hat{S}_K(\tilde{\alpha}_K)
\end{bmatrix}
$$

$$
= \quad \sqrt{n}
\begin{pmatrix}
n^{-1}\sum_{i=1}^{N}\xi_{1i} \\
\vdots \\
n^{-1}\sum_{i=1}^{N}\xi_{Ki}
\end{pmatrix}
$$

Let $\vec{\xi}_i = (\xi_{1i}, \ldots, \xi_{Ki})$. Then

$$
\sqrt{n}\vec{S}(\tilde{\alpha}) \xrightarrow{d} \mathcal{N}\left(0, \vec{\Sigma}\right)
$$

where

$$
\vec{\Sigma} = \mathbb{E}\left(\vec{\xi}_i\vec{\xi}_i^{\mathsf{T}}\right)
$$

We can similarly formulate test statistics

$$
\begin{aligned}
\mathbf{G}_1 &= n\vec{\mathbf{S}}^\mathsf{T}(\tilde{\boldsymbol{\alpha}})\,\vec{\Sigma}^{-1}\vec{\mathbf{S}}(\tilde{\boldsymbol{\alpha}}) \sim \chi_K^2 \\
\mathbf{G}_2 &= n\vec{\mathbf{S}}^\mathsf{T}(\tilde{\boldsymbol{\alpha}})\,\vec{\mathbf{S}}(\tilde{\boldsymbol{\alpha}}) \sim \sum_{i=1}^{K} \lambda_i\left(\vec{\Sigma}\right)\chi_1^2
\end{aligned}
$$

where $\lambda_i\left(\vec{\Sigma}\right)$ are the eigenvalues of $\vec{\Sigma}$.

## 2.10 Appendix E: Augmentation Methods

Let the augmentation term be $\hat{\boldsymbol{\epsilon}} = n^{-1}\sum_{i=1}^{n}\mathbf{e}\,(1-w_i)\,\mathbf{v}^\mathsf{T}\mathbf{z}_i^a \xrightarrow{p} 0$ for some coefficient $\mathbf{v}$ and auxiliary covariates $\mathbf{z}_i^a$. $\mathbf{e}$ is a vector of 0s and 1s where the 1s indicate the corresponding parameters of $\hat{\boldsymbol{\theta}}_k$ to be augmented. Let $\hat{\boldsymbol{\theta}}_k^a = \hat{\boldsymbol{\theta}}_k + \hat{\boldsymbol{\epsilon}}$. From Appendix B, we are able to write

$$
\sqrt{n}\left(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{k0}\right) = \mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\sqrt{n}\,n^{-1}\sum_{i=1}^{n} w_i\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^\mathsf{T}\mathbf{z}_i\right)\right]
$$

35

Then

$$
\sqrt{n}\left(\hat{\boldsymbol{\theta}}_k^a - \boldsymbol{\theta}_{k0}\right) = \sqrt{n}\left(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{k0}\right) + \sqrt{n}\hat{\boldsymbol{\epsilon}}
$$

$$
= \mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\sqrt{n}n^{-1}\sum_{i=1}^{n}\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right]
$$

$$
+\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\sqrt{n}n^{-1}\sum_{i=1}^{n}\left(w_i - 1\right)\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right]
$$

$$
+\sqrt{n}n^{-1}\sum_{i=1}^{n}\mathbf{e}\left(1 - w_i\right)\mathbf{v}^{\mathsf{T}}\mathbf{z}_i^a
$$

$$
= \sqrt{n}n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right]
$$

$$
+\sqrt{n}n^{-1}\sum_{i=1}^{n}\left(w_i - 1\right)\left[\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^{\mathsf{T}}\mathbf{z}_i^a\right]
$$

Now let

$$
\mathbf{F}_0 = \sqrt{n}n^{-1}\sum_{i=1}^{n}\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right]
$$

$$
\mathbf{F}_a = \sqrt{n}n^{-1}\sum_{i=1}^{n}\left(w_i - 1\right)\left[\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\mathbf{z}_i\left[y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^{\mathsf{T}}\mathbf{z}_i^a\right]
$$

Here $\mathbf{F}_0$ and $\mathbf{F}_a$ converge to a joint normal distribution and are independent. First,

$$
\mathbf{F}_0 \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\Sigma_0\mathbb{E}\left(\boldsymbol{\psi}'_{\boldsymbol{\theta}_{0k}}\right)^{-1}\right)
$$

where

$$
\Sigma_0 = \mathbb{E}\left[\mathbf{z}_i\mathbf{z}_i^{\mathsf{T}}\left(y_{ki} - g\left(\boldsymbol{\theta}_k^{\mathsf{T}}\mathbf{z}_i\right)\right)^2\right]
$$

36

There is nothing we can do to reduce $\Sigma_0$. Next,

$$\mathbf{F}_a \xrightarrow{d} \mathcal{N}\left(0, \Sigma_a\right)$$

Let $\mathcal{D}$ represent the available data. Then

$$
\begin{aligned}
\Sigma_a &= \mathit{Var}\left[\left(w_i - 1\right)\left(\mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^\intercal \mathbf{z}_i^a\right)\right] \\
&= \mathbb{E}\left[\mathit{Var}\left[\left(w_i - 1\right)\left(\mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^\intercal \mathbf{z}_i^a\right) | \mathcal{D}\right]\right] \\
&= \mathbb{E}\left[\mathit{Var}\left(w_i | \mathcal{D}\right)\left(\mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^\intercal \mathbf{z}_i^a\right)\right. \\
&\qquad \left.\left(\mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^\intercal \mathbf{z}_i^a\right)^\intercal\right] \\
&= \mathbb{E}\left[\frac{1 - \pi_i}{\pi_i}\left(\mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^\intercal \mathbf{z}_i^a\right)\right. \\
&\qquad \left.\left(\mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right] - \mathbf{e}\mathbf{v}^\intercal \mathbf{z}_i^a\right)^\intercal\right]
\end{aligned}
$$

Here $\mathcal{I} = \mathbb{E}\left(\psi'_{\theta_{0k}}\right)^{-1} \mathbf{z}_i \left[y_{ki} - g\left(\theta_k^\intercal \mathbf{z}_i\right)\right]$ is the influence function for $\hat{\theta}_k$. We care about reducing the diagonal elements of $\Sigma_a$ corresponding to the genetic variants. That is, the estimator $\hat{\beta}_k^a$ has a corresponding $\hat{\beta}_{ks}^a$ for each SNP $s$, and we can minimize $\mathit{Var}\left(\hat{\beta}_{ks}^a\right)$ by choosing an optimal $\mathbf{v}$. For an individual SNP, the row $s$ column $s$ element of $\Sigma_a$ depends on the $s$ element

of $\mathcal{I}$ through

$$
\begin{aligned}
\Sigma_a[s, s] &= \mathbb{E}\left[\frac{1 - \pi_i}{\pi_i}\left(\mathcal{I}[s] - \mathbf{v}^\mathsf{T}\mathbf{z}_i^a\right)^2\right] \\
&:= \mathbb{E}\left[\frac{1 - \pi_i}{\pi_i}\left(I_{si} - \mathbf{v}^\mathsf{T}\mathbf{z}_i^a\right)^2\right] \\
&= \mathbb{E}\left[w_i\frac{1 - \pi_i}{\pi_i}\left(I_{si} - \mathbf{v}^\mathsf{T}\mathbf{z}_i^a\right)^2\right]
\end{aligned}
$$

where $I_{si}$ is the influence function for $\hat{\beta}_{ks}^w$. This corresponds to a weighted linear regression with outcome $I_{si}$ and covariates $\mathbf{z}_i^a$ to find the optimal $\mathbf{v}$ for each SNP. Putting everything together gives

$$
\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}}_k^a - \boldsymbol{\theta}_{k0}\right) &= \mathbf{F}_0 + \mathbf{F}_a \\
&\xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left(\boldsymbol{\psi}_{\boldsymbol{\theta}_{0k}}'\right)^{-1}\Sigma_0\mathbb{E}\left(\boldsymbol{\psi}_{\boldsymbol{\theta}_{0k}}'\right)^{-1} + \Sigma_a\right)
\end{aligned}
$$

For testing, we can similarly reduce the variance of $\hat{\mathbf{S}}_k(\tilde{\boldsymbol{\alpha}}_k)$ by augmenting it with $\hat{\boldsymbol{\epsilon}}$. The derivation is the same except we use the influence function of $\hat{\mathbf{S}}_k(\tilde{\boldsymbol{\alpha}}_k)$ instead of $\hat{\boldsymbol{\theta}}_k$. Once

again, let $\mathbf{A}_k = \mathbb{E}\left[w_i \mathbf{s}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_{k0}^\mathsf{T} \mathbf{x}_i\right)\right]$ and $\mathbf{B}_k = \mathbb{E}\left[w_i \mathbf{x}_i \mathbf{x}_i^\mathsf{T} g'\left(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i\right)\right]$. Then,

$$
\begin{aligned}
\sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) + \sqrt{n}\hat{\boldsymbol{\epsilon}} \;=\; & \sqrt{n}n^{-1}\sum_{i=1}^{n} w_i\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right] \\
& + \sqrt{n}n^{-1}\sum_{i=1}^{n}\mathbf{e}\left(1 - w_i\right)\mathbf{v}^\mathsf{T}\mathbf{z}_i^a \\
\;=\; & \sqrt{n}n^{-1}\sum_{i=1}^{n}\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right] \\
& + \sqrt{n}n^{-1}\sum_{i=1}^{n}\left(w_i - 1\right)\left[\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right] - \mathbf{e}\mathbf{v}^\mathsf{T}\mathbf{z}_i^a\right]
\end{aligned}
$$

Using the same derivation we did for the asymptotic distribution of $\hat{\boldsymbol{\theta}}_k^a$, we get

$$
\sqrt{n}\hat{\mathbf{S}}_k\left(\tilde{\boldsymbol{\alpha}}_k\right) \;\xrightarrow{d}\; \mathcal{N}\left(0, \mathbf{W}_0 + \mathbf{W}_a\right)
$$

where

$$
\begin{aligned}
\mathbf{W}_0 \;=\; & \mathbb{E}\left(\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right]\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)^\mathsf{T}\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right]^\mathsf{T}\right) \\
\mathbf{W}_a \;=\; & \mathbb{E}\left(\frac{1 - \pi_i}{\pi_i}\left[\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right] - \mathbf{e}\mathbf{v}^\mathsf{T}\mathbf{z}_i^a\right]\right. \\
& \left.\times\left[\left(\mathbf{s}_i - \mathbf{A}_k\mathbf{B}_k^{-1}\mathbf{x}_i\right)\left[y_{ki} - g\left(\boldsymbol{\alpha}_{k0}^\mathsf{T}\mathbf{x}_i\right)\right] - \mathbf{e}\mathbf{v}^\mathsf{T}\mathbf{z}_i^a\right]^\mathsf{T}\right)
\end{aligned}
$$

# Bayesian Hierarchical Modeling of Substate Area Estimates from the Medicare CAHPS Survey

Tianyi Cai
Department of Biostatistics
Harvard University

Alan Zaslavsky
Department of Health Care Policy
Harvard Medical School

## 3.1 Introduction

The Medicare system in the United States is a government-sponsored program which provides health insurance to most legal residents over 65 years old, and younger residents with particular disabilities and conditions. In 2015, Medicare covered 55 million people for \$648 billion in total expenditures (Centers for Medicare & Medicaid, 2016). Medicare beneficiaries may be directly insured by the government through Fee-for-Service (FFS) Medicare, or enroll in a private "Medicare Advantage" (MA) plan reimbursed by the Medicare sponsoring agency, the Centers for Medicare and Medicaid Services (CMS).

Since 1997, quality experiences of beneficiaries enrolled in MA plans have been monitored through the Consumer Assessments of Healthcare Providers and Systems (CAHPS) survey for Medicare (Crofton et al., 1999), or MCAHPS, administered annually to samples of beneficiaries of each MA plan. A version of the MCAHPS survey in its current form has also been administered to FFS beneficiaries, annually since 2007. These surveys were designed as part of a monitoring system, aimed at motivating and guiding quality improvement activities and informing consumer choice (Goldstein et al., 2001). Most MA plans operate in a state, part of a state, or in a few cases, in two or three contiguous states. Thus, it is natural to similarly design the FFS survey to represent states. For national representativeness, and to provide more detailed information on states with large Medicare enrollments, state FFS samples are designed to yield response counts proportional to enrollment (given historical response rates), with a

floor sample size yielding about 1200 responses. Previous studies on MA and FFS survey data suggest geographic variations across areas and across plans within areas (Keenan et al., 2010).

A concise set of survey results is distributed in a handbook to each beneficiary, and a larger set is made available on the Medicare Compare website. Published MCAHPS scores are an information resource for beneficiaries choosing among locally available plans or FFS. Beneficiaries seeking information about available MA plans in comparison to FFS enter a ZIP (postal) code to look up locally relevant information on the Medicare Compare website, which serves up responses distinguished below the state level. For this purpose, 94 substate reporting areas were defined in the 32 states whose samples were designed to provide at least twice the standard 1200 responses targeted for collection in the smallest states. The areas were composed of contiguous counties, had approximately equal sample sizes, and were centered around major metropolitan areas. Thus, estimates for these areas were about as accurate as those for the states receiving the standard minimum sample. Reporting at the state level introduces bias, but direct domain estimates in smaller areas could have high variance. An annual decision to report survey measures at the state or substate level is made for each state, based on an ad hoc $F$-test procedure.

We estimate Bayesian random effects models for all substate area means, jointly modeling these domains of 94 areas in 32 states across 5 years. In the spirit of Reiter (2000), estimates from our best-fitting models are used to identify the proper amount of pooling for presentation of direct estimates (state or substate level in each state), as well as to propose alternative small area estimates superior to either direct estimate, if such are allowed. The best-performing

model was determined using log pseudo-marginal likelihood (Chen et al., 2008) and posterior predictive checks. Using a Bayesian hat matrix, we show heuristically how each domain estimate from our random effects model combines information from other domains.

In the sections that follow, Section 3.2 describes the survey data and the decision procedure used in production of reports since 2007. Section 3.3 presents our Bayesian random effects model, the Bayesian hat matrix, and methods for model comparison. Section 3.4 applies these methods to FFS CAHPS data. Section 3.5 discusses the implications of the results.

## 3.2 The FFS CAHPS Survey and Reporting Rule

The FFS CAHPS survey is administered in February to May of each year to a sample of 275,000 beneficiaries, allocated to states and areas as described in the Introduction. We analyzed data from 2012-2016, with 8 measures that were used in public reporting for part or all of this period. These include one yes/no item on immunization, two overall rating items on a 0-10 scale, and five composites of two or more items on an ordinal never/sometimes/usually/always scale, as shown in Table 3.1 and Zaslavsky & Cleary (2002). Survey measure means for each state and area were weighted to match the MA distribution over counties and adjusted for effects of covariates using linear regression models, and sampling variances of adjusted means (or composites of means) were estimated by Taylor linearization (Agency for Healthcare Research and Quality, 2012).

**Table 3.1:** FFS Data Survey Measures

| Measure | Range of Response Values | Mean[†] | SD[†] | Mean # Responses[†] |
|---|---|---|---|---|
| Coordination of Care[1] | 3.39-3.71 | 3.60 | 0.044 | 690 |
| Customer Service[1] | 3.12-3.83 | 3.50 | 0.089 | 782 |
| Get Care Quickly[1] | 3.06-3.50 | 3.26 | 0.082 | 744 |
| Get Needed Care[1] | 3.27-3.72 | 3.58 | 0.050 | 596 |
| Flu Immunization[2] | 0.56-0.85 | 0.73 | 0.041 | 792 |
| Doctor Communication[1] | 3.60-3.79 | 3.71 | 0.031 | 660 |
| Rating of Care | 8.16-8.85 | 8.55 | 0.117 | 790 |
| Rating of Plan | 7.60-8.77 | 8.31 | 0.182 | 763 |

[1]Composite measures of two or more items.
[2]Binary indicator of immunization.
[3]Rating on a scale from 1 to 10.
[†]Applied over all domains (substate sample means over five years).

For each measure, the weighted sample mean and its sampling variance, after adjusting for respondent covariate effects, were calculated for 94 areas nested in 32 states across the 5 years (2012-2016), for a total of 470 entries. Of the 32 states in the dataset, 17 have the minimum of 2 areas, while the largest states (California, Florida, New York, and Texas) have 5 or more

areas. For each measure, Table 3.1 summarizes the domain means and the mean number of responses, which varies across measures due to skip patterns in the survey (Klein et al., 2011).

Since 2007, the annual reporting decision has been made using an ad hoc *F*-test procedure for each state. If $y_{ijk}$ is the sample mean for area $j$ (of state $i$) in year $k$, $v_{ijk}$ is the sampling variance of $y_{ijk}$, and $n_i$ is the number of areas in state $i$, this hypothesis testing procedure uses the statistic

$$F_{ik} = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} v_{ijk}^{-1} \left( y_{ijk} - m \right)^2, \; m = \frac{\sum_{j=1}^{n_i} v_{ijk}^{-1} y_{ijk}}{\sum_{j=1}^{n_i} v_{ijk}^{-1}}$$

This statistic can be calculated from the sufficient statistics $y_{ijk}$ and $v_{ijk}$ instead of individual level data, which are undefined for composite measures. In the interest of predictive accuracy, we seek a difference in AIC (equivalent to $F_{ik}$ for linear models) greater than 2 for the two models corresponding to the two reporting methods, which would indicate improved accuracy of prediction of the substate estimates over the state estimates. Under this hypothesis testing approach, the annual reported survey measures within each state and its areas are estimated using only data from that state and year, which can lead to estimates with high variances for smaller areas.

## 3.3   Methods

The structure of the FFS data is naturally well-suited for a hierarchical model, with areas nested within states across years. We follow Fay & Herriot (1979) in estimating our hierarchical

model from aggregated data, so the complexities of unequal weighting and skipped items enter the hierarchical model only through the sufficient statistics of sample mean and sampling variance. Similar models using CAHPS data are discussed by O'Malley & Zaslavsky (2008), and Zaslavsky (2007).

### 3.3.1 Bayesian Random Effects Model

At the first stage of the hierarchical model, let $y_i \sim \mathcal{N}(\eta_i, v_i)$, where $\eta_i$ is the weighted, adjusted population mean for a domain defined by a substate area and year, and $y_i$ is the corresponding weighted, adjusted sample mean. We assume further that $y_i$ has negligible bias as an estimator of $\eta_i$, and that the variance $v_i$ is known. At the second stage of the hierarchical model, the population mean is modeled as

$$\eta_i = \mathbb{E}\left(y_i | \theta_{1,t(i,1)}, \ldots, \theta_{G,t(i,G)}\right) = \sum_{g=1}^{G} \theta_{g,t(i,g)}, i = 1, \ldots, n, g = 1, \ldots, G, t = 1, \ldots, T_g$$

where the parameters $\theta_{gt}$ are modeled as either fixed or random effects depending on the specification of their prior distributions. In most of our model specifications, the year effect is modeled as a fixed effect, since five years of data would not provide much information for estimating a variance component for a random year effect. The index $g$ labels the effect being modeled, while $t$ indexes the levels of that effect. For example, a state random effect would be labeled by a particular $g$, and in the design specification would take one of $T_g = 32$ levels, $t = 1, \ldots, 32$, where the state $t$ depends on the domain $i$. This parameterization expresses $\eta_i$

as the sum of $G$ independent $\theta$ parameters characterized by a matrix of indices. Similar parameterizations of multilevel models are presented by (Gelman & Hill, 2006). This formulation provides a concise and general framework for writing out the random effects model and deriving the corresponding conditional posterior distributions, while accommodating complex structures such as crossed and nested random effects. The indexing scheme is simple and efficient computationally, and facilitates adding or removing random effect terms by adding or removing columns from the index matrix.

The fixed effect, corresponding to $g = 1$, is modeled with a diffuse Normal prior $\mathcal{N}\left(0, \tau^2\right)$ for arbitrarily large $\tau$. Random effects are modeled with a Normal prior $\theta_{gt} \sim \mathcal{N}\left(0, \sigma^2_{gs(t)}\right)$. The index $s$ allows $\sigma^2_{gs(t)}$ to potentially vary by a set of indices $\left\{s\left(1\right), \ldots, s\left(T_g\right)\right\} \subset \left\{1, \ldots, T_g\right\}$. For example, a random effect for the substate area could have a separate variance parameter for each state, rather than one variance parameter for all areas. However, in most model specifications, we use a single variance for each random effect group $g$, or $s\left(t\right) = 1, t = 1, \ldots, T_g$.

For each $y_i$, fixing $\theta_{gt}$ for $g \neq g_0$, we have $G$ residual components defined by

$$r_{ig_0} \quad = \quad y_i - \sum_{g \neq g_0} \theta_{g, t(i,g)} \sim \mathcal{N}\left(\theta_{g_0, t(i,g_0)}, v_i\right)$$

At the third stage of the hierarchical model, we give variance components $\sigma^2_{gs(t)}$ a noninformative Inverse-Gamma prior $\mathcal{IG}\left(\alpha, \beta\right)$ for small $\alpha$ and $\beta$. Using the likelihood for data $r_{ig}$ and

our prior distributions, we can write the conditional posteriors as

$$\theta_{1t} \sim \mathcal{N}\left(\left(\sum_{i \in \mathcal{I}_{1t}} r_{i1} v_i^{-1}\right)\left(\tau^{-2} + A_{1t}\right)^{-1}, \left(\tau^{-2} + A_{1t}\right)^{-1}\right), t = 1, \ldots, T_1$$

$$\theta_{gt}|\sigma_{gs}^2 \sim \mathcal{N}\left(\left(\sum_{i \in \mathcal{I}_{gt}} r_{ig} v_i^{-1}\right)\left(\sigma_{gs}^{-2} + A_{gt}\right)^{-1}, \left(\sigma_{gs}^{-2} + A_{gt}\right)^{-1}\right), g = 2, \ldots, G, t = 1, \ldots, T_g$$

$$\sigma_{gs}^2|\theta_{gk} \sim \mathcal{IG}\left(\alpha + \frac{\#\{\mathcal{T}_{gs}\}}{2}, \beta_g + \frac{1}{2}\sum_{t \in \mathcal{T}_{gs}} \theta_{gt}^2\right), g = 2, \ldots, G, s = 1, \ldots, S_g$$

Derivations appear in Appendix A, with a general form for multiple outcomes in Appendix B. Here we have defined $\mathcal{I}_{gt} = \{i : t(i, g) = t\}$ as all the domain-level data belonging to level $t$ of group $g$, and $A_{gt} = \sum_{i \in \mathcal{I}_{gt}} v_i^{-1}, g = 1, \ldots, G, t = 1, \ldots, T_g$. Similarly, we defined $\mathcal{T}_{gs} = \{t : s(t(g)) = s\}$ as all the $t$-level data belonging to level $s$ of group $g$, where $\#\{\mathcal{T}_{gs}\}$ is the cardinality of the set $\mathcal{T}_{gs}$.

To fit these models, we used Markov chain Monte Carlo methods with Gibbs sampling steps to update parameters for the conditional posterior distributions (Casella & George, 1992). At each iteration, we individually sample the $\theta_{gt}$ for each $g$ and $t$, followed by a sample of the $\sigma_{gs}^2$ for each $g$ and $s$. The Gibbs procedure was run for 50,000 iterations, and samples were thinned to reduce autocorrelation by taking every 50th sample.

### 3.3.2 Analysis of Borrowing Strength

The concept of "borrowing strength," introduced by Tukey, is prevalent today in empirical Bayes and other areas of statistics (Brillinger, 2002). In small-area estimation, it involves im-

proving prediction for a domain using information from other domains. We illustrate this concept with a "Bayesian hat matrix" (a term proposed by Steece (1989) in the context of single-level regression with a prior) that identifies the leverage of each area's data on each (same or other) area's estimate. The $n \times n$ Bayesian hat matrix $\mathbf{H}$ satisfies $\hat{\boldsymbol{\eta}} = \mathbf{Hy}$, or $\hat{\eta}_i = \mathbf{H}_i^{\mathsf{T}} \mathbf{y}$, where $\hat{\eta}_i$ is the posterior mean of $\eta_i$, and $\mathbf{H}$ has rows $\mathbf{H}_i$ that sum to 1 (Appendix C). Here, $\hat{\eta}_i$ is a linear combination of all the sample estimates for each domain $(\mathbf{y})$, where the sum of the coefficients of that linear combination is 1. This indicates that the results are "translation-invariant," as expected: if we increased every observed sample mean by the same amount, the posterior means would also shift by nearly the same amount, a consequence of the vague prior on the yearly fixed effects. We summarize these coefficients by adding them up based on relationship between domain $i$ and all other domains. This expresses $\hat{\eta}_i$ as a linear combination of weighted sample estimates within relationship group, and helps us to understand how our data are combined to estimate domain values.

Following up from Section 3.3.1, we can formulate our data in vector form as $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\theta}, \mathbf{y} \sim \mathcal{N}(\boldsymbol{\eta}, \mathbf{V})$, where $\mathbf{X}$ is an $n \times p$ design matrix of 0s and 1s with $p$ columns for each of the $p$ $(g, t)$ pairs, $\boldsymbol{\theta}$ is a $p \times 1$ vector of all $\theta_{gt}$, and $\mathbf{V}$ is an $n \times n$ diagonal matrix with elements $v_i$. Then $\boldsymbol{\theta} \sim \mathcal{N}(0, \boldsymbol{\Omega})$ where $\boldsymbol{\Omega}$ is a $p \times p$ diagonal matrix made up of the corresponding $\boldsymbol{\sigma}^2$ parameters of the $\boldsymbol{\theta}$ ($\sigma_{gs}^2$ for $\theta_{gt}$, and $\tau^2$ for the fixed effect). For a fixed $\boldsymbol{\Omega}$,

$$
\begin{aligned}
f(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\Omega}) \quad &\propto \quad f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{\Omega}) \\
&\sim \quad \mathcal{N}\left[\left(\boldsymbol{\Omega}^{-1} + \mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{y}, \left(\boldsymbol{\Omega}^{-1} + \mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\right]
\end{aligned}
$$

Then the estimated $\hat{\boldsymbol{\eta}}$ can be written as

$$\hat{\boldsymbol{\eta}} = \mathbb{E}\left(\boldsymbol{\eta}|\Omega, \mathbf{y}\right) = \mathbb{E}\left(\mathbf{X}\boldsymbol{\theta}|\Omega, \mathbf{y}\right) \quad = \quad \mathbf{X}\left(\Omega^{-1} + \mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{V}^{-1}\mathbf{y}$$

$$:= \quad \mathbf{H}\left(\Omega\right)\mathbf{y}$$

To get the hat matrix for unconditional posterior means, we average $\mathbf{H}\left(\Omega\right)$ over draws of the $\Omega$ components to get $\mathbf{H}$. Note that $\mathbf{H}$ here is not the traditional hat matrix from linear regression, as we have an additional $\Omega$ component, and $\mathbf{X}$ is not the standard design matrix. Nonetheless, the rows of $\mathbf{H}$ sum to 1, as shown in Appendix C.

Because this $470 \times 470$ matrix is too big for ready examination, we summarize the coefficients by grouping them up based on the relationship between domain $i$ and each other domain, and summing them in each group. This expresses $\hat{\eta}_i$ as a linear combination of weighted sample estimates within relationship group. We then average the group weights over domains, yielding insight into how the data are combined to calculate domain estimates. The groups are:

1. Same area and year.

2. Same area, different year.

3. Different area in the same state, same year.

4. Different area in the same state, different year.

5. Different state, same year.

6. Different state, different year.

50

To summarize, we first view the estimated $\hat{\eta}_i$ as a linear combination of all the existing data **y** where the coefficients are from the corresponding row of **H** and sum to 1. We then would like to know how much weight is attributed to the existing data from each of the six relationship categories, revealing where and how much information is being borrowed from other domains.

### 3.3.3   Cross-Validation for Hierarchical Models

Traditional Bayesian model comparison methods such as Bayes factors (Kass & Raftery, 1995) could be difficult to compute or approximate for complex hierarchical models. Additionally, it is difficult to formulate our model structure to make entirely comparable models without resorting to the use of arbitrary constants, such as the variance parameter of our fixed effects. Since our primary interest is predictive accuracy, we would like to perform cross-validation. However, this is difficult due to the computational burden of re-fitting our model $n$ times for leave-one-out cross-validation.

We instead propose using importance sampling weighting to compute a cross-validated prediction, motivated by outlier detection methods by Zaslavsky & Bradlow (2010). Instead of re-fitting a model $n$ times, importance sampling allows us to re-weigh our posterior draws of $\theta$ to calculate leave-one-out cross-validated prediction errors. For a new $\tilde{y}_i$ and model $\mathcal{M}$, we select the model that maximizes the log pseudo-marginal likelihood $\ell\left(\mathcal{M}\right) = \sum_{i=1}^{n} \log\left[p\left(y_i|\mathbf{y}_{(-i)}\right)\right]$ (Chen et al., 2008), where $p\left(y_i|\mathbf{y}_{(-i)}\right)$ is estimated using importance sampling, shown in Ap-

pendix D. We can then estimate $\hat{\ell}(\mathcal{M}) = \sum_{i=1}^{n} \log \left( T \left[ \sum_{t=1}^{T} \frac{1}{p\left(y_i | \theta_{(-i)}^{(t)}\right)} \right]^{-1} \right)$, where $T$ is

the number of draws of from the posterior distribution. Similar derivations are provided by

(Gelfand & Dey, 1994).

## 3.4 Application to FFS CAHPS Data

A comparison of posterior distributions of the variance components quantifies the contributions of random effects at different levels to variation in the domain means. State and substate random-effects variances are larger than the substate by year interaction random-effects variances, by an average factor of 7 over all measures. As is typical for a shrinkage estimator (Ghosh & Rao, 1994), the Bayesian estimators from the model had about 1/3 the variance of the direct estimates, over all measures. The hat-matrix analysis revealed a large weight attributed to data from the same domain across years (groups 1 and 2), but this amount varied by measure. Finally, we were able to use the model estimates to evaluate the performance of the naive state or substate estimators through mean squared error calculations.

### 3.4.1 Model Selection

We fit five Bayesian random effects models using the flexible parameterization described in Section 3.3.1. In order of simplicity, these models were:

1. An annual model separately for each year, with an intercept fixed effect, state random effect, and substate random effect.

2. A "base" model with a year fixed effect, state random effect, substate random effect, and substate by year random effect.

3. A substate by year model, which is the base model with a substate by year random effect.

4. A "rich" model, which is the substate by year model whose substate random effect has a variance parameter that varies by state.

5. A "full" model with an intercept fixed effect, year random effect, state random effect, substate random effect, state by year random effect, and substate by year random effect.

The model specifications are presented in Table 3.2. In most cases we estimated the year effect as fixed, with a large prior variance, since it is more feasible to estimate the annual trend using the ample data than it is to try to estimate variance components for a parameter with only five data points. These models were applied to all 8 measures, with data from 94 areas in 32 states, spanning 2012 to 2016.

**Table 3.2:** Specification of Candidate Models

| Parameter | # Levels $(T_g)$ | Annual Model | Base Model | State by Year Model[‡] | Rich Model[‡] | Full Model |
|---|---|---|---|---|---|---|
| intercept $\theta_0$ | 1 | ✓ | | | | ✓ |
| year effect $\theta_{1t}$ | 5 | ✓ | ✓ | ✓ | ✓ | ✓ |
| state effect $\theta_{2t}$ | 32 | ✓ | ✓ | ✓ | ✓ | ✓ |
| substate effect $\theta_{3t}$ | 94 | | ✓ | ✓ | ✓ | ✓ |
| state by year effect $\theta_{4t}$ | 160 | | | ✓ | ✓ | ✓ |
| substate by year effect $\theta_{5t}$ | 470 | | ✓ | ✓ | ✓ | ✓ |

[‡]These two models differ in the variance specification of the substate random effect $\theta_{3t}$. In the state by year model, $\theta_{3t} \sim \mathcal{N}\left(0, \sigma_3^2\right)$. In the rich model, $\theta_{3t} \sim \mathcal{N}\left(0, \sigma_{3s}^2\right)$, where $s$ is the state defined by substate $t$.

Table 3.3 shows the calculation of the log pseudo-marginal likelihood for our five candidate models, where the base model has the highest $\hat{\ell}\left(\mathcal{M}\right)$ across virtually all measures. Additionally, a posterior predictive check on the sample variance of all domains was performed and shown in Table 3.4, where posterior predictive $p$-values for the base model are the most reasonable out of all candidate models. Based on these diagnostics, we decided to use our base model.

**Table 3.3:** Values and Ranks of $\hat{\ell}(\mathcal{M})$ for Candidate Models

| Measure | Annual Model | Base Model | State by Year Model | Rich Model | Full Model |
|---|---|---|---|---|---|
| Coordination of Care | 896 (5) | 1033 (1) | 1027 (2) | 1023 (4) | 1027 (3) |
| Customer Service | 516 (1) | 516 (2) | 515 (4) | 501 (5) | 516 (3) |
| Get Care Quickly | 623 (5) | 937 (1) | 928 (2) | 928 (4) | 928 (3) |
| Get Needed Care | 823 (5) | 888 (1) | 880 (3) | 875 (4) | 880 (2) |
| Flu Immunization | 875 (5) | 1114 (1) | 1113 (3) | 1112 (4) | 1114 (2) |
| Doctor Communication | 990 (5) | 1057 (1) | 1050 (3) | 1041 (4) | 1050 (2) |
| Rating of Care | 400 (5) | 508 (1) | 504 (3) | 503 (4) | 504 (2) |
| Rating of Plan | 293 (5) | 446 (1) | 444 (3) | 442 (4) | 444 (2) |

The log pseudo-marginal likelihood $\hat{\ell}(\mathcal{M})$ is shown, along with its rank in parentheses among the five candidate models.

**Table 3.4:** Posterior Predictive Check of Empirical Standard Deviation

| Measure | $\widehat{sd}\,(\mathbf{y})$ | Posterior Predictive $p$-values | | | | |
| | | Annual Model | Base Model | State by Year Model | Rich Model | Full Model |
|---|---|---|---|---|---|---|
| Coordination of Care | 0.044 | 0.239 | 0.276 | 0.139 | 0.023 | 0.138 |
| Customer Service | 0.089 | 0.082 | 0.231 | 0.170 | 0.011 | 0.189 |
| Get Care Quickly | 0.082 | 0.408 | 0.336 | 0.235 | 0.153 | 0.260 |
| Get Needed Care | 0.050 | 0.449 | 0.483 | 0.380 | 0.051 | 0.418 |
| Flu Immunization | 0.041 | 0.288 | 0.157 | 0.080 | 0.016 | 0.075 |
| Doctor Communication | 0.031 | $< 0.001$ | 0.017 | 0.005 | $< 0.001$ | 0.001 |
| Rating of Care | 0.117 | 0.381 | 0.398 | 0.324 | 0.115 | 0.342 |
| Rating of Plan | 0.182 | 0.359 | 0.471 | 0.464 | 0.275 | 0.478 |

Draws of the posterior predictive distribution are generated from samples of the posterior distribution of $\eta$.

## 3.4.2 Base Model Results

The posterior distributions of the variance components $\sigma^2_{gs}$ of each random effect is standardized by the mean $\bar{v}$ of the sampling variance for each measure, and summarized on a log scale

in Figure 3.1. Similar patterns are apparent for all measures except the Customer Service measure. The posterior means of variances of the state and substate random effects range from 0.34 to 4.71 and are generally larger than the corresponding variances of the substate by year interaction random effect, which range from 0.10 to 0.28. The consistently small variation in substate by year effects suggest that these measures are fairly stable from year to year. Get Care Quickly stands out as having very high variability in its state and substate random effects, but not in year interactions, possibly reflecting persistent local variations in the adequacy of the supply of medical professionals. Flu immunization similarly shows large state and substate variation, possibly due to climate-related variations in immunization rates. Customer Service is an outlier from these typical patterns. Its state and substate random effects show much less variation than for any other measure, even less than the substate by year interaction; this is consistent with the absence of a locally-oriented customer service staff for FFS Medicare. The year main effects range from 3 to 4 for the composite measures, 0.65 to 0.80 for the binary flu measure, and 8 to 9 for the ratings of care and plan. These differences among the year fixed effects are substantially larger than the corresponding estimates for random effects.

57

**Figure 3.1:** Posterior Distributions of Random-Effects Variances
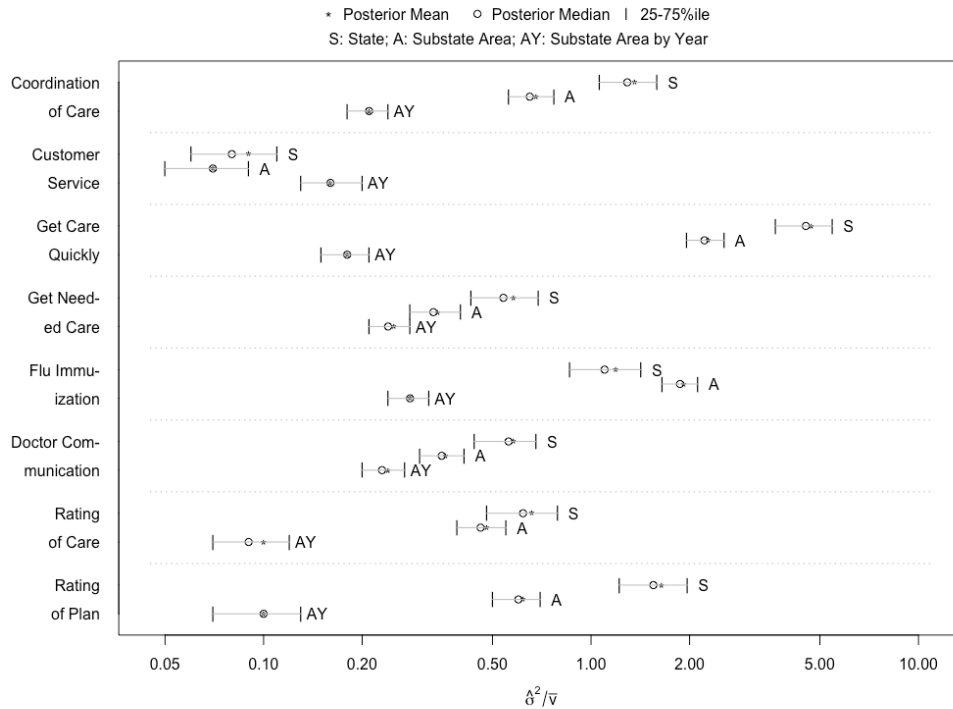


Each of the four random-effects variances are summarized in a different row for each measure, separated by the dotted grey lines. The posterior mean, median, and 25% and 75% quantiles are plotted. To remove the uninteresting differences in variance components due to the use of difference scales for different measures, the samples of the posterior variances were scaled relative to the mean $\bar{v}$ of the sampling variance for each measure. To account for the large differences in the variability of random effects among the different measures, the samples are plotted on a log scale.

Figure 3.2 compares the posterior variances of the domain estimates from the Bayesian model to the sampling variances of the corresponding direct estimates. Within each measure, areas in small states see the most variance reduction, while areas in larger states have the least. Customer Service shows greater variance reduction than all other measures, with pos-

58

terior variances of the domain estimates at roughly 20-30% of their corresponding sampling variances. This is consistent with the random-effects variances for Customer Service, which are relatively small (Figure 3.1), indicating that data from other areas, states, and years are informative for model estimates in each domain. By contrast, Flu Immunization has posterior variances at roughly 40-45% of their corresponding sampling variances, which is consistent with its higher state and substate random-effects variances, and the relatively small sampling variances due to the high item response rate to this measure.

**Figure 3.2:** Variance Reduction, Bayesian vs. Direct Area Estimates
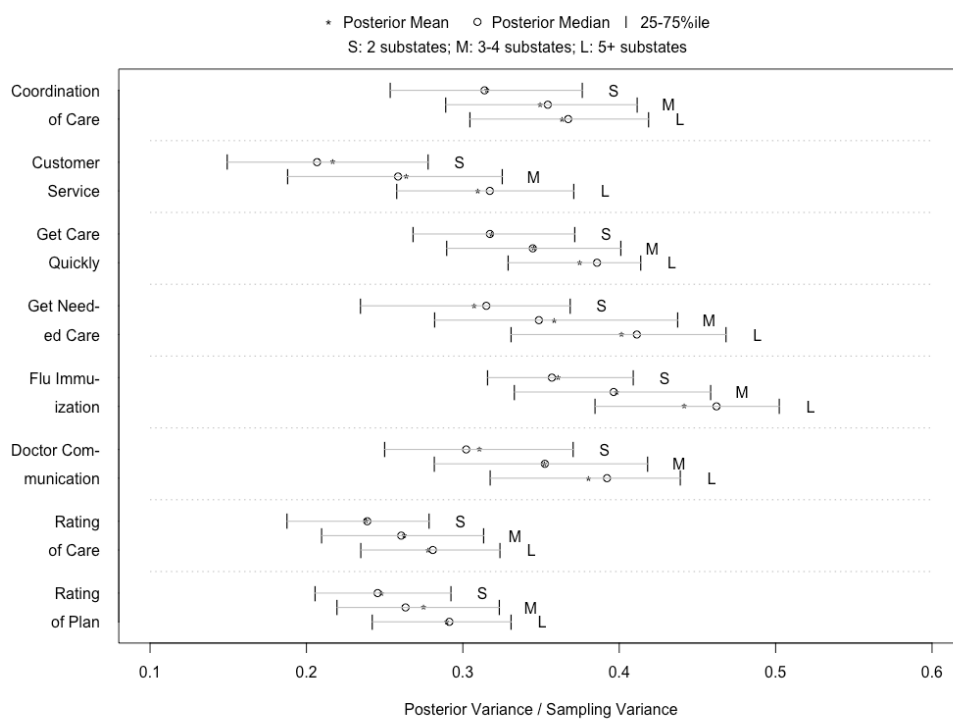


For each measure, the ratio of the posterior variance and sampling variance for domains are summarized and stratified based on the size of the domain's state.

### 3.4.3 Bayesian Hat Matrix Analysis

We compute the hat matrix **H** for each measure. Results for Get Care Quickly are summarized in Table 3.5, and are typical of most measures. The more distinctive patterns for Customer Service are also shown in Table 3.6.

**Table 3.5:** Hat Matrix Results for Get Care Quickly

| | Get Care Quickly | | | |
|---|---|---|---|---|
| Relationship[1] | Average # Domains[2] | Mean of Weight[3] | 5%-ile of Weight[3] | 95%-ile of Weight[3] |
| 1 | 1 | 0.34 | 0.20 | 0.47 |
| 2 | 4 | 0.60 | 0.50 | 0.71 |
| 3 | 2.6 | 0.03 | 0.01 | 0.06 |
| 4 | 10.2 | 0.01 | -0.01 | 0.04 |
| 5 | 90.4 | 0.63 | 0.50 | 0.76 |
| 6 | 361.8 | -0.62 | -0.74 | -0.49 |

[1]Relationships are coded as 1: same area and year, 2: same area, different year, 3: different area in the same state, same year, 4: different area in the same state, different year, 5: different state, same year, 6: different state, different year.
[2]Average over all rows of the number of domains in each relationship group.
[3]Weight is the sum of coefficients by group for a particular row. Mean and percentiles are over all rows of **H**.

**Table 3.6:** Hat Matrix Results for Customer Service

| | | Customer Service | | |
|---|---|---|---|---|
| Relationship[1] | Average # Domains[2] | Mean of Weight[3] | 5%-ile of Weight[3] | 95%-ile of Weight[3] |
| 1 | 1 | 0.26 | 0.10 | 0.42 |
| 2 | 4 | 0.24 | 0.17 | 0.32 |
| 3 | 2.6 | 0.06 | 0.02 | 0.12 |
| 4 | 10.2 | 0.15 | 0.06 | 0.26 |
| 5 | 90.4 | 0.68 | 0.51 | 0.86 |
| 6 | 361.8 | -0.39 | -0.51 | -0.28 |

[1]Relationships are coded as 1: same area and year, 2: same area, different year, 3: different area in the same state, same year, 4: different area in the same state, different year, 5: different state, same year, 6: different state, different year.
[2]Average over all rows of the number of domains in each relationship group.
[3]Weight is the sum of coefficients by group for a particular row. Mean and percentiles are over all rows of **H**.

For a particular model estimate for a domain, group 1 represents the direct estimate for that particular domain (substate area and target year). Group 2 contains direct estimates of the same area, but from the four other years in the dataset. The average of the group 2 domain means can be viewed as as an estimator for the domain mean outside the target year. Group 5 represents direct estimates of out-of-state areas from the same year, while group 6 repre-

sents direct estimates of out-of-state areas from different years. Thus, the average difference between the group 5 and 6 domain means can be viewed as estimate of the overall trend between the target year domain mean and the mean of other years. This trend, in combination of the group 2 domain means, provide a trend-corrected estimate for the target domain. The small magnitude of our substate by year interaction random effects suggests that this trend adjustment should be about the same in every area. Similarly, group 3 and 4 provide an analogous trend, but within the target state.

Group weight means and percentiles were computed for all domains, or rows of the hat matrix. Group 1 weights range from a mean of 0.25 to 0.39 across all measures. Group 2 weights range from a mean of 0.24 to 0.57 across all measures, making each of the four domains in group 2 less informative than the single domain in group 1. When combined, group 1 and 2 account for a mean weight of 0.77 to 0.94 for all measures except Customer Service, whose combined weight for groups 1 and 2 is 0.50. This indicates that direct estimates for the same area in the same year and over other years are both very informative for our small-domain estimates.

The average number of domains used in computing each weight was displayed to demonstrate that a large number of minor effects can sum up to have a significant weight. For example, the group 2 effect is represented by four domain estimates (the other four years from the same area), and combine to have a weight of 0.60 for Get Care Quickly. The group 5 effect for Get Care Quickly is 0.63, but is a combination of (on average) 90 other domains, depending on the area and year. Although the workings of the estimators under the Bayesian model

62

are entirely mechanical, this analysis provides a heuristic explanation for the model's results, which can be important when the results are publicly reported and have policy significance.

### 3.4.4 Formulating a Decision Rule by Comparing MSE

The original motivating problem was to formulate a decision rule for reporting either at the state (pooled) or substate area (unpooled) level, for each state $i$ and year $k$. In this scenario, we suppose that the direct estimates are the only allowable estimates that could be published, so the role of the Bayesian model is to choose between pooled or unpooled estimates in each state and year. We evaluate the naive direct estimates by estimating predictive mean square error (MSE). This uses the Bayesian estimates to determine which states would benefit from pooling in the long run, given the variance components and sampling errors of each state's design, but without looking at the specific estimates for each year.

Let $\eta^*$ be the future $\eta$ for a particular year and state $s$, where $\mathcal{I}_s = \{i : t(i, 2) = s\}$ are the domains in state $s$ (recall that $g = 2$ refers to the state effect), and $n_s$ is the number of areas in state $s$. Let $\bar{y}_s^*$ and $\bar{\eta}_s^*$ be the corresponding pooled state $s$ averages of $y_i^*$ and $\eta_i^*$, respectively. To account for the future covariation of $\eta^*$ and the corresponding future data $\mathbf{y}^*$, we compute
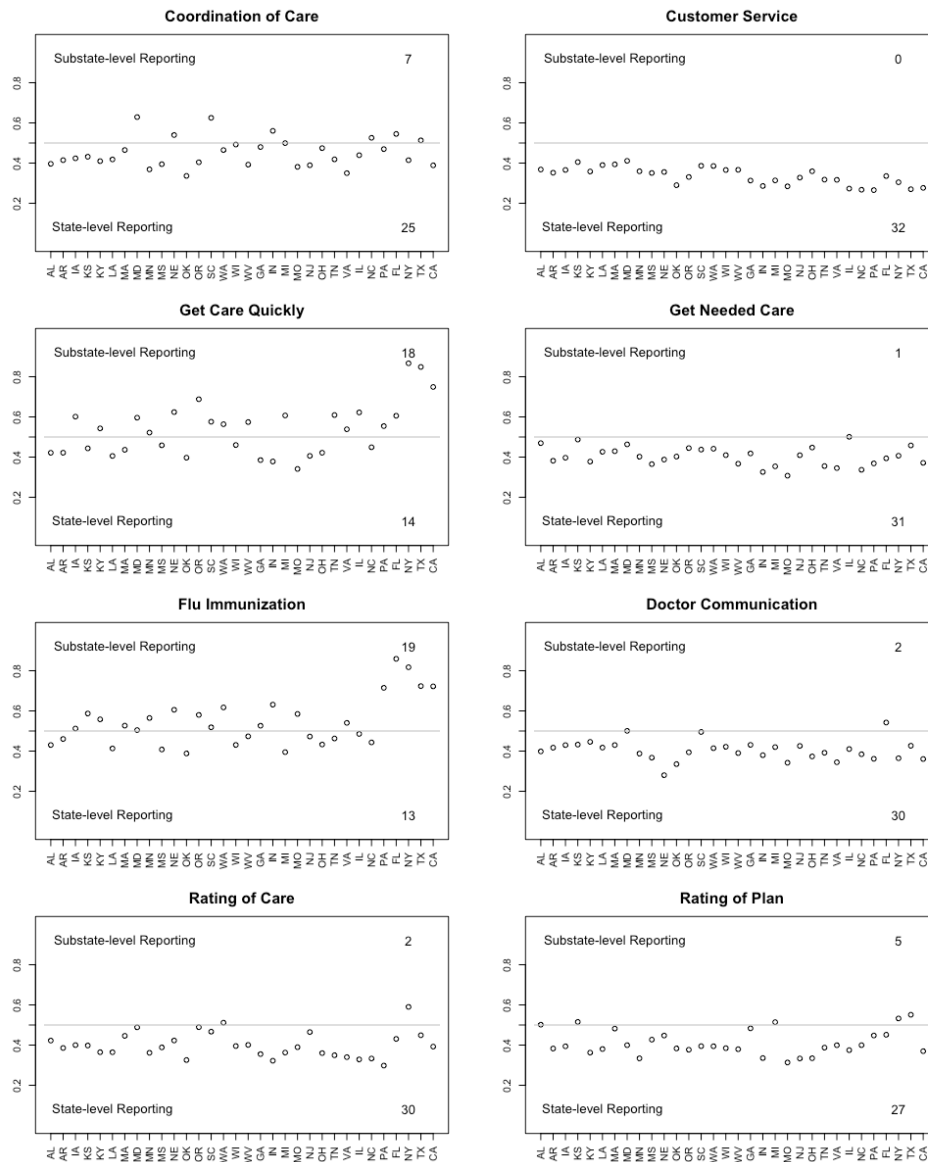
for each state $s$

$$
\begin{aligned}
\text{MSE}_{\text{unpooled}} \;&=\; \sum_{i \in \mathcal{I}_s} \mathbb{E}\left[(y_i^* - \eta_i^*)^2\right] = \sum_{i \in \mathcal{I}_s} Var(y_i^*) = \sum_{i \in \mathcal{I}_s} v_i^* \\
\text{MSE}_{\text{pooled}} \;&=\; \sum_{i \in \mathcal{I}_s} \mathbb{E}\left[(\bar{y}_s^* - \eta_i^*)^2\right] = \sum_{i \in \mathcal{I}_s} \left\{ \mathbb{E}\left[(\bar{y}_s^* - \bar{\eta}_s^*)^2\right] + \mathbb{E}\left[(\bar{\eta}_s^* - \eta_i^*)^2\right] \right\} \\
&=\; \sum_{i \in \mathcal{I}_s} \left\{ Var(\bar{y}_s^*) + \mathbb{E}\left[(\bar{\eta}_s^* - \eta_i^*)^2\right] \right\}
\end{aligned}
$$

$v_i^*$ is estimated as the value of $v_i$ for the most recent year of data, $Var(\bar{y}_s^*)$ is estimated as a weighted average of the $v_i^*$ based on sample size in each area, and $\mathbb{E}\left[(\bar{\eta}_s^* - \eta_i^*)^2\right]$ is estimated using draws from the distribution of $\eta_i^*$. To sample from the distribution of $\eta_i^*$, we sample from the predictive distributions of the future $\theta$ parameters, denoted by a '$*$' superscript. Because $\theta_{2t}^*$ (state random effect) and $\theta_{3t}^*$ (substate random effect) do not vary by year, we sample them directly from their joint posterior distribution. However, $\theta_{4t}^*$ (substate by year random effect) are sampled from the updated prior distribution $\mathcal{N}\left(0, \sigma_4^2|\mathbf{y}\right)$, where $\sigma_4^2|\mathbf{y}$ is sampled from its posterior distribution. We can then calculate $\eta_i^* = \sum_{g=1}^{G} \theta_{g, t(i,g)}^*$.

The ratios $\frac{\text{MSE}_{\text{pooled}}}{\text{MSE}_{\text{pooled}} + \text{MSE}_{\text{unpooled}}}$ puts the comparison of pooled and unpooled MSE on a scale from 0 to 1, where 0 and 1 indicate error-free estimation of pooled and unpooled means, respectively, and 0.5 represents equal MSE for pooled and unpooled estimates. These values are plotted in Figure 3.3 for each state, where each vertical series of five dots represent the ratios across the five years. A simple decision rule would be to report pooled state estimates for ratios below 0.5, and to report unpooled substate estimates for ratios above 0.5. In these plots, a higher ratio indicates higher heterogeneity of measure results within states for a particular

64

year.

**Figure 3.3:** MSE Ratios for Future Year



States are in increasing order by number of areas. For each measure, the top number is the number of states where unpooled substate estimates have lower MSE, and the bottom number is the number of states where pooled state estimates have lower MSE.

65

Not surprisingly, the larger states tend to have higher MSE for the pooled state estimates, which is most evident for Get Care Quickly and Flu Immunization. This trend is a reflection of the substate variation within large states. For most other measures, and especially Customer Service, the unpooled MSE is higher, suggesting reporting at the state level for all states and years.

## 3.5   Discussion

Our Bayesian random effects model for small area estimation provides a general framework for small area estimation of measures repeated over time in nested geographic domains. In our application to a major health care quality survey, model-based small area estimates are substantially more precise than the corresponding direct estimates. We can either report these improved Bayesian estimates in place of the direct estimates, or use our model to evaluate the direct state and substate estimates, providing a decision rule for reporting either one.

Variance component estimates under the Bayesian model have important policy implications. The substate by year interaction random effects have generally small posterior variances, highlighting the stability of quality differences over time within each area. On the other hand, state and substate random effects are generally more variable. Customer service stands out as a measure with particularly low variability. This is an unsurprising consequence of the absence of a distinctive local customer service "presence" in FFS comparable to that of private plans in Medicare Advantage, where customer service assessments are both more variable and

more volatile. Other composite measures and ratings tend to have the most variability at the state level, followed by the substate level, which indicates some heterogeneity between areas in the same state, but not as much as the amount between states. Flu Immunization, potentially provided by a wide range of local institutions with differing resources and effectiveness, has the most variability at the substate level. These differences in statistical properties among the measures present an additional challenge, as it may not be feasible to select different reporting procedures for different measures.

A Bayesian hat matrix summarizes the way information is borrowed across domains to obtain the model estimates. Summaries of the hat matrix allow us to see how much weight data from other domains have in the estimate in each domain, as well as how much variability there is among these weights. The hat matrix demonstrates a novel way of understanding the way model estimates borrow information across domains. In most cases, the majority of the weight for the estimate of the domain means is allotted to the direct estimates of that domain, along with the corresponding direct estimates from past years. If the model-based estimates are reported, these summaries may help to give some transparency to what otherwise is a rather opaque procedure.

The options available under the present system are very limiting. The same reporting rule must be used for all measures, even though it is clear from our results that certain measures behave differently than others. Additionally, we are currently forced to choose between state or substate estimates for each state, across all measures. This could be troublesome if some measures demonstrate higher heterogeneity within states than others. However, these conditions

may be reasonable from a policy point of view, as direct state or substate estimates are easy

for beneficiaries to understand, compared to a more sophisticated model. Ultimately, it will be

up to policymakers to weigh the advantages and disadvantages of using our methodologies.

The correlations among the multiple measures are ignored when they are analyzed sepa-

rately. Using the relationships among the measures can lead to more precise estimates, and

possibly more consistent recommendations for reporting of all measures than would come

from unlinked analyses. A natural extension of this work would be to develop a multivariate

version of the existing model.

## 3.6 Appendix A: Univariate Posterior Distribution

The joint posterior can be formulated as

$$
\begin{aligned}
p\left(\boldsymbol{\theta}|\mathbf{r}\right) \propto \mathcal{L}\left(\mathbf{r}|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}|\boldsymbol{\sigma}^2\right) \pi\left(\boldsymbol{\sigma}^2\right) \quad = \quad & \prod_{i=1}^{N}\prod_{g=1}^{G} \frac{1}{\sqrt{2\pi v_i}} \exp\left[-\frac{1}{2v_i}\left(r_{ig}-\theta_{gt}\right)^2\right] \\
& \prod_{t=1}^{T_1} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2}\theta_{1t}^2\right) \text{ (fixed effect)}, g=1 \\
& \prod_{g=2}^{G}\prod_{t=1}^{T_g} \frac{1}{\sqrt{2\pi\sigma_{gs}^2}} \exp\left(-\frac{1}{2\sigma_{gs}^2}\theta_{gt}^2\right) \text{ (random effects)} \\
& \prod_{g=2}^{G}\prod_{s=1}^{S_g} \left(\sigma_{gs}^2\right)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma_{gs}^2}\right)
\end{aligned}
$$

Fixing a $t$ fixes a particular set of $i$ as well. Define $\mathcal{I}_{gt} : \{i : t(i, g) = t\}$. Fixing an $s$ fixes a particular set of $t$ as well. Define $\mathcal{T}_{gs} : \{t : s(t(g)) = s\}$. Then the conditional posteriors are

$$p\left(\theta_{gt}|\sigma_{gs}^2\right) \propto \exp\left[-\sum_{i\in\mathcal{I}_{gt}} \frac{1}{2v_i}\left(r_{ig} - \theta_{gt}\right)^2 - \frac{1}{2\sigma_{gs}^2}\theta_{gt}^2\right]$$

$$g = 2, \ldots, G, t = 1, \ldots, T_g$$

$$\sim \mathcal{N}\left(\left(\sum_{i\in\mathcal{I}_{gt}} r_{ig}v_i^{-1}\right)\left(\sigma_{gs}^{-2} + \sum_{i\in\mathcal{I}_{gt}} v_i^{-1}\right)^{-1}, \left(\sigma_{gs}^{-2} + \sum_{i\in\mathcal{I}_{gt}} v_i^{-1}\right)^{-1}\right)$$

$$p\left(\theta_{1t}\right) \sim \mathcal{N}\left(\left(\sum_{i\in\mathcal{I}_{1t}} r_{i1}v_i^{-1}\right)\left(\tau^{-2} + \sum_{i\in\mathcal{I}_{1t}} v_i^{-1}\right)^{-1}, \left(\tau^{-2} + \sum_{i\in\mathcal{I}_{1t}} v_i^{-1}\right)^{-1}\right)$$

$$p\left(\sigma_{gs}^2|\theta_{gt}\right) \propto \left(\sigma_{gs}^2\right)^{-\frac{1}{2}\left(\sum_{t\in\mathcal{T}_{gs}} 1\right)-\alpha-1} \exp\left(-\frac{1}{2\sigma_{gs}^2}\sum_{t\in\mathcal{T}_{gs}} \theta_{gt}^2 - \frac{\beta_g}{\sigma_{gs}^2}\right)$$

$$g = 2, \ldots, G, s = 1, \ldots, S_g$$

$$\sim \mathcal{IG}\left(\alpha + \frac{\#\{\mathcal{T}_{gs}\}}{2}, \beta_g + \frac{1}{2}\sum_{t\in\mathcal{T}_{gs}} \theta_{gt}^2\right)$$

70

## 3.7 Appendix B: Multivariate Posterior Distribution

The joint posterior can be formulated as

$$p\left(\boldsymbol{\theta}|\mathbf{r}\right) \propto \mathcal{L}\left(\mathbf{r}|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}|\boldsymbol{\sigma}^2\right) \pi\left(\boldsymbol{\sigma}^2\right)$$

$$= \prod_{i=1}^{N}\prod_{g=1}^{G} (2\pi)^{-\frac{M}{2}} |\mathbf{V}_i|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\mathbf{r}_{ig} - \boldsymbol{\theta}_{gt}\right)^{\mathsf{T}} \mathbf{V}_i^{-1} \left(\mathbf{r}_{ig} - \boldsymbol{\theta}_{gt}\right)\right]$$

$$\prod_{t=1}^{T_1} \left(2\pi\tau^2\right)^{-\frac{M}{2}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\theta}_{1t}^{\mathsf{T}}\boldsymbol{\theta}_{1t}\right) \text{ (fixed effect)}, g = 1$$

$$\prod_{g=2}^{G}\prod_{t=1}^{T_g} (2\pi)^{-\frac{M}{2}} |\boldsymbol{\Sigma}_{gs}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\boldsymbol{\theta}_{gt}^{\mathsf{T}}\boldsymbol{\Sigma}_{gs}^{-1}\boldsymbol{\theta}_{gt}\right) \text{ (random effects)}$$

$$\prod_{g=2}^{G}\prod_{s=1}^{S_g} |\boldsymbol{\Sigma}_{gs}|^{-\frac{v+M+1}{2}} \exp\left[-\frac{1}{2}\text{tr}\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_{gs}^{-1}\right)\right]$$

Fixing a $t$ fixes a particular set of $i$ as well. Define $\mathcal{I}_{gt} : \{i : t(i, g) = t\}$. Fixing an $s$ fixes a

particular set of $t$ as well. Define $\mathcal{T}_{gs} : (t : s(t(g)) = s)$. Then the conditional posteriors are

71

$$p\left(\boldsymbol{\theta}_{gt}|\boldsymbol{\Sigma}_{gs}\right) \quad \propto \quad \exp\left[-\sum_{i\in\mathcal{I}_{gt}}\frac{1}{2}\left(\mathbf{r}_{ig}-\boldsymbol{\theta}_{gt}\right)^{\mathsf{T}}\mathbf{V}_i^{-1}\left(\mathbf{r}_{ig}-\boldsymbol{\theta}_{gt}\right)-\frac{1}{2}\boldsymbol{\theta}_{gt}^{\mathsf{T}}\boldsymbol{\Sigma}_{gs}^{-1}\boldsymbol{\theta}_{gt}\right]$$

$$g=2,\dots,G,\, t=1,\dots,T_g$$

$$\sim \quad \mathcal{N}\left(\left(\boldsymbol{\Sigma}_{gs}^{-1}+\sum_{i\in\mathcal{I}_{gt}}\mathbf{V}_i^{-1}\right)^{-1}\left(\sum_{i\in\mathcal{I}_{gt}}\mathbf{V}_i^{-1}\mathbf{r}_{ig}\right),\left(\boldsymbol{\Sigma}_{gs}^{-1}+\sum_{i\in\mathcal{I}_{gt}}\mathbf{V}_i^{-1}\right)^{-1}\right)$$

$$p\left(\boldsymbol{\theta}_{1t}\right) \quad \sim \quad \mathcal{N}\left(\left(\tau^{-2}\mathbf{I}+\sum_{i\in\mathcal{I}_{1t}}\mathbf{V}_i^{-1}\right)^{-1}\left(\sum_{i\in\mathcal{I}_{1t}}\mathbf{V}_i^{-1}\mathbf{r}_{i1}\right),\left(\tau^{-2}\mathbf{I}+\sum_{i\in\mathcal{I}_{1t}}\mathbf{V}_i^{-1}\right)^{-1}\right)$$

$$p\left(\boldsymbol{\Sigma}_{gs}|\boldsymbol{\theta}_{gt}\right) \quad \propto \quad \left|\boldsymbol{\Sigma}_{gs}\right|^{-\frac{1}{2}\left(\sum_{t\in\mathcal{T}_{gs}}1\right)-\frac{v+M+1}{2}}\exp\left[-\frac{1}{2}\sum_{t\in\mathcal{T}_{gs}}\boldsymbol{\theta}_{gt}^{\mathsf{T}}\boldsymbol{\Sigma}_{gs}^{-1}\boldsymbol{\theta}_{gt}-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Psi}\boldsymbol{\Sigma}_{gs}^{-1}\right)\right]$$

$$g=2,\dots,G,\, s=1,\dots,S_g$$

$$\sim \quad \mathcal{IW}\left(v+\#\left\{\mathcal{T}_{gs}\right\},\boldsymbol{\Psi}+\sum_{t\in\mathcal{T}_{gs}}\left(\boldsymbol{\theta}_{gt}\boldsymbol{\theta}_{gt}^{\mathsf{T}}\right)\right)$$

## 3.8   Appendix C: Proof of $\sum_i \mathbf{H}_i = 1$

We would like to show that $\mathbf{H}1 = 1$, where $1$ is an $n \times 1$ vector where each element is equal to

1. Note that by construction, $1 \in \mathrm{col}\,(\mathbf{X})$, so we are able to write $1 = \mathbf{Xe}$ for some $p \times 1$ vector

$\mathbf{e}$. Then

$$\mathbf{H}1 = \mathbf{X} \left(\Omega^{-1} + \mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{Xe} \tag{3.1}$$

$$1 = \mathbf{Xe} = \mathbf{X} \left(\Omega^{-1} + \mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \left(\Omega^{-1} + \mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right) \mathbf{e} \tag{3.2}$$

It suffices to show that the difference $\mathbf{X} \left(\Omega^{-1} + \mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \Omega^{-1}\mathbf{e}$ between (3.1) and (3.2) ap-

proaches $0$.

We require the entries of $\Omega$ to be strictly positive with probability $1$, which is ensured

through the Inverse-Gamma specification for the variance components. Assume that we have

one fixed effect, whose corresponding entries in $\Omega$ approach $\infty$. We show that under this

condition, $\Omega^{-1}\mathbf{e} \to 0$, which would complete the proof. Note that having one fixed effect is a

consequence of the model, as $\mathbf{H}$ would not be well-defined with two or more fixed effects. Par-

tition $\Omega = \begin{bmatrix} \Omega_1 & 0 \\ 0 & \Omega_2 \end{bmatrix}, \Omega_1 = \tau^2 \mathbf{I}$, where $\tau^2$ is the variance of the diffuse Normal prior used for

the fixed effect. Similarly, we can partition $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where the columns of $\mathbf{X}_1$ correspond

to the fixed effect terms in $\theta$. This partition allows us to write $\mathbf{Xe} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \\ 0 \end{bmatrix} = 1$ for

some vector $\mathbf{e}_1$, since $1 \in \text{col}(\mathbf{X}_1)$. That is, the elements in $\mathbf{e}$ corresponding to random effects are all 0.

Then

$$\boldsymbol{\Omega}^{-1}\mathbf{e} = \begin{bmatrix} \boldsymbol{\Omega}_1^{-1} & 0 \\ 0 & \boldsymbol{\Omega}_2^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 \\ 0 \end{bmatrix} = \begin{bmatrix} \tau^{-2}\mathbf{I}\mathbf{e}_1 \\ 0 \end{bmatrix} \to 0 \text{ as } \tau^2 \to \infty$$

## 3.9 Appendix D: Log Pseudo-Marginal Likelihood

Let $\boldsymbol{\theta}_{(-i)}$ be the set of $\theta_{gt}$ parameters without the substate by year random effect corresponding to domain $i$, and let $\mathbf{y}_{(-i)}$ be the data without $y_i$, so that

$$p\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}_{(-i)}\right) \propto \frac{p\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}\right)}{p\left(y_i|\boldsymbol{\theta}_{(-i)}\right)}$$

where $y_i|\boldsymbol{\theta}_{(-i)} = y_i|\boldsymbol{\theta} - \theta_{g't} \sim \mathcal{N}\left(\sum_{g \neq g'} \theta_{g,t(i,g)}, v_i + \sigma_{g's}^2\right)$, where $g'$ is the substate by year interaction random effect group. Suppose we let $p_u\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}_{(-i)}\right) = \frac{p\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}\right)}{p\left(y_i|\boldsymbol{\theta}_{(-i)}\right)}$ so that $p_u\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}_{(-i)}\right)p\left(y_i\right) = p\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}_{(-i)}\right)$. Then we know $p\left(\boldsymbol{\theta}_{(-i)}|\mathbf{y}_{(-i)}\right)$ up to a proportionality

constant, and can use self-normalized importance sampling to get, for a new $\tilde{y}_i$,

$$
\begin{aligned}
p\left(\tilde{y}_i|\mathbf{y}_{(-i)}\right) &= \int p\left(y_i|\mathbf{\theta}_{(-i)}\right) p\left(\mathbf{\theta}_{(-i)}|\mathbf{y}_{(-i)}\right) d\mathbf{\theta}_{(-i)} \\[2ex]
&\doteq \frac{\sum_{t=1}^{T} p\left(y_i|\mathbf{\theta}_{(-i)}^{(t)}\right) \frac{p_u\left(\mathbf{\theta}_{(-i)}^{(t)}|\mathbf{y}_{(-i)}\right)}{p\left(\mathbf{\theta}_{(-i)}^{(t)}|\mathbf{y}\right)}}{\sum_{t=1}^{T} \frac{p_u\left(\mathbf{\theta}_{(-i)}^{(t)}|\mathbf{y}_{(-i)}\right)}{p\left(\mathbf{\theta}_{(-i)}^{(t)}|\mathbf{y}\right)}} \\[2ex]
&= \frac{\sum_{t=1}^{T} p\left(y_i|\mathbf{\theta}_{(-i)}^{(t)}\right) \frac{1}{p\left(y_i|\mathbf{\theta}_{(-i)}^{(t)}\right)}}{\sum_{t=1}^{T} \frac{1}{p\left(y_i|\mathbf{\theta}_{(-i)}^{(t)}\right)}} \\[2ex]
&= T \left[\sum_{t=1}^{T} \frac{1}{p\left(y_i|\mathbf{\theta}_{(-i)}^{(t)}\right)}\right]^{-1}
\end{aligned}
$$

where $t$ represents draws from the posterior distribution of $\mathbf{\theta}_{(-i)}$.

## 3.10    Appendix E: Alternative Decision Rule

In this "opportunist" method, we seek which set of direct estimates (state or substate area) are closest to the Bayesian estimates. This allows us to capitalize on chance for the data in the given year. In one year the substate estimates could happen to be close to the Bayesian estimates, and subsequently reported, but in another year they could be far from the Bayesian estimates and not selected for reporting.

We are looking to estimate the MSE of the state or substate direct estimates, given the current data. For the substate direct estimate $y_i$, this is

$$\mathbb{E}_{\eta_i|y_i}\left[(y_i - \eta_i)^2\right] = \left(\mathbb{E}_{\eta_i|y_i}[y_i - \eta_i]\right)^2 + Var_{\eta_i|y_i}(y_i - \eta_i)$$
$$= (y_i - \hat{\eta}_i)^2 + \widehat{Var}(\eta_i)$$

where $\hat{\eta}_i$ is the posterior mean and $\widehat{Var}(\eta_i)$ is the posterior variance. Similarly, for the state direct estimate $y_{i\cdot}$, the MSE is

$$\mathbb{E}_{\eta_i|y_i}\left[(y_{i\cdot} - \eta_i)^2\right] = (y_{i\cdot} - \hat{\eta}_i)^2 + \widehat{Var}(\eta_i)$$

Then the MSE for each direct estimate is calculated by summing across areas within a state, to produce a value for each state and year. A decision rule can then be formulated by identifying the lower MSE among the pooled or unpooled estimator, for each state and year.

# A Hospital Profiling Approach to Determine Association Between Palliative Care and Aggressiveness of End-of-Life Treatments

Tianyi Cai
Department of Biostatistics
Harvard University

Sherri Rose
Department of Health Care Policy
Harvard Medical School

Deborah Schrag
Department of Medical Oncology
Dana Farber Cancer Institute

Francesca Dominici
Department of Biostatistics
Harvard University

## 4.1  Introduction

Hospital profiling has a widespread impact on government, insurers, physicians, and patients (Gatsonis, 1998; Normand, 2005; Normand & Shahian, 2007). Methods used to profile hospitals are not standard for a myriad of reasons. First, hospital profiling needs to account for both patient and hospital-level characteristics, as well as the correlation between outcomes of patients at the same hospital. This necessitates the use of multi-level models (Gelman & Hill, 2006). Second, an adjustment for case-mix bias at the patient level is necessary to ensure that hospitals are not profiled negatively by virtue of treating sicker individuals (Salem-Schatz et al., 1994; Shahian & Normand, 2008; Ash et al., 2012). Lastly, hospital size can vary substantially, and methods need to account for the varying size of each hospital unit.

Traditionally, hospitals are profiled based on a binary outcome of interest that characterizes patient health, such as estimated mortality or morbidity rates. Hierarchical models have long been recommended to account for the nested structure of the data (Goldstein & Spiegelhalter, 1996; Normand et al., 1997; Leyland & Boddy, 1998). Typically, there is a case-mix adjustment using patient characteristics at the first level of the model, and hospital characteristics included in the second level of the model to obtain a standardized outcome for each hospital. Frequentist methods generally use hypothesis testing or confidence intervals for ratios or differences of observed and expected outcome rates. Common models include logistic regression, where hospital effects can be specified as fixed (using indicator variables), random (using multi-level

78

models), or not specified (DeLong et al., 1997). Bayesian methods for hospital profiling include Bayesian hierarchical models (Normand et al., 1997; Shen & Louis, 1998; Racz & Sedransk, 2010) and empirical Bayes (Thomas et al., 1994).

Classical methods to adjust for measured confounding include propensity score matching (Rosenbaum & Rubin, 1983; Austin, 2011), inverse probability weighting (Robins et al., 2000; Joffe et al., 2004), and doubly robust estimators (Bang & Robins, 2005; Kennedy et al., 2016). Varewyck et al. (2014) and Shahian & Normand (2015) have structured the hospital profiling question within a causal inference framework. However, all of these methods estimate an average causal effect at the individual (patient) level, and have not been developed in the context of nested data where the focus is on estimation of a causal effect at the second (hospital) level of the hierarchical model. In our scenario, we need the first level to provide the patient case-mix adjustment, while our scientific interest lies in hospital-level associations at the second level, but not necessarily causal associations.

In this chapter, we use a Bayesian hierarchical model in the spirit of Normand et al. (1997) to estimate the hospital-level association between palliative care received and the aggressiveness of end-of-life treatments. At the first stage of the model, we provide a patient-level case-mix adjustment. We include a propensity score adjustment at the second level of the hierarchical model, modeling the receipt of palliative care treatment as either binary, or continuous using a normal and quasi-binomial propensity function (Imai & Van Dyk, 2004). Our methods are applied to a dataset of deceased Medicare FFS beneficiaries diagnosed with brain, lung, colon, or pancreatic cancer.

The remainder of this chapter is organized as follows. In Section 4.2, we descibe the working data, study population, our hospital-level definition of palliative care received, and confirm the existence of confounding in our working data. Section 4.3 includes our Bayesian hierarchical models used to analyze our data. Section 4.4 contains the results of our data analysis, which are discussed with concluding remarks in Section 4.5.

## 4.2 Data

Our data are collected from Medicare inpatient and post-acute care claims, and medicare enrollment files. This includes a cohort of Medicare FFS beneficiaries 65+ discharged with principal discharge diagnosis code of lung (162.xx), pancreas (157.xx), colorectal (153.xx), or brain (191.xx) cancers (from Medicare Part A claims), with a 2-year follow-up period. From there, we applied the inclusion and exclusion criteria presented in Table 4.1.

**Table 4.1:** Inclusion and Exclusion Criteria

|     Inclusion     |     Exclusion     |
| --- | --- |

- principal diagnosis code (and confirmation that these are new diagnoses) for the aforementioned cancers in 2009

- FFS beneficiaries only

- enrollment in FFS at least 12 months prior to first condition-specific admission

- continuous enrollment in Medicare after diagnosis, at least 12 months of follow-up after diagnosis

- discharged from a short-term acute care hospital

- beneficiaries who resided in 50 states plus Puerto Rico and DC

- in-hospital biopsy/resection

- initial or first condition-specific hospitalization

- receipt of chemotherapy within 120 days of diagnosis

- continuous enrollment in Medicare FFS plan after initial diagnosis

- metastatic cancer patients only (second malignant neoplasm ICD-9 code)

- diagnosis of any other cancer

- receipt of prior major cancer surgery

- failed to merge with denominator files

- transferred to another acute-care hospital

- in-hospital death

- unreliable death data

- discharged from a critical access hospital

- receipt of prior major cancer surgical procedures at any time during the study period

- patients who had another cancer within 3 years prior to the initial cancer diagnosis

## 4.2.1 Study Population

As part of the inclusion criteria, patients from the data were selected based on having a first hospital discharge between January 2009 and December 2010, corresponding to initial hospital admission dates between October 2008 and December 2010, and death dates between January 2009 and December 2012. In order to define end-of-life outcomes and treatments for patients and hospitals, our study population only included deceased patients from hospitals with 30 or more discharged patients. Figure 4.1 shows the admission, discharge, and death dates for a random sample of 100 patients in our study population of 20,400 patients over 408 hospitals.

**Figure 4.1:** Discharge, Admission, and Death Dates for Study Population
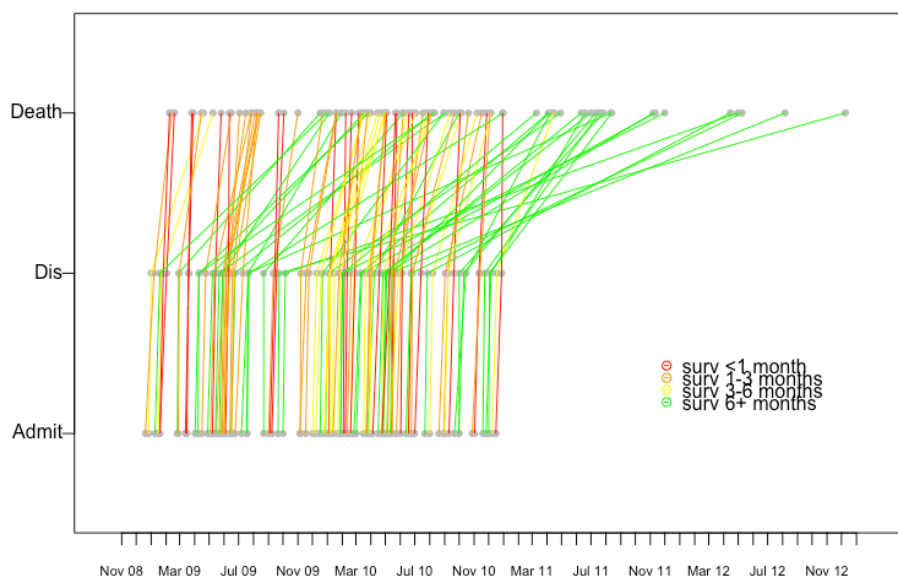
We are interested in various binary outcomes related to the aggressiveness of end-of-life care, as shown in Table 4.2. Due to the low prevalence of chemotherapy and radiotherapy within 30 days of death, the outcome for our analysis is a binary indicator of receiving either of these two forms of EOL care within 30 days of death. Chemotherapy and radiotherapy can begin before or after palliative care interventions. We are concerned with whether chemotherapy and radiotherapy continued through to end-of-life, defined as within 30 days of death.

**Table 4.2:** End-of-Life Outcomes of Interest

| Outcome | Prevalence |
|---|---|
| Chemotherapy within 30 days of death | 0.103 |
| Radiotherapy within 30 days of death | 0.114 |
| Hospital readmission within 30 days of discharge | 0.229 |

Prevalence is based on number of patients with the outcome in our study population.

## 4.2.2  Defining Palliative Care Received

Palliative care is a specialized form holistic health care targeted at patients with severe illnesses such as cancer, cardiac disease, and psychiatric disorders. The primary intent of palliative care is to improve the quality of life by assuaging physical and mental pain stemming from the disease (Rome et al., 2011). Palliative care can begin at any stage of a serious illness, and can co-exist with traditional curative treatment, or act as a segue to hospice for incurable illnesses.

Patient-level palliative care encounters are defined by the ICD-9 code V66.7 for "encounter for palliative care." For hospitals defined by their six digit MEDPAR provider number, we define the hospital-level receipt of palliative care as $a_i = \frac{p_i}{n_i}$, where $p_i$ is the number of patients discharged by hospital $i$ in 2009 or 2010 who received palliative care (based on the ICD-9 code) from hospital $i$ during the 2-year follow-up period, and $n_i$ is the number of patients discharged by hospital $i$ in 2009 or 2010. We can then threshold the level of palliative care received to get a binary treatment $PC_i = \mathbb{I}(a_i > \omega)$ for each hospital $i$.

The amount of palliative care received is plotted for the 408 hospitals in Figure 4.2. We see a high amount of variation in the geographic location of the hospitals, the amount of palliative care by geographic area, and the size of hospitals by geographic area. In general, there appears to be more palliative care received in the northeast and midwest.
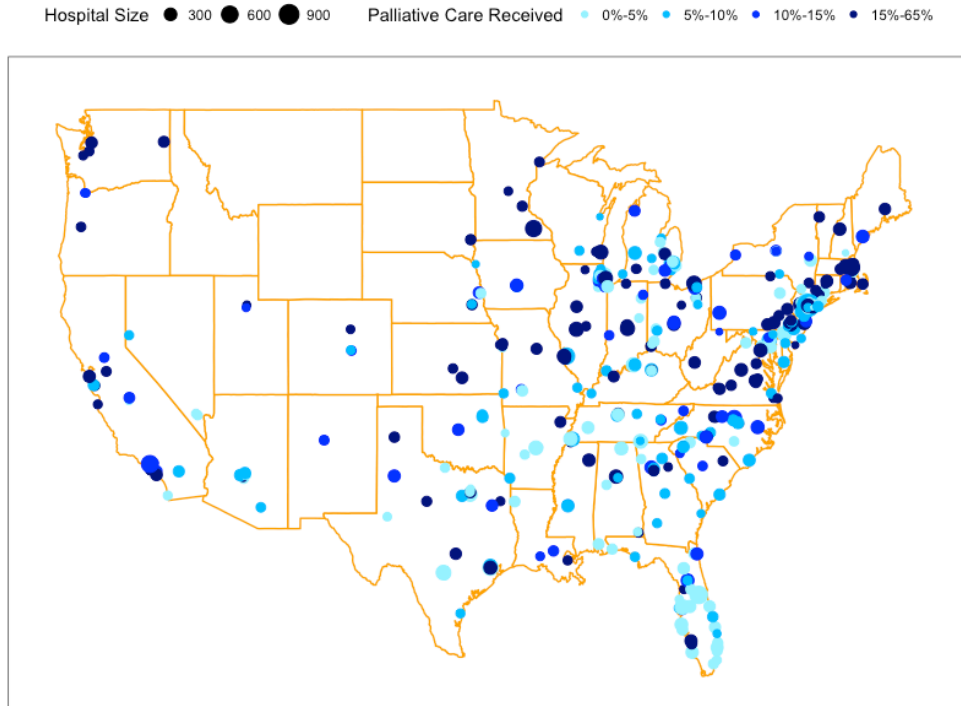
## 4.2.3   Assessment of Confounding using Binary Treatment

We define the treated population as the 203 hospitals with palliative care received $a_i$ greater

than $\omega = 0.1$, and the control population as the remaining 205 hospitals with $a_i \leq 0.1$. We

chose 0.1 as a starting value for $\omega$ since it balances the number of treated and untreated hos-

pitals. To look for possible sources of confounding, we looked at the balance of potential

confounders in the treated and untreated groups before and after propensity score adjustment

(Rosenbaum & Rubin, 1983), where the propensity score model is fitted with all the available

hospital-level covariates. For each covariate $z$, we can compute the absolute standardized dif-

ference (ASD) between the treated (trt) and control (ctr) populations as

$$d = \frac{\left| \bar{z}_{\text{trt}} - \bar{z}_{\text{ctr}} \right|}{\sqrt{\frac{(n_{\text{trt}}-1)s_{\text{trt}}^2 + (n_{\text{ctr}}-1)s_{\text{ctr}}^2}{n_{\text{trt}}+n_{\text{ctr}}-2}}}$$

The ASDs before propensity score adjustment are sorted in descending order and represented by the red dots in Figure 4.3, so that the most unbalanced covariates are at the top of the figure. For propensity score adjustment, we considered matching, and stratification. The ASDs after propensity score adjustment are plotted in Figure 4.3, where the green dots represent 1 to 1 matching based on log odds of the propensity score, and the blue dots represent stratifying based on quartiles of the propensity score. In the latter case, the ASDs in each quartile are calculated and averaged based on the size of the quartiles.

**Figure 4.3:** Absolute Standardized Differences of Hospital Covariates before and after Propensity Score Matching
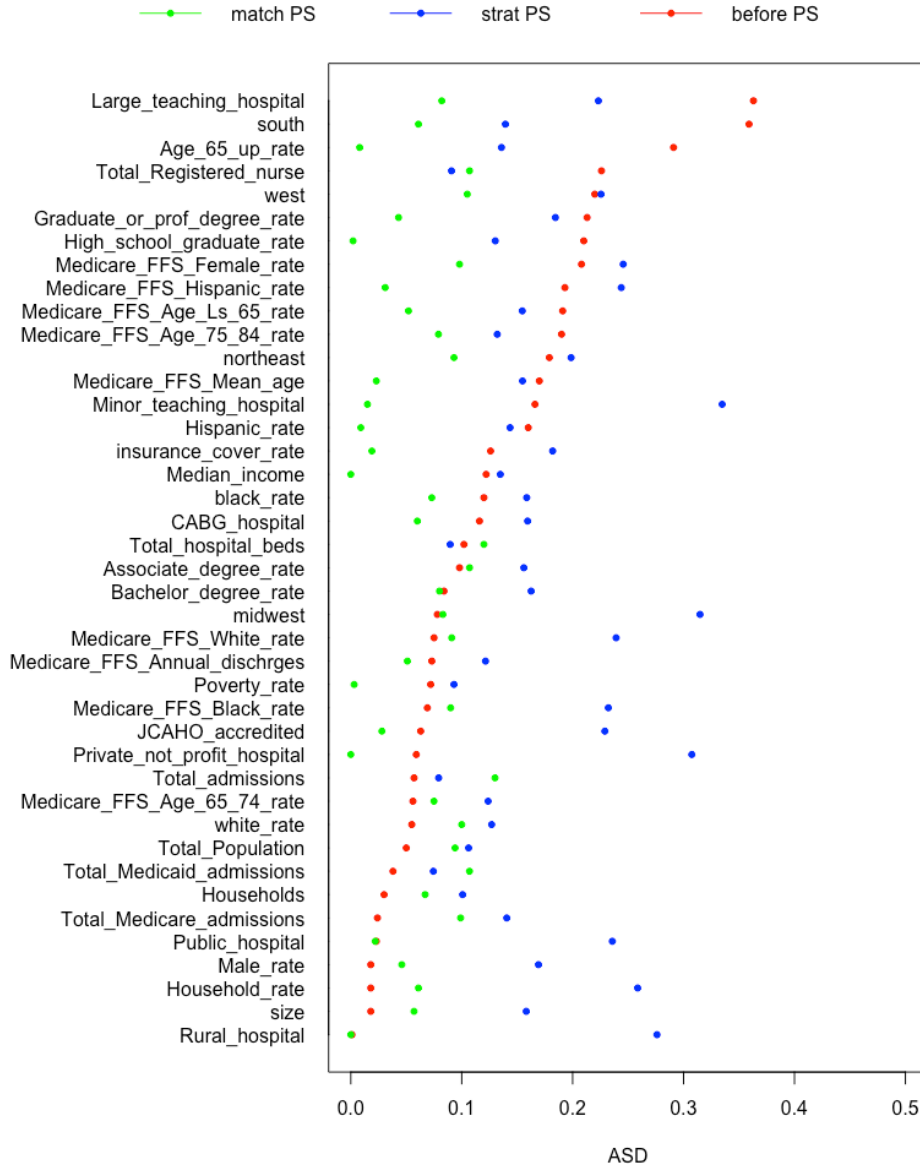
## 4.2.4 Assessment of Confounding using Continuous Treatment

As an alternative, we can directly use the palliative care received $a_i$ as a continuous treatment. Imai & Van Dyk (2004) showed that a uniquely parameterized propensity function $p(a_i|\mathbf{z}_i)$ satisfies the properties of a balancing score and the strong ignorability of treatment assignment assumption. We consider a propensity function

$$p(a_i|\mathbf{z}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(a_i - \boldsymbol{\gamma}^T\mathbf{z}_i\right)^2\right]$$

For each covariate, calculating the ASD between two populations is akin to testing the significance of a binary treatment on that covariate. With a continuous treatment, we are not able to use the ASD. Instead, to assess the balance of covariates, we proceed as in Imai & Van Dyk (2004) by regressing each covariate on $a_i$, using logistic regression for binary covariates and linear regression for continuous ones. Figure 4.4 shows the standard normal quantile plot of the 34 $t$-statistics (362 degrees of freedom). There is a lack of balance in the covariates due to the extreme values of the $t$-statistics shown in the tails, indicating strong association between the treatment and certain covariates. After adjusting for the propensity score $\hat{a}_i = \hat{\boldsymbol{\gamma}}^T\mathbf{z}_i$, we see $t$-statistics closer to 0, indicating balance after adjustment.

**Figure 4.4:** Quantile Plots of $t$-Statistics for the Coefficient of $a_i$
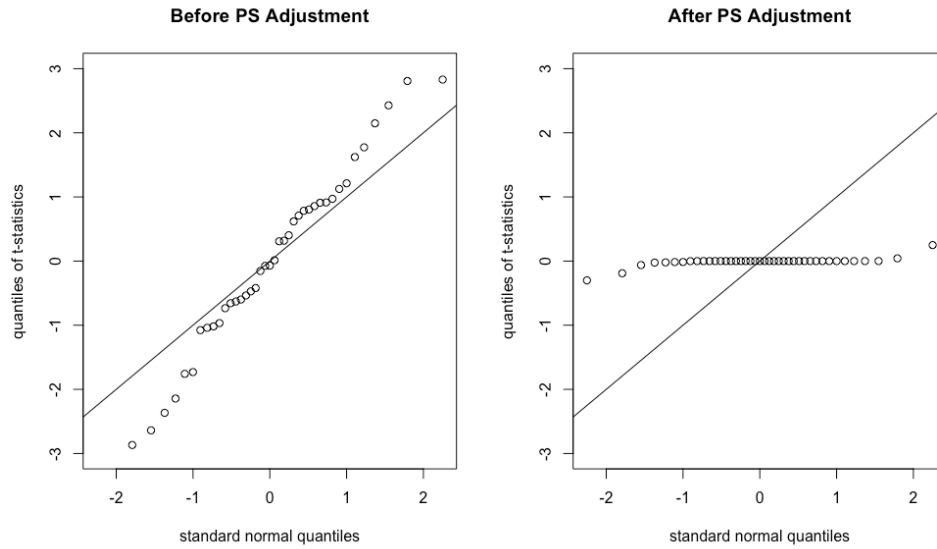


## 4.3   Methods

We consider two methods that differ based on the modeling of the palliative care received variable $a_i$. First, for a binary treatment, we extend the random intercept Bayesian hierarchical model proposed by Normand et al. (1997) to incorporate a propensity score adjustment at the second level of the model, with a sensitivity analysis on the threshold for defining the binary treatment. Second, we use the same Bayesian hierarchical model, but with a continuous treatment and a propensity function, as defined by Imai & Van Dyk (2004). This amounts to a comparison between four Bayesian hierarchical models differing based on adjustment for confounding: one with regression adjustment, one with propensity score adjustment, and two

89

different propensity functions for adjustment.

## 4.3.1   Model Specification

The Bayesian hierarchical model makes a strong assumption that we can adjust for confounding using a linear regression model, and we describe its most general form as follows. At the first level of the hierarchical model, for hospital $i$ and patient $j$, assume the outcome $y_{ij} \sim \text{Bern}\left(p_{ij}\right)$, with

$$\text{logit}\left(p_{ij}\right) = \alpha_i + \boldsymbol{\beta}^T \mathbf{x}_{ij}, \boldsymbol{\beta} \in \mathbb{R}^P, i = 1, \ldots, I, j = 1, \ldots, n_i$$

The patient-level covariates $\mathbf{x}_{ij}$ provide the necessary case-mix adjustment to ensure that the $\alpha_i$ represent the hospital-specific random intercepts, after adjusting for patient-level case-mix bias. In our data, the case-mix adjustment is performed using a univariate estimated severity of sickness (details in Appendix C). At the second level of the hierarchical model, we specify priors

$$
\begin{aligned}
\alpha_i &\sim \text{skew-}t\left(\mu = \theta_0 + \theta_a g\left(a_i\right) + \sum_{q=1}^{Q} \theta_q z_{qi}, \nu, k, \gamma\right), i = 1, \ldots, I \\
\boldsymbol{\beta} &\sim \mathcal{N}\left(0, \tau_\beta^2 \mathbf{I}_P\right), \tau_\beta = 10^3
\end{aligned}
$$

The $z_{1i}, \ldots z_{Qi}$ are hospital-level confounders associated with $\alpha_i$, and our coefficient of interest is $\theta_a$. Receipt of palliative care is $a_i$, where $g\left(a_i\right) = \mathbb{I}\left(a_i > \omega\right)$ for models using a binary treat-

90

ment, and $g(a_i) = a_i$ for models with continuous treatment. Instead of a normal distribution,

we specify a more flexible skew-$t$ distribution (Fernández & Steel, 1998; Lee & Thompson, 2008)

for $\alpha_i$, where the normal distribution is a special case with $k = \infty$, $\gamma = 1$. This specification

allows for heavier tails, skewed random effects, and allows us to check the quality of model

fit for the random effects $\alpha_i$. The final level of the hierarchical model specifies priors

$$\theta_0, \theta_a, \theta_q \quad \sim \quad \mathcal{N}\left(0, \tau_\theta^2\right), \tau_\beta = 10^3, q = 1, \dots, Q$$

$$v \quad \sim \quad \mathcal{IG}\left(a, b\right), a, b = 10^{-3}$$

We include a propensity score adjustment at the second level of the hierarchical model, namely

$$\alpha_i \sim \text{skew-}t\left(\mu = \theta_0 + \theta_a g(a_i) + h(\text{PS}_i) + \sum_{q=1}^{Q} \theta_q z_{qi}, v, k, \gamma\right), i = 1, \dots, I$$

for some function $h(\cdot)$ of the propensity score $\text{PS}_i$. Details of the derivation of the posterior

distributions are presented in Appendix A, along with information about the MCMC sampler

in Appendix B.

In order to have a meaningful analysis, we had to choose a threshold for the minimum

hospital size to include in our working data. A low threshold meant more hospitals in the

data, providing more information to estimate hospital-level effects, but it is not informative to

include hospitals with only a few patients (a large number of hospitals contained exactly one

patient). Conversely, a high threshold meant many patients in each hospital, but not many

data points to estimate hospital-level effects.

## 4.3.2 Binary Treatment

We define the binary palliative care treatment as $PC_i = \mathbb{I}(a_i > \omega)$ for hospital $i$ with palliative care received $a_i$. To begin, we set $\omega = 0.1$ to achieve an approximately equal proportion of treated and control hospitals, and a sensitivity analysis is conducted for various levels of $\omega$. In the case of a binary treatment, we use a classic propensity score model $\text{logit}(PS_i) = \gamma^T v_i$ for confounders $v_i$ and $PS_i = \mathbb{P}(PC_i = 1) = \mathbb{P}(a_i > \omega)$. At the second level of the hierarchical model, we can adjust for propensity score by including it directly, or by including indicators of propensity score strata.

## 4.3.3 Continuous Treatment

We first consider a normal propensity function defined by

$$a_i | z_i \sim \mathcal{N}\left(\gamma^T z_i, \sigma_a^2\right),$$

where $\mathbb{E}(a_i | z_i) = \hat{\gamma}^T z_i$ can be estimated by linear regression. By the uniquely parameterized propensity function assumption (Imai & Van Dyk, 2004), we can perform propensity score adjustment using $\hat{\gamma}^T z_i$, which characterizes the normal propensity function. One drawback to this method is that $a_i$ is bounded between $[0, 1]$, and may not be suited for a normal distribution.

We thus consider a quasi-binomial propensity function defined by

$$a_i | \mathbf{z}_i \sim \text{Bern}\left(\pi_i\right),$$

where $\mathbb{E}\left(a_i | \mathbf{z}_i\right) = \pi_i$, $Var\left(a_i | \mathbf{z}_i\right) = \varphi \pi_i \left(1 - \pi_i\right)$, and $\varphi$ is the dispersion parameter to account for the fact that $a_i$ is not binary. In this case, we can perform propensity score adjustment through $\hat{\pi}_i$ estimated by logistic regression with $a_i$ as the independent variable.

### 4.3.4   Model Comparison

As our primary interest is at the hospital-level, we would like to examine the adequacy of model fit for the hospital random effects $\alpha_i$. We consider posterior predictive checks (Gelman et al., 1996) to compute a posterior predictive $p$-value for each hospital using the test statistic $y_{i+} = \sum_{j=1}^{n_i} y_{ij}$ (Venturini et al., 2017). A small $p$-value for $y_{i+}$ would correspond to hospitals whose observed data is not predicted well by the model, along with possible misspecification of the hospital random effects distribution for $\alpha_i$.

The model comparison portion contains two steps. In the first step, we only consider a normal specification for the prior of $\alpha_i$, which is equivalent to fixing $k = \infty$, $\gamma = 1$ in the skew-$t$ prior. Posterior predictive checks are used to select a model out of the four proposed in Section 4.3, based on having the fewest small $p$-values among the hospitals. Secondly, the best-performing model from the first step is re-run 9 times with $k = \{10, 50, \infty\}$ and $\gamma = \{0.8, 1, 1.25\}$, and posterior predictive checks are performed on those set of models to select

values for $k$ and $\gamma$.

## 4.4  Application to Medicare Data

The hospital random intercepts $\alpha_i$ are interpreted as the log odds of receiving the EOL outcome for a patient at hospital $i$ with average severity, which can be viewed as a standardized risk of the EOL outcome. Analyses are provided for all EOL outcomes presented in Table 4.2. In all analyses, the parameter of interest is $\theta_a$, indicating the association (after adjusting for hospital-level confounders) between palliative care received and $\alpha_i$.

### 4.4.1  Using a Normal Prior for $\alpha_i$

We applied four Bayesian hierarchical models with a normal prior on $\alpha_i$ to the Medicare dataset of 20,400 patients nested in 408 hospitals. The EOL outcome in this case is chemotherapy or radiotherapy within 30 days of death. The results are summarized in Table 4.3.

**Table 4.3:** Results with Normal Prior for $\alpha_i$

| Model | Model Specification | $\hat{\theta}_a$ | 95% Credible Interval |
|---|---|---|---|
| Regression | $\omega = 0.10$ | -0.052 | (-0.126, 0.016) |
| | $\omega = 0.05$ | -0.035 | (-0.121, 0.043) |
| | $\omega = 0.10$ | -0.046 | (-0.126, 0.026) |
| Binary Treatment | $\omega = 0.15$ | -0.064 | (-0.129, -0.003) |
| | $\omega = 0.20$ | -0.048 | (-0.118, 0.016) |
| | $\omega = 0.25$ | -0.027 | (-0.141, 0.081) |
| Continuous Treatment | Normal | -0.054 | (-0.130, 0.005) |
| | Quasi-Binomial | -0.048 | (-0.117, 0.011) |

The parameter of interest is the log odds $\theta_a$, indicating the association between palliative care received and $\alpha_i$. Across all models, the posterior mean of $\theta_a$ ranged from -0.027 to -0.064 odds ratios of 0.97 and 0.94, respectively. For most models, the 95% credible interval for $\theta_a$ contains 0. The sensitivity analysis on $\omega$ showed minor differences when $\omega = 0.10, 0.15, 0.20$, and more significant differences when $\omega = 0.05, 0.25$, mostly due to the imbalance of treated and controls for those extreme thresholds; $\omega = 0.05$ corresponds to 73% of hospitals receiving palliative care, while $\omega = 0.25$ corresponds to 14% receiving palliative care. Based on the normal prior for $\alpha_i$, there does not seem to be a strong association between palliative care received and EOL chemotherapy/radiotherapy.

## 4.4.2   Model Comparison Stage 1

Posterior predictive checks were performed on the test statistic $y_{i+} = \sum_{j=1}^{n_i} y_{ij}$. The model with binary treatment, propensity score adjustment, and $\omega = 0.1$ generally had the fewest low $p$-values, indicating the fewest number of hospitals with potentially misspecified hospital-level random effects distributions for $\alpha_i$. Figure 4.5 shows scatterplots between the $p$-values for this base model and all other candidate models.

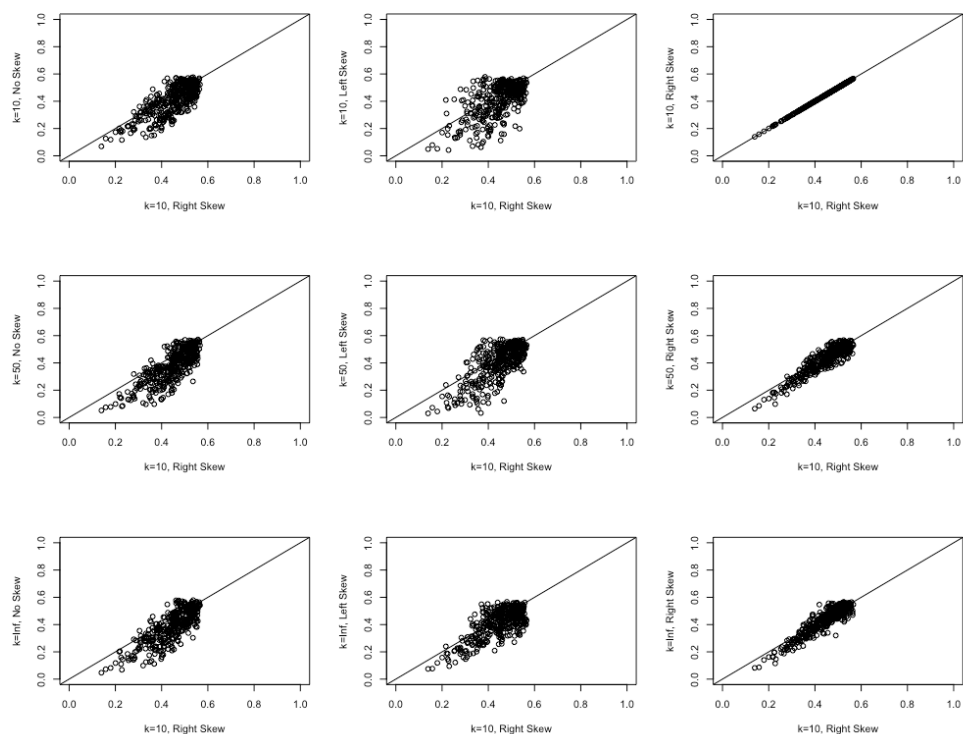**Figure 4.5:** Comparison of Posterior Predictive $p$-values
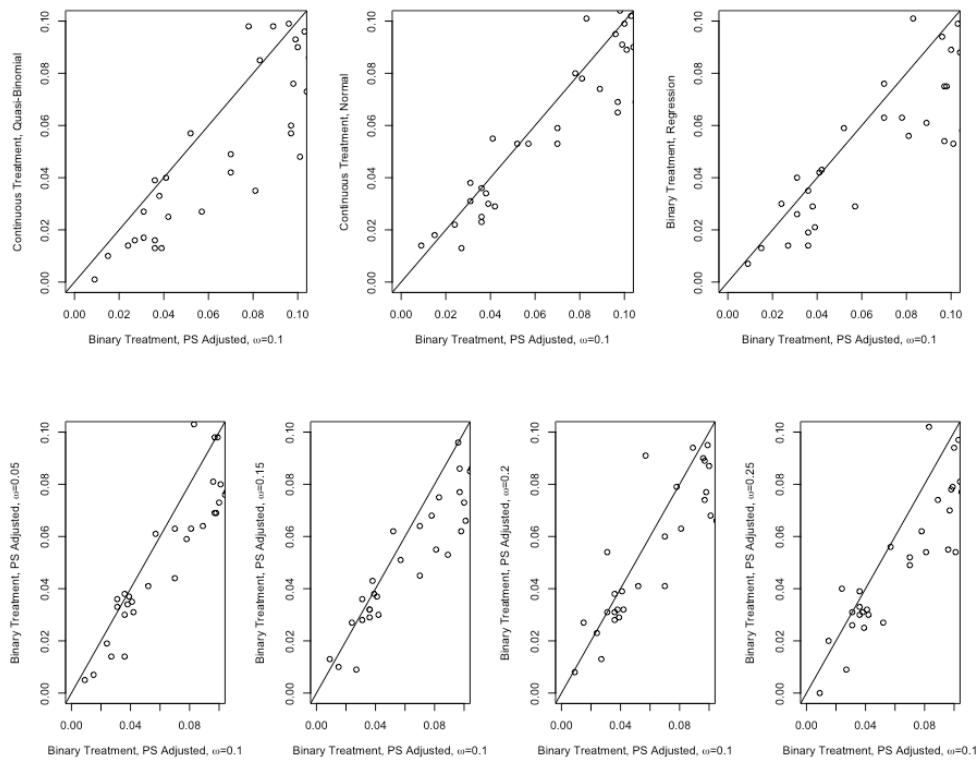


The scatterplots only consider low posterior predictive $p$-values below 0.1, as seen in the

axes. Points below the diagonal line in the scatterplots indicate hospitals that had an undesirably lower posterior predictive $p$-value in the candidate model being compared against our base model. We proceed with this base model by varying the parameters $k$ and $\gamma$ to specify a more flexible skew-$t$ prior for $\alpha_i$.

### 4.4.3 Model Selection Stage 2: Using a Skew-$t$ Prior for $\alpha_i$

A similar model comparison was made between our base model (binary treatment, propensity score adjustment, and $\omega = 0.1$) and 9 other models specified by $k = \{10, 50, \infty\}$ and $\gamma = \{0.8, 1, 1.25\}$. The values for $\gamma$ correspond to left skew, no skew, and right skew, respectively. Once again, posterior predictive checks were performed on the test statistic $y_{i+} = \sum_{j=1}^{n_i} y_{ij}$. Posterior predictive checks strongly favor the model with $k = 10$, $\gamma = 1.25$, indicating a heavy-tailed right-skew prior for $\alpha_i$, which is reasonably consistent with the exploratory data analysis. This indicates that there are a number of hospitals with high log odds of chemotherapy or radiotherapy within 30 days of death, after adjusting for patient-level case-mix bias. Figure 4.6 shows scatterplots between the $p$-values for various parameters of $k$ and $\gamma$.

**Figure 4.6:** Comparison of Posterior Predictive *p*-values



Based on these posterior predictive checks, we proceed with the final model using a right-skew-*t* prior on $\alpha_i$ with $k = 10$, $\gamma = 1.25$. For additional comparison, a regression-only version (without propensity score adjustment) of the final model is included, as well as another regression-only version of the final model using categorical variables for palliative care based on its quartiles.

### 4.4.4　Final Model

Credible intervals lying entirely above or below zero are highlighted in green, the credible interval for $\theta_a$ is highlighted in blue, and the remaining intervals are highlighted in red. Results from our final model are shown in the middle column of Figure 4.7, where the outcome of interest is chemotherapy or radiotherapy within 30 days of death. Also shown in Figure 4.7 are results from the final model applied the other EOL outcome of interest: hospital readmission within 30 days of discharge. Figure 4.8 presents results from the two regression-only models (without propensity score adjustment). The first column presents results from the regression-only final model with a binary indicator for palliative care, while the second column presents results from the regression-only final model with an ordinal palliative care variable based on quartiles.

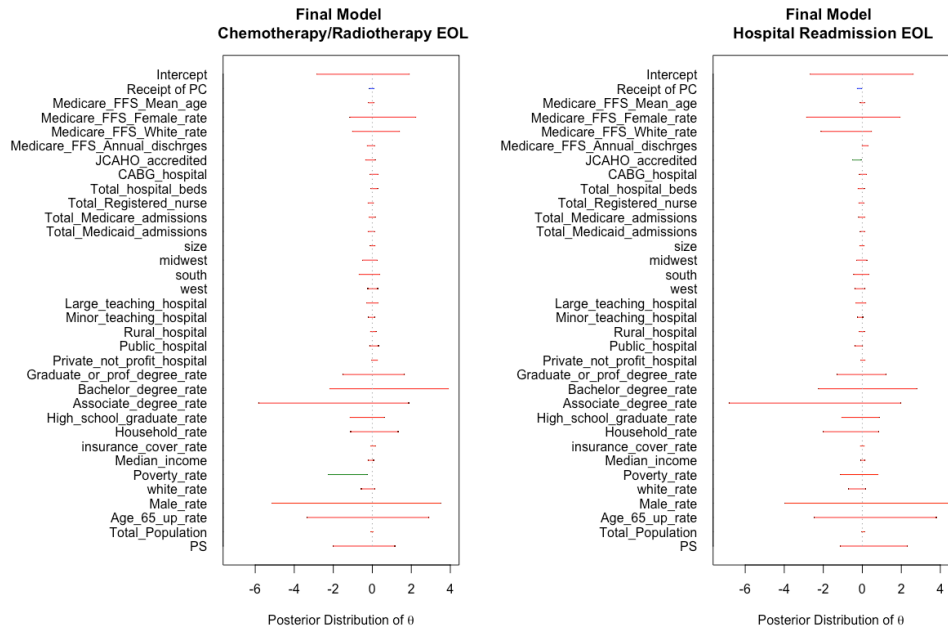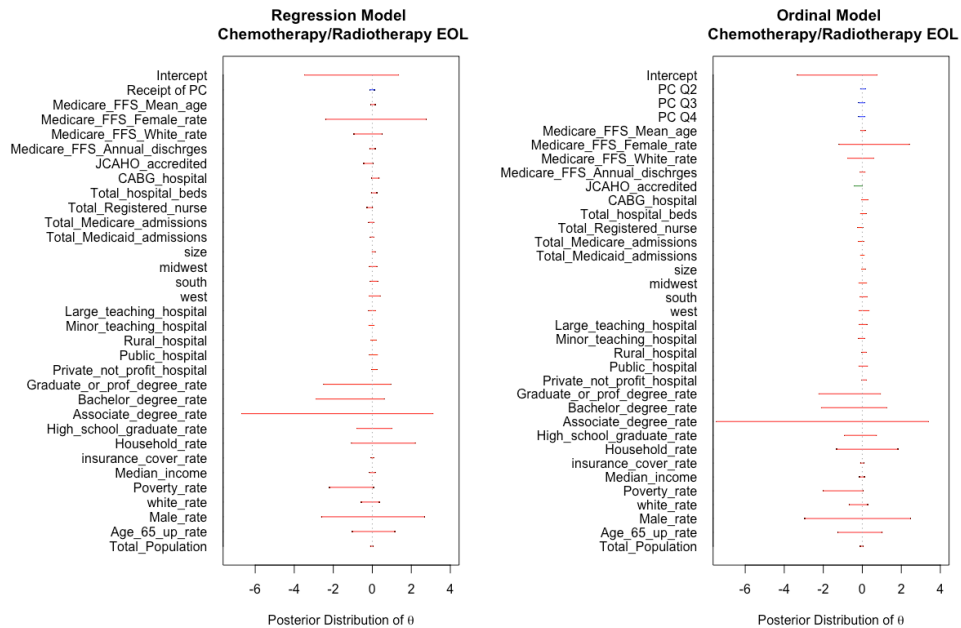**Figure 4.7:** 95% Credible Intervals for Hospital-Level Covariates, Models with PS



**Figure 4.8:** 95% Credible Intervals for Hospital-Level Covariates, Models without PS



100

A more detailed summary of the posterior distribution of $\theta_a$ along with all green intervals is presented in Table 4.4.

**Table 4.4:** Summary of Hospital-Level Covariates of Interest

| | Covariate | Posterior Mean $\theta_a$ | Description |
|---|---|---|---|
| Regression Model | | | |
| | Palliative Care Received | -0.008 (-0.110, 0.116) | having receipt of palliative care over 0.1 |
| Chemo/Radio EOL | | | |
| | Palliative Care Q2 | 0.026 (-0.089, 0.176) | 2nd quartile of palliative care |
| Ordinal Model | Palliative Care Q3 | -0.038 (-0.162, 0.129) | 3rd quartile of palliative care |
| Chemo/Radio EOL | Palliative Care Q4 | -0.027 (-0.185, 0.124) | 4th quartile of palliative care |
| | JCAHO Hospital | -0.199 (-0.391, -0.031) | Joint Commission accredited |
| Final Model | Palliative Care Received | -0.022 (-0.142, 0.085) | having receipt of palliative care over 0.1 |
| Chemo/Radio EOL | Poverty Rate | -1.198 (-2.247, -0.270) | rate of poverty in hospital's zipcode |
| Final Model | Palliative Care Received | -0.135 (-0.256, -0.022) | having receipt of palliative care over 0.1 |
| Readmission EOL | JCAHO Hospital | -0.282 (-0.520, -0.049) | Joint Commission accredited |

Once again, the parameter of interest is the log odds $\theta_a$, indicating the association between palliative care received and $\alpha_i$. The regression models suggest no association between hospital-level palliative care received and EOL chemotherapy/radiotherapy. However, in the ordinal model, there is a very slight negative association between Joint Commission accreditation status and EOL chemotherapy/radiotherapy. Among the final models, hospital-level palliative

care does not appear to be associated with EOL chemotherapy/radiotherapy, but does seem to have a slight negative association with EOL hospital readmissions. Two confounders that stand out in these final models include the poverty rate and the JCAHO status, affecting EOL chemotherapy/radiotherapy and EOL hospital readmissions, respectively.

## 4.5   Discussion

We extended the work done by Normand et al. (1997) by introducing a causal framework at the second (hospital) level of the Bayesian hierarchical model. This allowed us to adjust for patient-level case-mix bias at the first level of the model, and to adjust for hospital-level confounding at the second level of the model. We considered candidate models in two steps. In the first step, for a normal prior on $\alpha_i$, we looked at a regression only model, a binary treatment model with propensity score adjustment, and two continuous treatment models using propensity functions. Posterior predictive checks revealed the binary treatment model to have the best performance. In the second step, for the binary treatment model, we looked at left and right skew-$t$ distributions as the prior on $\alpha_i$, and found an ideal right-skewed heavy-tailed final model. Regression-only models based on this final model were provided for comparison purposes.

Estimating parameters for the final model revealed no significant association between receipt of palliative care and EOL chemotherapy/radiotherapy, and a small negative association between receipt of palliative care and hospital readmissions. Additionally, it revealed that the

poverty rate in a particular hospital's zipcode has a reasonably negative association with EOL chemotherapy/radiotherapy, which is consistent with various articles on the importance of considering poverty in conjunction with palliative care (Kumar, 2007; Kikule, 2003; Hughes, 2005). For hospital readmissions, the JCAHO status of a hospital had a small negative association. Including these confounders as interaction effects (with palliative care), or stratified analyses using these confounders can provide further insight.

The aggressiveness of EOL treatments alone is just one aspect of quality of life. One area not considered in this study is the impact of palliative care on the overall quality of life, which is related to the aggressiveness of EOL treatments, but contains other aspects such as patient surveys and mood outcomes. While it does not appear that implementation of palliative care seems to be significantly helpful in reducing the aggressiveness of EOL treatments, palliative care could have a significant impact on the overall quality of life (Temel et al., 2010).

Hospital-level associations were considered because relying on patient-level billing codes is not ideal, as doctors who provider services that could be defined as palliative care may classify those services under a more familiar billing code. Additionally, different doctors have varying opinions on how to classify services. These are the primary reasons for not performing an analysis at the patient level, since control groups could include patients who received palliative care but had it billed under a more common code. One caveat is that this definition does not account for patients having access to palliative care but not actually receiving it. To address these issues, we aggregate the patient-level palliative care into a hospital-level treatment. Assuming that the billing inconsistencies are reasonably stable across hospitals, a hospital-level

103

palliative care treatment is more robust to patient-level billing differences, and is more in line with our scientific question of interest.

Of additional consideration is the correlation between the multiple EOL outcomes (hospital chemotherapy, radiotherapy, hospital readmission, home death, hospice enrollment, etc.), which is not considered when looking at these outcomes separately. Jointly modeling these outcomes could yield more precise estimates, and could identify hospital-level covariates that affect all EOL outcomes. This work could naturally be extended to consider such multivariate outcomes.

## 4.6   Appendix A: MCMC Derivations

The likelihood function is

$$
\begin{aligned}
\mathcal{L}\left(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2\right) &= \prod_{i=1}^{I}\prod_{j=1}^{n_i} p_{ij}^{y_{ij}}\left(1-p_{ij}\right)^{1-y_{ij}} \\
&= \prod_{i=1}^{I}\prod_{j=1}^{n_i}\left[\frac{\exp\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)}{1+\exp\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)}\right]^{y_{ij}}\left[\frac{1}{1+\exp\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)}\right]^{1-y_{ij}} \\
&= \prod_{i=1}^{I}\prod_{j=1}^{n_i}\frac{\exp\left[\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)y_{ij}\right]}{1+\exp\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)}
\end{aligned}
$$

Then the joint posterior distribution is

$$
\begin{aligned}
p\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},v|\mathbf{y}\right) &\propto \mathcal{L}\left(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},v\right)\pi\left(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},v\right) \\
&= \mathcal{L}\left(\mathbf{y}|\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\theta},v\right)\pi\left(\boldsymbol{\alpha}|\boldsymbol{\theta},v\right)\pi\left(\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\theta}\right)\pi\left(v\right) \\
&= \prod_{i=1}^{I}\prod_{j=1}^{n_i}\frac{\exp\left[\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)y_{ij}\right]}{1+\exp\left(\alpha_i+\boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}_{ij}\right)} \\
&\quad \prod_{i=1}^{I}\frac{2}{\gamma+\frac{1}{\gamma}}\left\{t\left(\frac{\alpha_i}{\gamma}\bigg|\mu=\boldsymbol{\theta}^{\mathsf{T}}\mathbf{z}_i,v,k\right)\mathbb{I}_{[\mu,\infty)}\left(\alpha_i\right)+t\left(\gamma\alpha_i|\mu=\boldsymbol{\theta}^{\mathsf{T}}\mathbf{z}_i,v,k\right)\mathbb{I}_{(-\infty,\mu)}\left(\alpha_i\right)\right\} \\
&\quad \left(2\pi\right)^{\frac{P}{2}}\left(\tau_{\beta}^2\right)^{\frac{P}{2}}\exp\left(-\frac{1}{2\tau_{\beta}^2}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{\beta}\right) \\
&\quad \left(2\pi\right)^{\frac{Q}{2}}\left(\tau_{\theta}^2\right)^{\frac{Q}{2}}\exp\left(-\frac{1}{2\tau_{\theta}^2}\boldsymbol{\theta}^{\mathsf{T}}\boldsymbol{\theta}\right)v^{-a-1}\exp\left(-\frac{b}{v}\right)
\end{aligned}
$$

where

$$
t\left(x|\mu,v,k\right)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k v}}\left[1+\frac{1}{k v}\left(x-\mu\right)^2\right]^{-\frac{(k+1)}{2}}
$$

is the $t$ distribution with location parameter $\mu$ and scale parameter $v$. Then the conditional posterior for $\alpha_i$ is

$$
\begin{aligned}
p\left(\alpha_i|\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2,\mathbf{y}\right) \quad \propto \quad & \prod_{j=1}^{n_i} \frac{\exp\left[\left(\alpha_i+\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)y_{ij}\right]}{1+\exp\left(\alpha_i+\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)} \\
& \frac{2}{\gamma+\frac{1}{\gamma}}\left\{t\left(\left.\frac{\alpha_i}{\gamma}\right|\mu=\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i,v,k\right)\mathbb{I}_{[\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i,\infty)}\left(\alpha_i\right)\right. \\
& \left.+t\left(\gamma\alpha_i\right|\mu=\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i,v,k)\mathbb{I}_{(-\infty,\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i)}\left(\alpha_i\right)\right\}, i=1,\ldots,I \\
\propto \quad & \prod_{j=1}^{n_i}\frac{\exp\left(\alpha_i y_{ij}\right)}{1+\exp\left(\alpha_i+\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)} \\
& \left\{\left[1+\frac{1}{kv}\left(\frac{\alpha_i}{\gamma}-\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}}\mathbb{I}_{[\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i,\infty)}\left(\alpha_i\right)\right. \\
& \left.+\left[1+\frac{1}{kv}\left(\gamma\alpha_i-\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}}\mathbb{I}_{(-\infty,\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i)}\left(\alpha_i\right)\right\} \\
\log\left[p\left(\alpha_i|\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2,\mathbf{y}\right)\right] \quad = \quad & \sum_{j=1}^{n_i}\left\{\alpha_i y_{ij}-\log\left[1+\exp\left(\alpha_i+\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)\right]\right\} \\
& -\frac{(k+1)}{2}\log\left[1+\frac{\left(\frac{\alpha_i}{\gamma}-\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2}{kv}\right]\mathbb{I}_{[\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i,\infty)}\left(\alpha_i\right) \\
& -\frac{(k+1)}{2}\log\left[1+\frac{\left(\gamma\alpha_i-\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2}{kv}\right]\mathbb{I}_{(-\infty,\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i)}\left(\alpha_i\right)
\end{aligned}
$$

As a special case, when $\alpha_i$ is normally distributed ($k=\infty,\gamma=1$), we get

$$
\log\left[p\left(\alpha_i|\boldsymbol{\beta},\boldsymbol{\theta},\sigma^2,\mathbf{y}\right)\right]=\sum_{j=1}^{n_i}\left\{\alpha_i y_{ij}-\log\left[1+\exp\left(\alpha_i+\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)\right]\right\}-\frac{\left(\alpha_i-\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2}{2\sigma^2}
$$

The conditional posterior for $\boldsymbol{\beta}$ is

$$
\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{y}) \;\;&\propto\;\; \prod_{i=1}^{I}\prod_{j=1}^{n_i} \frac{\exp\left[(\alpha_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij})\, y_{ij}\right]}{1 + \exp\left(\alpha_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)} \exp\left(-\frac{1}{2\tau_\beta^2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}\right) \\
&\propto\;\; \prod_{i=1}^{I}\prod_{j=1}^{n_i} \frac{\exp\left(\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij} y_{ij}\right)}{1 + \exp\left(\alpha_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)} \exp\left(-\frac{1}{2\tau_\beta^2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}\right) \\
\log\left[p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{y})\right] \;\;&=\;\; \sum_{i=1}^{I}\sum_{j=1}^{n_i} \left\{\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij} y_{ij} - \log\left[1 + \exp\left(\alpha_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)\right]\right\} - \frac{1}{2\tau_\beta^2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}
\end{aligned}
$$

As a special case, when $\alpha_i$ is normally distributed ($k = \infty$, $\gamma = 1$), we get

$$
\log\left[p(\boldsymbol{\beta}|\boldsymbol{\alpha}, \mathbf{y})\right] = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \left\{\boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij} y_{ij} - \log\left[1 + \exp\left(\alpha_i + \boldsymbol{\beta}^\mathsf{T}\mathbf{x}_{ij}\right)\right]\right\} - \frac{1}{2\tau_\beta^2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{\beta}
$$

The conditional posterior for $\theta_q$ is

$$
\begin{aligned}
p\left(\theta_q|\boldsymbol{\alpha}, \sigma^2\right) \;\;&\propto\;\; \prod_{i=1}^{I} \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ t\left(\left.\frac{\alpha_i}{\gamma}\right| \mu = \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i, v, k\right) \mathbb{I}_{[\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i, \infty)}(\alpha_i) \right. \\
&\qquad\left. + t\left(\gamma\alpha_i| \mu = \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i, v, k\right) \mathbb{I}_{(-\infty, \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i)}(\alpha_i) \right\} \exp\left(-\frac{1}{2\tau_\theta^2}\theta_q^2\right) \\
&\propto\;\; \prod_{i=1}^{I} \left\{ \frac{1}{\sqrt{v}}\left[1 + \frac{1}{kv}\left(\frac{\alpha_i}{\gamma} - \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{[\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i, \infty)}(\alpha_i) \right. \\
&\qquad\left. + \frac{1}{\sqrt{v}}\left[1 + \frac{1}{kv}\left(\gamma\alpha_i - \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{(-\infty, \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i)}(\alpha_i) \right\} \exp\left(-\frac{1}{2\tau_\theta^2}\theta_q^2\right) \\
\log\left[p\left(\theta_q^2|\boldsymbol{\alpha}, \sigma^2\right)\right] \;\;&=\;\; \sum_{i=1}^{I} \log\left\{ \frac{1}{\sqrt{v}}\left[1 + \frac{1}{kv}\left(\frac{\alpha_i}{\gamma} - \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{[\boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i, \infty)}(\alpha_i) \right. \\
&\qquad\left. + \frac{1}{\sqrt{v}}\left[1 + \frac{1}{kv}\left(\gamma\alpha_i - \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{(-\infty, \boldsymbol{\theta}^\mathsf{T}\mathbf{z}_i)}(\alpha_i) \right\} - \frac{1}{2\tau_\theta^2}\theta_q^2
\end{aligned}
$$

As a special case, when $\alpha_i$ is normally distributed ($k = \infty$, $\gamma = 1$), we get conjugacy with

$$p\left(\boldsymbol{\theta} | \boldsymbol{\alpha}, \sigma^2\right) \sim \mathcal{N}\left(\left(\sum_{i=1}^{I} \frac{\mathbf{z}_i \mathbf{z}_i^{\mathsf{T}}}{\sigma^2} + \frac{1}{\tau_\theta^2} \mathbf{I}\right)^{-1} \sum_{i=1}^{I} \frac{\alpha_i \mathbf{z}_i}{\sigma^2}, \left(\sum_{i=1}^{I} \frac{\mathbf{z}_i \mathbf{z}_i^{\mathsf{T}}}{\sigma^2} + \frac{1}{\tau_\theta^2} \mathbf{I}\right)^{-1}\right)$$

The conditional posterior for $v$ is

$$
\begin{aligned}
p(v | \boldsymbol{\alpha}, \boldsymbol{\theta}) \quad \propto \quad & \prod_{i=1}^{I} \frac{2}{\gamma + \frac{1}{\gamma}} \left\{ t\left(\frac{\alpha_i}{\gamma} \middle| \mu = \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i, v, k\right) \mathbb{I}_{[\boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i, \infty)}(\alpha_i) \right. \\
& \left. + t\left(\gamma \alpha_i \middle| \mu = \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i, v, k\right) \mathbb{I}_{(-\infty, \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i)}(\alpha_i) \right\} v^{-a-1} \exp\left(-\frac{b}{v}\right) \\
\propto \quad & \prod_{i=1}^{I} \left\{ \frac{1}{\sqrt{v}} \left[1 + \frac{1}{kv}\left(\frac{\alpha_i}{\gamma} - \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{[\boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i, \infty)}(\alpha_i) + \right. \\
& \left. + \frac{1}{\sqrt{v}} \left[1 + \frac{1}{kv}(\gamma \alpha_i - \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{(-\infty, \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i)}(\alpha_i) \right\} v^{-a-1} \exp\left(-\frac{b}{v}\right) \\
\log\left[p\left(v | \boldsymbol{\alpha}, \boldsymbol{\theta}\right)\right] \quad = \quad & \sum_{i=1}^{I} \log\left\{ \frac{1}{\sqrt{v}} \left[1 + \frac{1}{kv}\left(\frac{\alpha_i}{\gamma} - \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i\right)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{[\boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i, \infty)}(\alpha_i) \right. \\
& \left. + \frac{1}{\sqrt{v}} \left[1 + \frac{1}{kv}(\gamma \alpha_i - \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i)^2\right]^{-\frac{(k+1)}{2}} \mathbb{I}_{(-\infty, \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i)}(\alpha_i) \right\} - (a+1)\log(v) - \frac{b}{v}
\end{aligned}
$$

As a special case, when $\alpha_i$ is normally distributed ($k = \infty$, $\gamma = 1$), we get conjugacy with

$$p(v | \boldsymbol{\alpha}, \boldsymbol{\theta}) \sim \mathcal{IG}\left(a + \frac{I}{2}, \sum_{i=1}^{I} \frac{(\alpha_i - \boldsymbol{\theta}^{\mathsf{T}} \mathbf{z}_i)^2}{2} + b\right)$$

## 4.7    Appendix B: MCMC Sampler

For all conditional posterior distributions, a random walk Metropolis algorithm was applied with a normal proposal distribution. Univariate parameters were tuned to have acceptance rates just under 0.5, while multivariate parameters were tuned to have acceptance rates just under 0.3 (Rosenthal et al., 2011). Implementation of the MCMC sampler was done in R.

## 4.8    Appendix C: Case-Mix Adjustment

It was important to minimize the number of parameters estimated in our model, and an important part was limiting the number of case-mix adjustment variables. As our data contained deceased patients, we were able to estimate the severity of their condition based on regressing survival time against age, comorbidity, gender, and race.

# References

Agency for Healthcare Research and Quality (2012). Instructions for Analyzing Data from CAHPS Surveys. URL: `https://goo.gl/jzVqFI`. Accessed: 2016-10-15.

Ash, A. S., Fienberg, S. F., Louis, T. A., Normand, S.-L. T., Stukel, T. A., & Utts, J. (2012). Statistical Issues in Assessing Hospital Performance.

Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3), 399–424.

Bang, H. & Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4), 962–973.

Bollinger, C. R. & David, M. H. (1997). Modeling Discrete Choice with Response Error: Food Stamp Participation. *Journal of the American Statistical Association*, 92(439), 827–835.

Brillinger, D. R. (2002). John W. Tukey: His Life and Professional Contributions. *Annals of Statistics*, 30(6), 1535–1575.

Casella, G. & George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3), 167–174.

Centers for Medicare & Medicaid (2016). 2016 Medicare Trustees Report. URL: `https://goo.gl/aj8WBf`. Accessed: 2016-10-15.

Chen, M.-H., Huang, L., Ibrahim, J. G., & Kim, S. (2008). Bayesian Variable Selection and Computation for Generalized Linear Models with Conjugate Priors. *Bayesian Analysis*, 3(3), 585.

Crofton, C., Lubalin, J. S., & Darby, C. (1999). Foreword. *Medical Care*, 37(3), MS1–MS9.

DeLong, E. R., Peterson, E. D., DeLong, D. M., Muhlbaier, L. H., Hackett, S., & Mark, D. B. (1997). Comparing Risk-Adjustment Methods for Provider Profiling. *Statistics in Medicine*, 16(23), 2645–2664.

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3), e1003348.

Duncan, L., Ratanatharathorn, A., Aiello, A., Almli, L., Amstadter, A., Ashley-Koch, A., Baker, D., Beckham, J., Bierut, L., Bisson, J., et al. (2017). Largest GWAS of PTSD (N= 20070) Yields Genetic Overlap with Schizophrenia and Sex Differences in Heritability. *Molecular Psychiatry*.

Edwards, J. K., Cole, S. R., Troester, M. A., & Richardson, D. B. (2013). Accounting for Misclassified Outcomes in Binary Regression Models using Multiple Imputation with Internal Validation Data. *American Journal of Epidemiology*, 177(9), 904–912.

Fay, R. E. & Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74(366a), 269–277.

Fernández, C. & Steel, M. F. (1998). On Bayesian Modeling of Fat Tails and Skewness. *Journal of the American Statistical Association*, 93(441), 359–371.

Gatsonis, C. (1998). Profiling Providers of Medical Care. *Encyclopedia of Biostatistics*.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society: Series B*, 56(3), 501–514.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica*, (pp. 733–760).

Ghosh, M. & Rao, J. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9(1), 55–76.

Gilbert, R., Martin, R. M., Donovan, J., Lane, J. A., Hamdy, F., Neal, D. E., & Metcalfe, C. (2014). Misclassification of Outcome in Case-Control Studies: Methods for Sensitivity Analysis. *Statistical Methods in Medical Research*, (pp. 0962280214523192).

Goldstein, E., Cleary, P. D., Langwell, K. M., Zaslavsky, A. M., & Heller, A. (2001). Medicare Managed Care CAHPS: A Tool for Performance Improvement. *Health Care Financing Review*, 22(3), 101.

Goldstein, H. & Spiegelhalter, D. J. (1996). League Tables and their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society: Series A*, (pp. 385–443).

Hughes, A. (2005). Poverty and Palliative Care in the US: Issues Facing the Urban Poor. *International Journal of Palliative Nursing*, 11(1).

Imai, K. & Van Dyk, D. A. (2004). Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association*, 99(467), 854–866.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., & Kimmel, S. E. (2004). Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications. *The American Statistician*, 58(4), 272–279.

Jurek, A. M., Maldonado, G., & Greenland, S. (2013). Adjusting for Outcome Misclassification: The Importance of Accounting for Case-Control Sampling and Other Forms of Outcome-Related Selection. *Annals of Epidemiology*, 23(3), 129–135.

Kass, R. E. & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Keenan, P. S., Cleary, P. D., O'Malley, A. J., Landon, B. E., Zaborski, L., & Zaslavsky, A. M. (2010). Geographic Area Variations in the Medicare Health Plan Era. *Medical Care*, 48(3), 260–266.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2016). Non-Parametric Methods for Doubly Robust Estimation of Continuous Treatment Effects. *Journal of the Royal Statistical Society: Series B*.

Kessler, R. C., Colpe, L. J., Fullerton, C. S., Gebler, N., Naifeh, J. A., Nock, M. K., Sampson, N. A., Schoenbaum, M., Zaslavsky, A. M., Stein, M. B., et al. (2013). Design of the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *International Journal of Methods in Psychiatric Research*, 22(4), 267–275.

Kikule, E. (2003). A Good Death in Uganda: Survey of Needs for Palliative Care for Terminally Ill People in Urban Areas. *BMJ*, 327(7408), 192–194.

Klein, D. J., Elliott, M. N., Haviland, A. M., Saliba, D., Burkhart, Q., Edwards, C., & Zaslavsky, A. M. (2011). Understanding Nonresponse to the 2007 Medicare CAHPS Survey. *The Gerontologist*, 51(6), 843–855.

Kuehn, B. M. (2009). Soldier Suicide Rates Continue to Rise. *Journal of the American Medical Association*, 301(11), 1111–1113.

Kumar, S. K. (2007). Kerala, India: A Regional Community-Based Palliative Care Model. *Journal of Pain and Symptom Management*, 33(5), 623–627.

Kurreeman, F., Liao, K., Chibnik, L., Hickey, B., Stahl, E., Gainer, V., Li, G., Bry, L., Mahan, S., Ardlie, K., et al. (2011). Genetic Basis of Autoantibody Positive and Negative Rheumatoid Arthritis Risk in a Multi-Ethnic Cohort Derived from Electronic Health Records. *The American Journal of Human Genetics*, 88(1), 57–69.

Lee, K. J. & Thompson, S. G. (2008). Flexible Parametric Models for Random-Effects Distributions. *Statistics in Medicine*, 27(3), 418–434.

Leyland, A. H. & Boddy, F. A. (1998). League Tables and Acute Myocardial Infarction. *The Lancet*, 351(9102), 555–558.

Liao, K. P., Cai, T., Savova, G. K., Murphy, S. N., Karlson, E. W., Ananthakrishnan, A. N., Gainer, V. S., Shaw, S. Y., Xia, Z., Szolovits, P., et al. (2015). Development of Phenotype Algorithms using Electronic Medical Records and Incorporating Natural Language Processing. *BMJ*, 350, h1885.

Lin, D. & Zeng, D. (2009). Proper Analysis of Secondary Phenotype Data in Case-Control Association Studies. *Genetic Epidemiology*, 33(3), 256–265.

Lyles, R. H., Tang, L., Superak, H. M., King, C. C., Celentano, D. D., Lo, Y., & Sobel, J. D. (2011). Validation Data-Based Adjustments for Outcome Misclassification in Logistic Regression: An Illustration. *Epidemiology*, 22(4), 589.

Magder, L. S. & Hughes, J. P. (1997). Logistic Regression when the Outcome is Measured with Uncertainty. *American Journal of Epidemiology*, 146(2), 195–203.

Monsees, G. M., Tamimi, R. M., & Kraft, P. (2009). Genome-Wide Association Scans for Secondary Traits using Case-Control Samples. *Genetic Epidemiology*, 33(8), 717–728.

Neuhaus, J. M. (1999). Bias and Efficiency Loss due to Misclassified Responses in Binary Regression. *Biometrika*, 86(4), 843–855.

Nock, M. K., Stein, M. B., Heeringa, S. G., Ursano, R. J., Colpe, L. J., Fullerton, C. S., Hwang, I., Naifeh, J. A., Sampson, N. A., Schoenbaum, M., et al. (2014). Prevalence and Correlates of Suicidal Behavior among Soldiers: Results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry*, 71(5), 514–522.

Normand, S.-L. T. (2005). Quality of Care. *Encyclopedia of Biostatistics*.

Normand, S.-L. T., Glickman, M. E., & Gatsonis, C. A. (1997). Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association*, 92(439), 803–814.

Normand, S.-L. T. & Shahian, D. M. (2007). Statistical and Clinical Aspects of Hospital Outcomes Profiling. *Statistical Science*, (pp. 206–226).

O'Malley, A. J. & Zaslavsky, A. M. (2008). Domain-Level Covariance Analysis for Multilevel Survey Data with Structured Nonresponse. *Journal of the American Statistical Association*, 103(484), 1405–1418.

Pepe, M. S. (1992). Inference using Surrogate Outcome Data and a Validation Sample. *Biometrika*, 79(2), 355–365.

Prentice, R. L. & Pyke, R. (1979). Logistic Disease Incidence Models and Case-Control Studies. *Biometrika*, 66(3), 403–411.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies. *Nature Genetics*, 38(8), 904–909.

Racz, M. J. & Sedransk, J. (2010). Bayesian and Frequentist Methods for Provider Profiling using Risk-Adjusted Assessments of Medical Outcomes. *Journal of the American Statistical Association*, 105(489), 48–58.

Reilly, M. & Pepe, M. S. (1995). A Mean Score Method for Missing and Auxiliary Covariate Data in Regression Models. *Biometrika*, 82(2), 299–314.

Reiter, J. P. (2000). Borrowing Strength when Explicit Data Pooling is prohibited. *Journal of Official Statistics*, 16(4), 295.

Richardson, D. B., Rzehak, P., Klenk, J., & Weiland, S. K. (2007). Analyses of Case-Control Data for Additional Outcomes. *Epidemiology*, 18(4), 441–445.

114

Robins, J. M., Hernan, M. A., & Brumback, B. (2000). Marginal Structural Models and Causal Inference in Epidemiology.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429), 122–129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90(429), 106–121.

Rome, R. B., Luminais, H. H., Bourgeois, D. A., & Blais, C. M. (2011). The Role of Palliative Care at the End of Life. *The Ochsner Journal*, 11(4), 348–352.

Rosenbaum, P. R. & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, (pp. 41–55).

Rosenthal, J. S. et al. (2011). Optimal Proposal Distributions and Adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, 4.

Salem-Schatz, S., Moore, G., Rucker, M., & Pearson, S. D. (1994). The Case for Case-Mix Adjustment in Practice Profiling: When Good Apples Look Bad. *Journal of the American Medical Association*, 272(11), 871–874.

Schifano, E. D., Li, L., Christiani, D. C., & Lin, X. (2013). Genome-Wide Association Analysis for Multiple Continuous Secondary Phenotypes. *The American Journal of Human Genetics*, 92(5), 744–759.

Schoenbaum, M., Kessler, R. C., Gilman, S. E., Colpe, L. J., Heeringa, S. G., Stein, M. B., Ursano, R. J., & Cox, K. L. (2014). Predictors of Suicide and Accident Death in the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS): Results from the Army Study to Assess Risk and Resilience in Servicemembers (Army STARRS). *JAMA Psychiatry*, 71(5), 493–503.

Shahian, D. M. & Normand, S.-L. T. (2008). Comparison of "Risk-Adjusted" Hospital Outcomes. *Circulation*, 117(15), 1955–1963.

Shahian, D. M. & Normand, S.-L. T. (2015). What is a Performance Outlier?

Shen, W. & Louis, T. A. (1998). Triple-Goal Estimates in Two-Stage Hierarchical Models. *Journal of the Royal Statistical Society: Series B*, 60(2), 455–471.

Siegmund, K. D., Whittemore, A. S., & Thomas, D. C. (1999). Multistage Sampling for Disease Family Registries. *Journal of the National Cancer Institute Monographs*, 26, 43–8.

Sinnott, J. A., Dai, W., Liao, K. P., Shaw, S. Y., Ananthakrishnan, A. N., Gainer, V. S., Karlson, E. W., Churchill, S., Szolovits, P., Murphy, S., et al. (2014). Improving the Power of Genetic Association Tests with Imperfect Phenotype Derived from Electronic Medical Records. *Human Genetics*, 133(11), 1369–1382.

Steece, B. M. (1989). Leverage in Bayesian Regression. *Biometrical Journal*, 31(7), 811–819.

Stein, M. B., Chen, C.-Y., Ursano, R. J., Cai, T., Gelernter, J., Heeringa, S. G., Jain, S., Jensen, K. P., Maihofer, A. X., Mitchell, C., et al. (2016). Genome-Wide Association Studies of Post-traumatic Stress Disorder in 2 Cohorts of US Army Soldiers. *JAMA Psychiatry*.

Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: Simulation of Multiple Disease SNPs. *Bioinformatics*, 27(16), 2304–2305.

Temel, J. S., Greer, J. A., Muzikansky, A., Gallagher, E. R., Admane, S., Jackson, V. A., Dahlin, C. M., Blinderman, C. D., Jacobsen, J., Pirl, W. F., et al. (2010). Early Palliative Care for Patients with Metastatic Non-Small-Cell Lung Cancer. *New England Journal of Medicine*, 363(8), 733–742.

Thomas, N., Longford, N. T., & Rolph, J. E. (1994). Empirical Bayes Methods for Estimating Hospital-Specific Mortality Rates. *Statistics in Medicine*, 13(9), 889–903.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Varewyck, M., Goetghebeur, E., Eriksson, M., & Vansteelandt, S. (2014). On Shrinkage and Model Extrapolation in the Evaluation of Clinical Center Performance. *Biostatistics*, 15(4), 651–664.

Venturini, S., Franklin, J. M., Morlock, L., Dominici, F., et al. (2017). Random Effects Models for Identifying the most Harmful Medication Errors in a Large, Voluntary Reporting Database. *The Annals of Applied Statistics*, 11(2), 504–526.

Wacholder, S. (1996). The Case-Control Study as Data Missing by Design: Estimating Risk Differences. *Epidemiology*, 7(2), 144–150.

Wooldridge, J. M. (2007). Inverse Probability Weighted Estimation for General Missing Data Problems. *Journal of Econometrics*, 141(2), 1281–1301.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics*, 89(1), 82–93.

Zaslavsky, A. M. (2007). Using Hierarchical Models to Attribute Sources of Variation in Consumer Assessments of Health Care. *Statistics in Medicine*, 26(8), 1885–1900.

Zaslavsky, A. M. & Bradlow, E. T. (2010). Posterior Predictive Outlier Detection Using Sample Reweighting. *Journal of Computational and Graphical Statistics*, 19(4), 790–807.

Zaslavsky, A. M. & Cleary, P. D. (2002). Dimensions of Plan Performance for Sick and Healthy Members on the Consumer Assessments of Health Plans Study 2.0 Survey. *Medical Care*, 40(10), 951–964.