2018

Addressing Geographic Confounding through Spatial Propensity Score Analysis for Hierarchical Data

Melanie Davis
Medical University of South Carolina

Addressing Geographic Confounding through Spatial Propensity Score Analysis for Hierarchical Data

Melanie Davis

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

College of Graduate Studies
Department of Public Health Sciences
March 2018

Approved by:

Brian Neelon, Ph.D. (Chair)

Lane Burgette, Ph.D.

Leonard Egede, M.D., M.S.

Kelly Hunt, Ph.D.

Andrew Lawson, Ph.D.

Paul Nietert, Ph.D.

Acknowledgments

I greatly appreciate the guidance of my committee and the opportunity to conduct this research under grant CIN 13-418 (PI: Leonard Egede) funded by the VHA Health Services Research and Development (HSR&D) program.

I am especially honored to work with Dr. Brian Neelon as my advisor and am immensely grateful for the time and commitment he has given to this research.

I would like to acknowledge the support I have received from my family and friends who have absolutely no idea what I actually do but, because of me, are now under the impression that it takes a quarter of your life just to get a degree that says you can do it. It helps to have people who are important enough in your life to put this all in perspective.

I would like to express the utmost gratitude for the patience and support of my boyfriend, Christopher Harrison. Throughout this endeavor, he has spent thousands of hours on the phone 750 miles away listening to the highest highs and lowest lows of this process. I am absolutely certain he has been conditioned to perspire at the mere mention of the word "simulation".

And lastly, I would like to acknowledge the life-long love and support of my

parents, Jeff and Donna Davis. It is because of them that up until entering this doctoral program, I never doubted myself or what I was capable of. And when I began to, they never wavered in their dedication to my education and reminded me of the light at the end of the tunnel even when I could not see it. It is because of their sacrifices and support that I have been able to pursue and endure this endeavor and I am so very thankful.

Abstract

Motivated by recent work exploring cluster-level confounding in multilevel observational data, we develop methods specifically addressing geographic confounding, which occurs when measured or potentially unmeasured confounding factors vary by geographic location. Accounting for this source of confounding achieves spatially-balanced global estimates of the treatment effect of interest, allowing researchers to compare individuals as if they were residentially similar and leading to policy decisions that benefit patients and areas most in need. This dissertation consists of three aims: 1. To develop a hierarchical spatial doubly robust estimator in propensity score analysis framework; 2. To develop spatial propensity score matching methods for hierarchical data; 3. To apply spatial propensity score matching to more complex analyses of spatially varying, zero-inflated outcomes. Each of these aims strives to explore the issue of geographic confounding and contribute to its resolution. Aim 1 seeks to build upon multilevel propensity score methods through augmentation of modeling with spatial random effects to create a spatially balanced estimator that is demonstrated in simulation to exhibit favorable performance under various sample sizes and levels of spatial heterogeneity. Aim 2 seeks to develop methods in a

propensity score matching framework, allowing for a more complete understanding of geographic confounding remediation techniques and extensions to additional applications. Finally, as modeling non-binary, spatially varying outcomes can prove challenging, Aim 3 seeks to incorporate spatial matching to alleviate geographic imbalance to allow for a minimally confounded analysis. We apply the spatial matching approach to the analysis of zero-inflated count outcomes.

Contents

1 Literature Review **8**

1.1 Causal Inference and the Potential Outcomes Framework 8

1.2 Propensity Score Analysis 12

1.2.1 Overview 12

1.2.2 Methods 13

1.2.3 PSA for Multi-Level Data 19

1.3 Using Spatial Propensity Scores to Address Geographic Confounding 20

1.3.1 Geographic Confounding 20

1.3.2 Existing Spatial PSA Methods 21

2 Proposed Methodology **23**

2.1 Overview 23

2.2 Motivating Example 25

2.2.1 Background 25

2.2.2 Data Description 26

2.3 Specific Aims 28

2.3.1 Aim 1 28
2.3.2 Aim 2 29
2.3.3 Aim 3 30

3 Aim 1 **31**

3.1 Introduction 33
3.2 Spatial Propensity Score Analysis 35
3.2.1 Overview of Propensity Score Weighting Methods 35
3.2.2 A Doubly Robust Estimator for Hierarchical Spatial Data . . 40
3.2.3 Model Fitting and Inference 42
3.3 Simulation Study 45
3.3.1 Data Description 45
3.3.2 Results 46
3.4 Analysis of Racial Disparities in Glycemic Control 49
3.5 Discussion 60
3.6 Acknowledgments 64

4 Aim 2 **65**

4.1 Introduction 67
4.2 Spatial Propensity Score Analysis 71
4.2.1 Overview of Propensity Score Matching Methods 71
4.2.2 Multi-level Spatial Matching 74
4.2.3 Model Fitting and Inference 76
4.3 Simulation Study 78

4.3.1 Data Description 78
4.3.2 Results 79
4.4 Analysis of Racial Disparities in Diabetes Care and Management . . . 82
4.5 Discussion 88

5 Aim 3 **92**
5.1 Introduction 92
5.2 Methods 96
5.2.1 Spatial Propensity Score Analysis 96
5.2.2 Two Part Spatial Hurdle Models 98
5.2.3 Treatment Effect Estimation 101
5.3 Model Fitting 102
5.4 Simulation Study 103
5.4.1 Data Description 103
5.4.2 Results 105
5.5 Analysis of Racial Disparities in Hospitalization and Inpatient Days . 106
5.5.1 Data Description 106
5.5.2 Analysis and Results 107
5.6 Discussion 111

6 Conclusions of Research **114**
6.1 Summary 114
6.2 Implications 115
6.3 Limitations and Extensions 115

Chapter 1

Literature Review

1.1 Causal Inference and the Potential Outcomes Framework

Evidence-based medicine seeks to incorporate cutting edge research in clinical decisions between a patient and a care-provider [1]. In order to contribute strong and convincing "evidence" of clinical benefit, researchers must strive to be able to attribute observed endpoint differences in outcome to the given treatment. If patient- or system-level differences related to the outcome exist between treatment groups, confounding may occur and it thus becomes challenging to determine with confidence that it was specifically the treatment that elicited, or caused, the effect. It is this desire to assert causal inference that has elevated randomized controlled trials (RCTs) to the top of the study evaluation pyramid, considering evidence generated from studies with this design more sound and convincing than evidence generated

from other study designs, thus designating RCTs the "gold standard" of health outcomes research [2]. Randomization seeks to eliminate selection bias. Along with the concurrent nature of the control group, randomization ensures the balance of covariates, some of which are unmeasured [3]. It has even been stated that randomization is the only method to control for imbalance with respect to measured and unmeasured factors [4]. Although issues such as bias should be addressed early in the study design phase and continue to be revisited through analysis, they are often overlooked in RCTs due to the confidence the community has in the study design. For example, in the acclaimed NINDS Tissue Plasminogen Activator for Acute Ischemic Stroke trial, differences in baseline stroke severity between treatment groups prevailed despite randomization [5]. Beyond the context of RCTs, there exist treatments that are inherently implausible, impossible or unethical to randomize, but that may be examined in observational studies. It is in observational studies that we must revisit the issue of confounding.

The potential outcomes framework put forth by Rubin [6] provides the theoretical foundation of contemporary causal inference. In the setting of clinical trials, let Z denote a binary treatment assignment. Each individual is assumed to have two potential outcomes: Y_1 and Y_0. If $Z = 1$, then Y_1 is observed and Y_0 is unobserved, while if $Z = 0$, then Y_0 is observed and Y_1 is unobserved. [6] The observed response Y can be expressed by $Y = ZY_1 + (1 - Z)Y_0$, yielding $Y = Y_1$ when $Z = 1$ and $Y = Y_0$ when $Z = 0$. The average treatment effect (ATE) is defined as $\Delta = E(Y_1) - E(Y_0)$, and, since in a randomized controlled trial the treatment groups are in theory balanced with respect to all but treatment assignment, (Y_1, Y_0) are

stochastically independent of treatment assignment. In essence, those who did not receive treatment can serve as the counterfactual observation for those who did, and vice versa.

The average treatment effect on the treated (ATT) is an alternative estimand of interest and can be defined as $\Delta_{ATT} = E(Y_1 - Y_0|Z = 1)$. Note the condition on treatment indicator $Z = 1$ in the case of the ATT that was not present in that of the ATE. Informally, the ATT is the difference in outcome between those who were treated and those who were not for those who actually participated in treatment [7].

In observational studies, it may be more appropriate to discuss "exposure" rather than "treatment" groups; in any case, these groups are not guaranteed to be balanced on a set of covariates that are also associated with the outcome Y and hence the issue of confounding arises. When we address confounding and can reasonably infer no unmeasured confounding is present, we may assume that (Y_1, Y_0) are conditionally independent of treatment assignment Z given a vector of covariates $\boldsymbol{X}$, expressed as $(Y_1, Y_0) \perp\!\!\!\perp Z|\boldsymbol{X} = \boldsymbol{x}$ [8].

In the special case where the group designation is an immutable characteristic such as race, it is important to note that the causal inference framework may not be appropriate since it is impossible to conceive a potential outcome corresponding to an alternative race designation. In this setting, Li et al. [9] propose using the average controlled difference (ACD) as a *descriptive* estimand analogous to the ATE. The ACD is defined as $\Delta = E_{\boldsymbol{X}}[E(Y|\boldsymbol{X} = \boldsymbol{x}, Z = 1) - E(Y|\boldsymbol{X} = \boldsymbol{x}, Z = 0)]$, where the outer expectation is taken with respect to the distribution of covariates $\boldsymbol{X}$ in the entire population and $\boldsymbol{x}$ is an observed realization of the random variable $\boldsymbol{X}$.

The ACD can be viewed as a weighted sum of the expected group differences formed within each stratum of $\boldsymbol{X}$. Under the assumption of no unmeasured confounding, Δ represents a controlled population-average difference between the two groups.

The ultimate decision among estimating the ATE, ACD or ATT has much to do with clinical and epidemiological questions of interest; however, statistical consideration must also be attended to. Interpretation of results should coincide with the estimand the data and analysis support.

Estimation of causal inference has assumptions that include no unmeasured confounding, a positive probability of each individual to receive treatment, and the stable unit treatment variable assumption (SUTVA) [10]. SUTVA assumes that treatment effect is uniform and that the treatment of one subject does not affect the outcome of another. Schwartz et al. [11] reiterate Little and Rubin's [12] classic example of psychotherapy: SUTVA could be violated if the effectiveness of the therapy varied among levels of expertise of the therapist or willingness of the patient to participate. Furthermore, SUTVA could be violated if a treated patient shared his insight with other patients, thus affecting their outcomes. SUTVA may also be violated in multilevel settings where it is likely that clusters of individuals interact in a way that may confer benefit or harm to a "neighbor" of the treated individual despite that neighbor not having been exposed to treatment himself.

1.2 Propensity Score Analysis

1.2.1 Overview

Propensity score analysis [8] (PSA) lends methodology to address confounding and under appropriately addressed assumptions, allows the researcher to make causal inferences in observational studies. The propensity score $e(\boldsymbol{x}) = \Pr(Z = 1|\boldsymbol{X} = \boldsymbol{x})$, is the conditional probability of exposure given a set of covariates $\boldsymbol{X}$ under the assumption of strong ignorability, i.e. the positivity condition $0 < e(\boldsymbol{x}) < 1$ and no unmeasured confounding. Variable selection for the propensity score model (1.1) is debated in the literature but tends to incorporate baseline covariates, potential confounders and true confounders [2] but should not include any factor that could have been affected by treatment assignment (post-treatment) [13]. It has been shown that there is minimal detriment when including variables not strongly associated with exposure and a small impact in increased variance estimates when including variables not strongly associated with the outcome. It is suggested that analysts suppress traditional concerns about collinearity and err on the side of inclusion to satisfy the condition of no unmeasured confounding [13]. Propensity scores are commonly generated from logistic regression [7]:

$$\text{logit}(e_i) = \text{logit}[e(\boldsymbol{x}_i)] = \text{logit}[\Pr(Z_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i)] = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad i = 1, \ldots, n. \tag{1.1}$$

Furthermore, it has been shown that the propensity score acts as a balancing score to achieve conditional independence of treatment assignment Z as noted above. [8]

Once generated from a model of the form of equation (1.1), propensity scores can be used in various settings to achieve this balance.

1.2.2 Methods

1.2.2.1 Inverse Probability of Treatment Weighting

Propensity score weighting, commonly referred to as inverse probability of treatment weighting (IPTW), assigns propensity score based weights to individuals in a sample, thus creating a new sample in which covariates are balanced between exposure groups. Weights are assigned by

$$w_i = \frac{Z_i}{\hat{e}_i} + \frac{(1 - Z_i)}{1 - \hat{e}_i} \tag{1.2}$$

and therefore an IPTW estimator of the ATE [14] is given by

$$\hat{\Delta}^{IPTW} = n^{-1} \sum_{i=1}^{n} \left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} \right] \tag{1.3}$$

where $\hat{e}_i$ denotes the estimated propensity score for subject i.

Issues in accuracy and stability may arise when estimated propensity scores approach 0 or 1. Stabilizing weights [15] have been suggested as a remedy to this issue but limit the analyst to estimating the ATT, which is not numerically equivalent to the ATE in non-randomized studies. Additionally, issues of misspecification of the propensity score model may deter confidence in the estimation of the ATE. If the estimated propensity score is not equal to the true propensity score, the aforementioned

formula (1.3) will not necessarily estimate Δ [16]. As a remedy, however, an augmentation of this formula leads to a semi-parametric doubly robust (DR) estimator [17] of the form

$$\begin{aligned} \hat{\Delta}^{DR} &= \frac{1}{N}\sum_{i=1}^{n}\hat{\Delta}_i \\ \hat{\Delta}_i &= \left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(Z_i-\hat{e}_i)\hat{Y}_{i1}}{\hat{e}_i}\right] - \left[\frac{(1-Z_i)Y_i}{1-\hat{e}_i} + \frac{(Z_i-\hat{e}_i)\hat{Y}_{i0}}{1-\hat{e}_i}\right]. \end{aligned} \tag{1.4}$$

This DR estimator provides a safeguard and has been shown to be a consistent estimator of the ATE if either the propensity score model or a model for the outcome is correctly specified. $\hat{Y}_{i1}$ and $\hat{Y}_{i0}$ are predictions of the potential outcomes under treatment and control, respectively, generated from a regression model of Y on Z and $\boldsymbol{X}$. It should be noted that in the case of $Z = 1$, $\hat{Y}_{i0}$ is counterfactual and is not observed while in the case of $Z = 0$, $\hat{Y}_{i1}$ is counterfactual and is not observed. The estimator $\hat{\Delta}^{DR}$ can be conceptualized as a difference in means $\hat{\mu}_1 - \hat{\mu}_0$ or a difference in proportions $\hat{p}_1 - \hat{p}_0$, i.e. a risk difference.

When the risk difference is of interest, the estimator $\hat{\Delta}^{DR} = \hat{p}_1 - \hat{p}_0$ where $\hat{p}_1$ estimates $E(Y_1)$ and $\hat{p}_0$ estimates $E(Y_0)$. In the case of $\hat{p}_1$, and similarly following for the case of $\hat{p}_0$, $\hat{p}_1$ estimates the following expression where the "postulated" propensity score model is expressed $e(\boldsymbol{X}, \beta)$ and the "postulated" outcome regression model is denoted $m_1(\boldsymbol{X}, \alpha_1)$ [16].

$$E\left[\frac{ZY}{e(\boldsymbol{X},\beta)} - \frac{Z-e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)} m_1(\boldsymbol{X},\alpha_1)\right]$$

$$= E\left[\frac{ZY_1}{e(\boldsymbol{X},\beta)} - \frac{Z-e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)} m_1(\boldsymbol{X},\alpha_1)\right]$$

$$= E\left[Y_1 + \frac{Z-e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - m_1(\boldsymbol{X},\alpha_1))\right]$$

$$= E(Y_1) + E\left[\frac{Z-e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - m_1(\boldsymbol{X},\alpha_1))\right]$$

and therefore $\hat{p}_1$ unbiasedly estimates $E(Y_1)$ when $E\left[\frac{Z-e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - m_1(\boldsymbol{X},\alpha_1))\right] = 0$.

When the propensity score model is correct and strong ignorability holds, i.e. $e(\boldsymbol{X},\beta) = e(\boldsymbol{X}) = E(Z|\boldsymbol{X}) = E(Z|Y_1,\boldsymbol{X})$, but the outcome model is misspecified, it can be shown that $E\left[\frac{Z-e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - Y_{ij1})\right] = 0$ through the following set of equations:

$$E\left[\frac{Z-e(\boldsymbol{X})}{e(\boldsymbol{X})}(Y_1 - m_1(\boldsymbol{X},\alpha_1))\right]$$

$$= E\left(E\left[\frac{Z-e(\boldsymbol{X})}{e(\boldsymbol{X})}(Y_1 - m_1(\boldsymbol{X},\alpha_1))|Y_1,\boldsymbol{X}\right]\right)$$

$$= E\left((Y_1 - m_1(\boldsymbol{X},\alpha_1))E\left[\frac{Z-e(\boldsymbol{X})}{e(\boldsymbol{X})}|Y_1,\boldsymbol{X}\right]\right)$$

$$= E\left((Y_1 - m_1(\boldsymbol{X},\alpha_1))\frac{E(Z|Y_1,\boldsymbol{X})-e(\boldsymbol{X})}{e(\boldsymbol{X})}\right)$$

$$= E\left((Y_1 - m_1(\boldsymbol{X},\alpha_1))\frac{E(Z|\boldsymbol{X})-e(\boldsymbol{X})}{e(\boldsymbol{X})}\right)$$

$$= E\left((Y_1 - m_1(\boldsymbol{X}, \alpha_1))\frac{e(\boldsymbol{X}) - e(\boldsymbol{X})}{e(\boldsymbol{X})}\right) = 0$$

When the outcome model is correct and unconfoundedness holds, i.e. $m_1(\boldsymbol{X}, \alpha_1) = E(Y|Z = 1, X) = E(Y_1|\boldsymbol{X})$, but the propensity score model is misspecified, it can once again be demonstrated that $E\left[\frac{Z - e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - Y_{ij1})\right] = 0$, as follows:

$$E\left[\frac{Z - e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - E(Y|Z = 1, \boldsymbol{X}))\right]$$

$$= E\left(\left[\frac{Z - e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(Y_1 - E(Y|Z = 1, \boldsymbol{X}))|Z, \boldsymbol{X}\right]\right)$$

$$= E\left(\frac{Z - e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}E\left[(Y_1 - E(Y|Z = 1, \boldsymbol{X}))|Z, \boldsymbol{X}\right]\right)$$

$$= E\left(\frac{Z - e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(E(Y_1|Z, \boldsymbol{X}) - E(Y|Z = 1, \boldsymbol{X}))\right)$$

$$= E\left(\frac{Z - e(\boldsymbol{X},\beta)}{e(\boldsymbol{X},\beta)}(E(Y_1|\boldsymbol{X}) - E(Y_1|\boldsymbol{X}))\right) = 0$$

Using similar arguments, it can be shown that $\hat{p}_0$ unbiasedly estimates $E(Y_0)$ under strong ignorability [16]. It should be noted that when the propensity score and outcome regression models are both misspecified, the DR estimator offers no protection and the estimates derived from such equations are not likely to be unbiased.

A large-sample approximate standard error of $\hat{\Delta}^{DR}$ is

$$s^2 = \frac{1}{n^2}\sum_{i=1}^{n}(\hat{\Delta}_i - \hat{\Delta}^{DR})^2. \tag{1.5}$$

Bootstrapping can also be employed to derive the standard error and surrounding confidence interval.

1.2.2.2 Propensity Score Matching

Propensity score matching is a technique that forms matched pairs between exposed and unexposed subjects based on the similarity of their estimated propensity scores [2, 8, 18]. As in all PSA, matching techniques require the analyst to first decide on the form of the propensity score model and the variables to be included in the model. After propensity scores have been generated, the analyst must then make decisions on the matching strategy: the uniqueness of pairs, number of controls to be matched to each exposed individual, the matching variable itself (propensity score, logit of propensity score, etc.), and the rules for designating acceptable matches. In terms of acceptable match designation, Austin [2] recommends a caliper width equal to 0.2 times the standard deviation of the logit of the propensity score as a valuable compromise between preserving match quality and minimizing mean square error (MSE) of the treatment effect.

Nearest neighbor k:1 matching [19] is implemented by matching treated subjects with k controls, although k = 1 may be most popular. This method results in a balanced sample if the propensity score is correctly specified; however, many controls may be discarded, resulting in a drastically reduced sample size and restricting the analyst to estimating the ATT [13]. Matching without replacement marries the treated and control subjects and precludes the control from further matches, while matching with replacement allows the control subject to be eligible for participation

in a new pair. Another distinction is "greedy" versus "optimal" matching. Greedy nearest neighbor matching may appear to be short-sighted as it is only concerned with the treated subject's best match without considering future matches, while optimal matching minimizes the difference in propensity scores in the overall sample, [20] although there is no strong indication that optimal matching is universally superior at producing well-matched groups [13].

Optimal full matching [21] is a special case of subclassification and may alleviate concerns regarding reduction in sample size. Full matching divides the sample into matched groups that contain at least one treated subject and any positive number of controls [22]. This method provides the analyst with a strategy to estimate the ATE in addition to the ATT.

The ATE, as described in detail in the earlier section concerning confounding, deserves further attention as its specific interpretation should not be conflated with that of the ATT. Based on the amount of overlap in the propensity scores of the treatment and control groups (i.e. common support), it may not be feasible to calculate the ATE even in a full matching setting. When the ratio of control:treated subjects is high, the technique of k:1 matching allows the analyst to construct the ATT; however, if the ratio is low, full matching may be necessary [13].

1.2.2.3 Assessing Balance

Propensity score methods should produce a well-balanced weighted or matched sample. Once propensity scores have been generated via equation (1.1) and decisions have been made with respect to its utility, assessment in balance can be achieved by

the calculation of the standardized difference for means and proportions, respectively [2].

$$d_{mean} = \frac{(\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{\frac{(s^2_{treatment} + s^2_{control})}{2}}} \tag{1.6}$$

$$d_{proportion} = \frac{(\hat{p}_{treatment} - \hat{p}_{control})}{\sqrt{\frac{(\hat{p}_{treatment}(1-\hat{p}_{treatment}) + \hat{p}_{control}(1-\hat{p}_{control})}{2}}} \tag{1.7}$$

Although commonly and inaccurately reported with test statistics and associated p-values, balance assessment is limited to sample-level description and should not be subject to fluctuations due to sample size reduction, hence the appropriateness of the standardized difference [23].

Balance of individual-level covariates may not be sufficient to remedy confounding if patients are clustered and measured or unmeasured confounders are associated with exposure at the cluster-level [9, 24].

1.2.3 PSA for Multi-Level Data

The DR estimator has recently been extended to multilevel data [9] where data is comprised of (Y_{ij}, Z_{ij}, X_{ij}) for the j^{th} subject in the i^{th} cluster:

$$\begin{aligned} \hat{\Delta}^{DR} &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\hat{\Delta}_{ij} \\ \hat{\Delta}_{ij} &= \left[\frac{Z_{ij}Y_{ij}}{\hat{e}_{ij}} - \frac{(Z_{ij} - \hat{e}_{ij})\hat{Y}_{ij1}}{\hat{e}_{ij}}\right] - \left[\frac{(1 - Z_{ij})Y_{ij}}{1 - \hat{e}_{ij}} + \frac{(Z_{ij} - \hat{e}_{ij})\hat{Y}_{ij0}}{1 - \hat{e}_{ij}}\right] \end{aligned} \tag{1.8}$$

In this formula (1.8), $\hat{e}_{ij}$ denotes the propensity score for the $(ij)^{th}$ individual and $N = \sum n_i$ where n_i indicates the sample size of cluster i. As before, generalized linear mixed models are used to estimate e_{ij}, Y_{ij0} and Y_{ij1}; however, in this case random effects are incorporated to represent between-cluster heterogeneity and account for unobserved cluster-level confounders. The results of extensive simulations by Li et al. signal the necessity to incorporate cluster level information in PSA.

Furthermore, Arpino and Mealli have recently extended propensity score matching to the multilevel setting [25]. In their work, they demonstrate that ignoring cluster assignment has deleterious effects when estimating the ATT.

1.3 Using Spatial Propensity Scores to Address Geographic Confounding

1.3.1 Geographic Confounding

Geographic confounding occurs when measured or unmeasured confounding factors vary by geographic location. Regional factors that contribute to geographic confounding are those associated with the exposure and associated with the outcome independently of the exposure and may include access to resources, community support, and policy influence among others. In observational studies, exposed individuals may be differentially geographically distributed compared to unexposed individuals. Ignoring this imbalance could lead to biased estimates of the treatment effect of interest.

1.3.2 Existing Spatial PSA Methods

Chagas et al. have developed a spatial propensity score matching method to address regional differences in sugarcane production in Brazil [26]. This work adopts a parsimonious approach in estimating the spatial propensity score, justified by sensitivity analysis. The method incorporates spatial information via spatial autocorrelation (SAC), spatial autoregressive (SAR) and spatial error models (SEMs). Additionally, the researchers explore spatial metrics such as distance to prominent landmark and an indicator for high density production. This work concludes that including spatial information is crucial to reduce bias.

Bayesian spatial-propensity score matching (BS-PSM) has recently been introduced by Gonzales et al. [27] as an extension of the regional-level spatial propensity score matching proposed by Chagas et al. BS-PSM maintains the goal of addressing uncertainty in the propensity score. Methods to address this uncertainty were previously explored in a non-spatial setting [28, 29, 30, 31]; however, Gonzales seeks to use Bayesian methods for proper standard error adjustment in a spatial setting. Utilizing spatial probit models to construct propensity score estimates, the authors then form matches based on a nearest neighbor algorithm that imposes a distance caliper (spatial caliper matching (SCM)) or neighbor requirement (spatial radius matching (SRM)) to avoid bad matches. Among the matched sample, a spatial average treatment effect (SATE) is estimated. Lastly, methods are applied to an application of the effect of microfinance in Bolivia.

These studies illustrate advances in PSA to incorporate spatial information; however, they conduct region-level analyses, failing to exploit valuable information at the

subject-level.

Chapter 2

Proposed Methodology

2.1 Overview

Building upon the recent work in multilevel and spatial PSA, we propose hierarchical spatial PSA to address geographic confounding, which occurs when measured or unmeasured confounding factors vary by geographic location. Patient-level covariates such as age, demographics, and comorbidities are traditional suspects in confounding; however, moreover, exposed individuals may in fact live in different regions than unexposed individuals.

Propensity score methods can be extended to the hierarchical spatial setting by incorporating spatial random effects into the propensity score and outcome models:

$$\text{logit}(e_{ij}) = \text{logit}[\Pr(Z_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{1i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \phi_{1i}, \tag{2.1}$$

where ϕ_{1i} is the spatial random effect for region i. Similarly, a logistic spatial model

for a binary outcome is expressed as

$$\text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{2i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + z_{ij}\alpha + \phi_{2i}, \tag{2.2}$$

where ϕ_{2i} denotes the spatial random effect for region i in the outcome model. To allow for maximal spatial smoothing, we assign the random effects ϕ_{1i} and ϕ_{2i} intrinsic conditional autoregressive (ICAR) priors [32] that take the conditional form

$$\phi_i \mid \boldsymbol{\phi}_{-i}, \sigma^2 \sim \text{N}\left(\frac{1}{m_i}\sum_{h\sim i}\phi_h, \sigma^2/m_i\right), \tag{2.3}$$

where $h \sim i$ indicates that region h is geographically adjacent to county i, m_i is the number of neighbors, and σ^2 is the conditional variance of ϕ_i given the remaining spatial effects, $\boldsymbol{\phi}_{-i}$. By way of this smooth spatial process modeling, we acknowledge and exploit the tendency of neighboring regions to be more similar than non-neighbors in terms of access to resources, regional policies and environmental conditions. Additionally, we allow for the estimation of region-level effects even when data in that region is sparse due to the allowance of "borrowing" information across adjacent regions.

Following Brook's Lemma [33], the joint distribution for $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)^T$ is given by

$$\pi(\boldsymbol{\phi} \mid \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\boldsymbol{\phi}^T \boldsymbol{Q} \boldsymbol{\phi}\right), \tag{2.4}$$

where $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{A}$ is a spatial structure matrix of rank $n-1$, with $\boldsymbol{M} = \text{diag}(m_1, \ldots, m_n)$ and $\boldsymbol{A}$ representing an $n \times n$ adjacency matrix with $a_{ii} = 0$, $a_{ih} = 1$ if $i \sim h$, and

$a_{ih} = 0$ otherwise. When a fixed intercept is included in the model, a sum-to-zero constraint must be applied to $\boldsymbol{\phi}$ to ensure an identifiable model.

Models 2.1 and 2.2 can now be used to construct a spatial version of the multilevel DR estimator suggested by Li et al. [9]. Such an extension would yield a global DR estimate between geographically balanced exposure groups. We could now assume ignorability of group assignment conditional on both observed individual-level covariates and spatial effect ϕ_{1i}. Alternatively, Model 2.1 can be used to generate propensity scores to be used in the setting of propensity score matching, creating a spatially balanced matched sample to be analyzed. Subsequently, Model 2.2 can be constructed using the matched sample, and an estimate of the risk difference could be derived using standardization.

In the case where the outcome of interest is not binary, a more complex spatial model may need to be adopted. Spatial matching can nevertheless be employed to ensure both patient-level and geographic balance in the resulting matched sample. It is then possible to derive a minimally biased effect estimate from the outcome model of choice.

2.2 Motivating Example

2.2.1 Background

Racial disparities in health outcomes persist even today despite decades of focus on deciphering the underlying causes [34]. For example, evidence consistently shows that racial minorities have a higher prevalence of diabetes, poorer diabetes outcomes,

higher risk of complications, and higher mortality rates compared to non-Hispanic whites [35, 36, 37]. While these disparities can be partially attributed to individual-level factors such as age, sex, marital status, and comorbidities [38, 39], there has been no so-called "silver bullet". It is plausible that many factors contribute modestly to observed disparities. More recent work has focused on the community, demonstrating that access to healthy food outlets and the availability of community health resources may also play a role [40, 41]. Therefore, incorporating geographically varying community factors in racial disparities analysis is important; however, some of these factors may be unmeasured due to lack of data availability or to their conceptual nature. Additionally, "race" is both an immutable characteristic and a socially and historically charged construct that has implications beyond phenotypic and biological characteristics. While the analysis of racial disparities must be sensitive and comprehensive and any conclusion drawn with restraint, researchers should not be dissuaded from seeking valuable knowledge that could help target vulnerable individuals.

2.2.2 Data Description

In order to study racial disparities in glycemic control, we obtain data for veterans with type 2 diabetes. Our analysis is based on a sample of 64,022 non-Hispanic Black (NHB) and non-Hispanic White (NHW) veterans with residential addresses in Alabama, Georgia or South Carolina in fiscal year (FY) 2014. Geographic boundaries are defined by the US Census county-level adjacency matrix irrespective of state membership [42]. This matrix contains $n = 272$ counties and 1528 pairwise adja-

cencies. In order to ascertain the severity of the disparity, we define poor glycemic control as indication of one or more hemoglobin A1c (HbA1c) measurements ≥ 8 in FY 2014. Within-county sample sizes range from 5 to 2,409, with a median of 108. Ten of the 272 counties in the study region have no NHB veterans. Overall, 36.5% of individuals in the study exhibit poor glycemic control (40.8% for NHBs, 33.2% for NHWs). We are able to identify potential patient-level confounders and can utilize indicated county of residence to incorporate geographic information.

To study the disparity in receiving diabetes education visits among type 2 diabetic veterans with poor glycemic control, we identify a sample of 20,636 NHB ($n = 9,277$) and NHW ($n = 11,359$) patients with a measure of HbA1c ≥ 8 in FY 2014. Once again, we restrict the sample to those veterans with residential addresses in Alabama, Georgia or South Carolina and utilize the same county-level adjacency matrix as in the study of poor glycemic control. Overall, approximately 13% of the patients in the sample receive a diabetes care visit following indication of poor control (15.0% for NHBs, 11.2% for NHWs).

Lastly, in order to assess the disparity in the number of inpatient days within the VA health care system, we identify veterans with type 2 diabetes and use their medical records to calculate the number of days spent as an inpatient in a VA facility in FY 2014. The data consist of observations for 23,533 NHB ($n = 9,695$) and NHW ($n = 13,838$) veterans with type 2 diabetes living in Georgia, Alabama and South Carolina in 2014. As the scope of the aforementioned studies in racial disparities among type 2 diabetic veterans has been restricted to the southeastern United States, we continue to use this geographic region and its associated county-level adjacency

matrix for this final study.

2.3 Specific Aims

The following section describes the structure of this research: it is contained in three separate but contiguous aims. Each aim is designed to stand alone while also complementing the other two. Uniquely, each aim will address an unanswered gap in the current literature, and cohesively, all three provide health researchers with a method to address geographic confounding with proof of concept in simulation, demonstration in application, and extension to current and innovative analysis methodology.

2.3.1 Aim 1

- To demonstrate the detriment of ignoring geographic confounding in causal inference by conducting simulation studies that reveal the bias, coverage, and root mean square error of estimates derived from exclusively subject-level models that do not account for geographic cluster
- To develop a method to ameliorate the aforementioned detriment by incorporating spatial random effects in the propensity score model stage and the outcome model stage of propensity score analysis and constructing a spatial doubly robust weighted estimator
- To examine the properties and performance of the novel methodology by conducting simulation studies that elucidate the bias, coverage, and root mean

square error of the spatial doubly robust weighted estimator under various degrees of spatial heterogeneity and per-region sample size

- To apply the methodology to a clinically relevant application by analyzing racial disparities in glycemic control, for which a balanced comparison between groups is desired, and examining the incremental effects of doubly robust estimation and spatial doubly robust estimation compared to an unadjusted estimate

2.3.2 Aim 2

- To develop methodology to incorporate spatial random effects in a propensity score matching framework

- To examine the performance of effect estimation in the spatially matched samples by conducting simulation studies under various degrees of spatial heterogeneity and per-region sample size

- To evaluate the effects of the inclusion of an outcome model for further regression adjustment compared to an unadjusted effect estimate

- To apply the methodology to a clinically relevant application by analyzing racial disparities in the receipt of diabetes education visits for which the ATT is the desired effect estimate and examining the incremental effects of estimation among a non-spatially matched sample and a spatially matched sample compared to an unadjusted estimate

2.3.3 Aim 3

- To demonstrate the utility of spatial propensity score matching in alternative analyses
- To utilize the methodology and conclusions developed in the prior aims to analyze a zero-inflated count outcome
- To combine spatial propensity score matching with a spatial negative binomial hurdle outcome model to derive a well-balanced, minimally biased estimate of clinically relevant ATTs

Chapter 3

Aim 1

Title: *Addressing geographic confounding through spatial propensity scores: a study of racial disparities in diabetes*

Authors: Melanie L. Davis, Brian Neelon, Paul J. Nietert, Kelly J. Hunt, Lane Burgette, Andrew B. Lawson, Leonard E. Egede

Status: Published in Statistical Methods in Medical Research (SMMR) November 2017

Abstract: Motivated by a study exploring differences in glycemic control between non-Hispanic black and non-Hispanic white veterans with type 2 diabetes, we aim to address a type of confounding that arises in spatially referenced observational studies. Specifically, we develop a spatial doubly robust (DR) propensity score estimator to reduce bias associated with geographic confounding, which occurs when measured or unmeasured confounding factors vary by geographic location, leading to imbalanced group comparisons. We augment the DR estimator with spatial random effects, which are assigned conditionally autoregressive priors to improve inferences by borrowing information across neighboring geographic regions. Through a series of simulations, we show that ignoring spatial variation results in increased absolute bias and mean squared error, while the spatial DR estimator performs well under various levels of spatial heterogeneity and moderate sample sizes. In the motivating application, we construct three global estimates of the risk difference between race groups: an unadjusted estimate, a DR estimate that adjusts only for patient-level information, and a hierarchical spatial DR estimate. Results indicate a gradual reduction in the risk difference at each stage, with the inclusion of spatial random effects providing a 20% reduction compared to an estimate that ignores spatial heterogeneity. Smoothed maps indicate poor glycemic control across Alabama and southern Georgia, areas comprising the so-called "stroke belt". These results suggest the need for community-specific interventions to target diabetes in geographic areas of greatest need.

3.1 Introduction

Diabetes is the seventh leading cause of death in the United States and is associated with a number of adverse health outcomes, including stroke, heart disease, kidney failure, and amputation [43]. Evidence consistently shows that racial minorities have a higher prevalence of diabetes, poorer diabetes outcomes, higher risk of complications, and higher mortality rates compared to non-Hispanic whites [35, 36, 37]. These disparities are explained in part by individual-level factors such as age, sex, marital status, and comorbidities [38, 39]. However, recent work has found that geographically varying community characteristics, such as access to healthy food outlets or the availability of community health resources, may also play a role [40, 41]. Given that racial disparity studies are inherently observational, it is critical to account for multiple sources of confounding, both at individual and neighborhood levels, in order to make comparisons between balanced race groups. This is especially relevant in diabetes research, as numerous recent studies have demonstrated associations between spatially varying confounding factors such as community environment and diabetes outcomes [44]. To obtain unbiased estimates of racial differences, it is necessary to account not only for individual-level confounding, but also *geographic confounding*, which occurs when the confounding factors, whether observed or unobserved, vary by geographic locations that share resources. The goal of this paper is to extend recent methods for multilevel causal inference to obtain minimally biased estimates of racial disparities in the presence of geographic confounding.

Propensity score analysis [8] (PSA) offers a principled approach to causal inference in observational studies, and has gained increasing traction in health dis-

parities studies in recent years [9, 45]. PSA is a multi-stage estimation strategy in which a propensity score model is first used to estimate the conditional probability of group assignment (i.e, the propensity score) given a set of covariates. The estimated propensity scores are then used to balance the groups according to important characteristics. Finally, an outcome model is fit in order to make balanced group comparisons. Common balancing methods include matching, stratification and inverse probability weighting. The balancing property of the propensity score ensures similar covariate distributions across groups under mild assumptions, allowing for a minimally confounded outcome analysis [2]. A particularly attractive weight-based estimator is the "doubly robust" (DR) estimator [46], which is a consistent estimator of the average treatment effect when either the propensity score model or the outcome model is correctly specified. Because racial identity is an immutable characteristic for which we desire a balanced comparison, the term "average controlled difference" is commonly used to denote the estimand of interest in racial disparity studies [9].

The central aim of this paper is to develop a spatial DR estimator that minimizes bias in the presence of observed and potentially unobserved geographic confounding. While there has been some recent work incorporating spatial information into PSA [26, 27, 47, 48], these methods have been limited to non-clustered data in which the response variable is a region-level proportion. Arpino and Mealli [25] and Li et al. [9] recently introduced PSA approaches for multilevel data. They fit propensity score and outcome models that included random effects to account for unobserved cluster-level confounding. Li et al. [9] additionally compared weighted estimators derived

from fixed and random effects models to demonstrate the benefit of incorporating cluster-level random effects in PSA, as well as the protective properties of the DR estimator. However, their approach did not incorporate spatial information.

Here, we propose a spatial DR estimator that incorporates available information at both the individual and region levels. We introduce a set of spatial random effects to account for variation due to unobserved geographic confounders. The random effects are assigned conditionally autoregressive (CAR) prior distributions that promote localized spatial smoothing by borrowing information from surrounding geographic areas. We adopt maximum likelihood (ML) as our initial estimation approach when fitting the spatial propensity score and outcome models. However, because ML-based numerical integration routines become unstable as the dimension of the random effects increases, we explore two alternative estimation methods: penalized quasi-likelihood and Bayesian inference. We conduct detailed simulation studies to compare the inferential properties of the three estimation methods under varying degrees of spatial heterogeneity. Finally, we apply the method to a study examining racial disparities in glycemic control among veterans with type 2 diabetes residing in the southeastern United States.

3.2 Spatial Propensity Score Analysis

3.2.1 Overview of Propensity Score Weighting Methods

We begin by briefly reviewing the inferential properties of PSA as outlined in Rosenbaum and Rubin [8] and summarized more recently in Lunceford and Davidian [14].

Let Z denote a group indicator taking values 0 or 1. In the context of clinical trials, Z commonly represents an assigned treatment group (e.g., $Z = 1$ if treated and 0 if control), while in epidemiologic settings, Z typically denotes a manipulable exposure group. In principle, Z can take more than two values, but since our focus in this paper is to estimate differences between only two groups, we assume throughout that Z is dichotomous. According to the causal framework outlined by Rubin [6], each individual is assumed to have two potential outcomes (Y_1, Y_0), where Y_1 and Y_0 denote the (potentially counterfactual) outcomes under $Z = 1$ and $Z = 0$, respectively. The observed response, Y, is given by $Y = ZY_1 + (1 - Z)Y_0$, so that $Y = Y_1$ if $Z = 1$ and $Y = Y_0$ otherwise. A common causal estimand of interest is the population average treatment effect (ATE), defined as $\Delta = E(Y_1) - E(Y_0)$. Because we observe only one of (Y_1, Y_0), unbiased estimation of the ATE, Δ, requires that we instead estimate the average effect conditional on *observed* treatment assignment, that is, $\Delta^* = E(Y_1|Z = 1) - E(Y_0|Z = 0)$.

In randomized controlled trials, the treatment groups are balanced with respect to relevant covariates, ensuring that the potential outcomes (Y_1, Y_0) are stochastically independent of the treatment assignment Z. In this case, $\Delta^* = \Delta$, and the observed treatment difference Δ^* serves as a suitable target for causal inference. In observational studies, however, the groups are not guaranteed to be balanced, and in this case we cannot conclude that $\Delta^* = \Delta$. Nevertheless, it may be reasonable to assume that (Y_1, Y_0) are conditionally independent of Z given a vector of covariates $\boldsymbol{X}$. This is commonly referred to as the "no unmeasured confounding" assumption [8]. Under this assumption, the ATE can be identified from the observed data $(Y, Z, \boldsymbol{X})$

through the equation

$$\begin{aligned}
\Delta &= E(Y_1) - E(Y_0) \\
&= E_{\boldsymbol{X}}[E(Y_1|\boldsymbol{X} = \boldsymbol{x}) - E(Y_0|\boldsymbol{X} = \boldsymbol{x})] \\
&= E_{\boldsymbol{X}}[E(Y_1|\boldsymbol{X} = \boldsymbol{x}, Z = 1) - E(Y_0|\boldsymbol{X} = \boldsymbol{x}, Z = 0)] \\
&= E_{\boldsymbol{X}}[E(Y|\boldsymbol{X} = \boldsymbol{x}, Z = 1) - E(Y|\boldsymbol{X} = \boldsymbol{x}, Z = 0)],
\end{aligned} \tag{3.1}$$

where the outer expectation is taken with respect to the distribution of covariates $\boldsymbol{X}$ in the entire population and $\boldsymbol{x}$ is an observed realization of the random variable $\boldsymbol{X}$. The third line of equation (3.1) follows from the conditional independence of (Y_1, Y_0) and Z under no unmeasured confounding, and the last line follows from the fact that $Y_k = Y$ if $Z = k$ $(k = 0, 1)$. Consequently, causal inference regarding the ATE can be made using the observed data.

When the "treatment" variable is an immutable characteristic such as race, the potential outcomes framework is not strictly applicable, since there is no well-defined potential outcome corresponding to an alternative race designation. This precludes formal causal inference in the context of racial disparity studies. In this setting, Li et al. [9] propose using the average controlled difference (ACD) as a *descriptive* estimand analogous to the ATE, where the ACD is defined as

$$\Delta = E_{\boldsymbol{X}}[E(Y|\boldsymbol{X} = \boldsymbol{x}, Z = 1) - E(Y|\boldsymbol{X} = \boldsymbol{x}, Z = 0)]. \tag{3.2}$$

Because the latter expression is identical to the last line of equation (3.1), we use Δ

throughout to denote either the ATE and ACD. However, the former is a causal estimand, whereas the latter is a purely descriptive one. When there is no unmeasured confounding, the ACD represents a population-average difference between two fully adjusted comparison groups. Although our focus here is on the ACD, the methods described below can equally apply to settings where the ATE is a more natural target of inference.

Under unconfoundedness, propensity score methods can be used to derive unbiased estimators of the ATE or ACD in observational studies. The propensity score, $e(\boldsymbol{x}) = \Pr(Z = 1|\boldsymbol{X} = \boldsymbol{x})$, is the conditional probability of exposure given $\boldsymbol{X}$, where the so-called "overlap" condition, $0 < e(\boldsymbol{x}) < 1$, is assumed to hold. Rosenbaum and Rubin [8] established that $e(\boldsymbol{x})$ functions as a balancing score such that

$$\Delta = E\left[\frac{ZY}{e(\boldsymbol{x})} - \frac{(1-Z)Y}{1-e(\boldsymbol{x})}\right], \tag{3.3}$$

when both the overlap and unconfoundedness assumptions hold. Hence, an unbiased, Horvitz-Thompson type [49] estimator can be obtained by correctly specifying a propensity score model. The propensity scores are typically estimated using a logistic regression model of the form

$$\text{logit}(e_i) = \text{logit}[e(\boldsymbol{x}_i)] = \text{logit}[\Pr(Z_i = 1|\boldsymbol{X}_i = \boldsymbol{x}_i)] = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad i = 1, \ldots, n. \tag{3.4}$$

If model (3.4) is correctly specified, an unbiased, inverse-probability weight (IPW)

estimator of the ACD is given by

$$\hat{\Delta}^{IPW} = n^{-1} \sum_{i=1}^{n} \left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} \right], \tag{3.5}$$

where $\hat{e}_i$ denotes the estimated propensity score for subject i. To guard against misspecification of the propensity score model, Robins et al.[17] developed a semi-parametric doubly robust (DR) estimator of the form

$$\begin{aligned} \hat{\Delta}^{DR} &= \frac{1}{n} \sum_{i=1}^{n} \hat{\Delta}_i \\ \hat{\Delta}_i &= \left[\frac{Z_i Y_i}{\hat{e}_i} - \frac{(Z_i - \hat{e}_i) \hat{Y}_{i1}}{\hat{e}_i} \right] - \left[\frac{(1 - Z_i) Y_i}{1 - \hat{e}_i} + \frac{(Z_i - \hat{e}_i) \hat{Y}_{i0}}{1 - \hat{e}_i} \right], \end{aligned} \tag{3.6}$$

where $\hat{Y}_{i1}$ and $\hat{Y}_{i0}$ are predicted outcomes obtained by regressing Y on $\boldsymbol{X}$ and Z, the former including the regression coefficient for Z and the latter excluding it. The doubly robust property derives from the fact that expression (3.6) is a consistent estimator of Δ if either the propensity model or the outcome model is correctly specified. The large-sample approximate variance of $\hat{\Delta}^{DR}$ is given by

$$s^2 = \frac{1}{n^2} \sum_{i=1}^{n} (\hat{\Delta}_i - \hat{\Delta}^{DR})^2. \tag{3.7}$$

Alternatively, bootstrapping by resampling with replacement can be used to estimate the standard error and associated confidence intervals.

3.2.2 A Doubly Robust Estimator for Hierarchical Spatial Data

Li et al. [9] recently extended the DR estimator to the multilevel setting, where $(Y_{ij}, Z_{ij}, \boldsymbol{X}_{ij})$ denote the data for the j-th subject in cluster i. Li et al. propose the following hierarchical DR estimator of the ACD:

$$\begin{aligned} \hat{\Delta}^{DR} &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\hat{\Delta}_{ij} \\ \hat{\Delta}_{ij} &= \left[\frac{Z_{ij}Y_{ij}}{\hat{e}_{ij}} - \frac{(Z_{ij}-\hat{e}_{ij})\hat{Y}_{ij1}}{\hat{e}_{ij}}\right] - \left[\frac{(1-Z_{ij})Y_{ij}}{1-\hat{e}_{ij}} + \frac{(Z_{ij}-\hat{e}_{ij})\hat{Y}_{ij0}}{1-\hat{e}_{ij}}\right], \end{aligned} \quad (3.8)$$

where e_{ij} denotes the propensity score for the (ij)-th individual, $N = \sum_{i=1}^{n} n_i$, and n_i is the sample size of the i-th cluster. Generalized linear mixed models are used to estimate e_{ij}, Y_{ij0}, and Y_{ij1}, with the random effects accommodating between-cluster heterogeneity and accounting for smoothly varying, unobserved cluster-level confounders. Using simulation studies, Li et al. demonstrate that incorporating the random effects yields improved inferences over models that ignore cluster-level variation or treat the cluster indicators as fixed effects. Analogous to equation (3.7), the large-sample variance estimator of $\hat{\Delta}^{DR}$ is given by

$$s^2 = \frac{1}{N^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}(\hat{\Delta}_{ij} - \hat{\Delta}^{DR})^2. \quad (3.9)$$

The multilevel estimator proposed by Li et al. is readily extended to the spatial setting by augmenting the propensity score and outcome models with spatial random

effects, resulting in a spatial version of the DR estimator given in equation (3.8). Turning to our motivating application, let Y_{ij} denote the presence of poor glycemic control for the j-th individual residing in the i-th county, let Z_{ij} denote an indicator variable taking a value of 1 if the individual is non-Hispanic black (NHB) and 0 if non-Hispanic white (NHW), and let $\boldsymbol{x}_{ij}$ represent a set of patient-level covariates. The spatial propensity score model is given by

$$\text{logit}(e_{ij}) = \text{logit}[\Pr(Z_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{1i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \phi_{1i}, \tag{3.10}$$

where ϕ_{1i} is the spatial random effect for county i. Similarly, the spatial outcome model is expressed as

$$\text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{2i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + z_{ij}\alpha + \phi_{2i}, \tag{3.11}$$

where ϕ_{2i} denotes the spatial random effect for county i in the outcome model. The spatial random effects can represent geographic variability in health care access, availability of community outreach and medical education programs, or access to other resources that may be associated with both race and diabetes management. To encourage maximal spatial smoothing, we assign each of the random effects ϕ_{1i} and ϕ_{2i} an intrinsic conditional autoregressive (ICAR) prior [32] that takes the conditional form

$$\phi_{ki} \mid \boldsymbol{\phi}_{k(-i)}, \sigma_k^2 \sim \text{N}\left(\frac{1}{m_i}\sum_{h \sim i} \phi_{kh}, \sigma_k^2/m_i\right), \quad k = 1, 2, \tag{3.12}$$

where $h \sim i$ indicates that county h is a geographic neighbor of county i, m_i is

the number of neighbors, and, for model k, σ_k^2 is the conditional variance of ϕ_{ki} given the remaining spatial effects, $\boldsymbol{\phi}_{k(-i)}$. Modeling between-county heterogeneity via a smooth spatial process is beneficial for two reasons. First, it recognizes the inherent tendency for neighboring regions to share health resources or experience similar environmental pressures that can lead to poor health outcomes. Second, it improves estimation of region-level effects by borrowing information from neighboring areas, thus reducing uncertainty in estimating the propensity scores and predicting the potential outcomes used to derive the spatial DR estimator.

Following Brook's Lemma [33], the joint distribution for $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kn})^T$ is given by

$$\pi(\boldsymbol{\phi}_k \mid \sigma_k^2) \propto \exp\left(-\frac{1}{2\sigma_k^2}\boldsymbol{\phi}_k^T \boldsymbol{Q} \boldsymbol{\phi}_k\right), \quad k = 1, 2, \tag{3.13}$$

where $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{A}$ is a spatial structure matrix of rank $n-1$, with $\boldsymbol{M} = \text{diag}(m_1, \ldots, m_n)$ and $\boldsymbol{A}$ representing an $n \times n$ adjacency matrix with $a_{ii} = 0$, $a_{ih} = 1$ if $i \sim h$, and $a_{ih} = 0$ otherwise. When a fixed intercept is included in the model, a sum-to-zero constraint must be applied to $\boldsymbol{\phi}_k$ to ensure an identifiable model.

3.2.3 Model Fitting and Inference

Because the DR estimator is a frequentist estimator, we adopt maximum likelihood as our default estimation approach. Maximum likelihood for models (3.10) and (3.11) can be easily implemented using off-the-shelf software such as SAS `PROC GLIMMIX` [50]. Maximum likelihood is selected by specifying the `METHOD=QUAD` option, which combines adaptive Gauss-Hermite quadrature for numerical integration with Newton-Raphson routines for maximization. The spatial covariance matrix is

introduced by first computing the Moore-Penrose generalized inverse of the structure matrix $\boldsymbol{Q}$ in expression (4.4), and then incorporating this as part of a user-defined covariance matrix in `PROC GLIMMIX`. Details can be found in Rasmussen [51]. The Moore-Penrose inverse is unique and serves the dual purpose of imposing the identifiability restriction $\sum_{i=1}^{n} \phi_i = 0$. Although adaptive quadrature tends to work well for low-dimensional random effects models (e.g., random intercept models), it becomes computationally burdensome as the dimension of the random effects grows, since an increasing number of quadrature points is required to accurately estimate the multivariate random effect distribution. For example, adaptive quadrature can pose challenges for models that include spatially varying covariates.

To address this potential limitation, we consider two computationally tractable estimation strategies: penalized quasi-likelihood (PQL) and Bayesian inference. PQL [52, 53] is an iterative estimation procedure achieved through Taylor series expansions of the response about current estimates of the fixed and random effects [54]. The expansion yields a "pseudo-response" that is linear in the model parameters. A linear mixed model is then fit to the pseudo-response using restricted maximum likelihood, thus avoiding computationally challenging numerical integration routines. PQL for the spatial propensity score and outcome models can be fit in `PROC GLIMMIX` using the default `METHOD=RSPL` option for restricted pseudo-maximum likelihood estimation.

Finally, we consider Bayesian estimation, the most common inferential approach for fitting spatial CAR models. Here, the propensity score and outcome models are estimated separately using approximate Bayesian methods. The propensity score e_{ij} is estimated using the posterior mean of the linear predictors from the propensity

score model given in equation (3.10). Likewise, the potential outcomes $\hat{Y}_{ij1}$ and $\hat{Y}_{ij0}$ are estimated (or, more accurately, "predicted") using the posterior mean linear predictors from the outcome model, the former including the posterior mean for α in equation (3.11) and the latter excluding it. The resulting estimates and predictions are fed into the spatial DR estimator for final inferences. In this context, the Bayesian approach should be viewed simply as an alternative way to estimate the propensity score e_{ij} and predict the potential outcomes Y_{ij0} and Y_{ij1} when forming the DR estimator. The DR estimator itself is a large-sample frequentist estimator, and hence our overall inferential approach should once again be regarded as frequentist. By fitting separate propensity score and outcome models, we avoid the so-called "feedback" issue that can arise when the models are fitted jointly under a fully Bayesian approach [55]. For our application, we adopt the efficient integrated nested Laplace approximation (INLA) proposed by Rue et al. [56]. INLA uses a Laplace approximation to estimate the joint posterior of the model parameters, yielding improved computational capacity over standard Markov chain Monte Carlo routines. This method can be easily implemented in the R package `INLA` (www.r-inla.org), where the `Besag` option is used to specify the ICAR prior. As a default, we assign weakly informative $\mathrm{N}(0, 1\mathrm{e}5)$ priors to fixed effects and $\mathrm{Ga}(1, 5\mathrm{e}\text{-}05)$ priors for the spatial precision (i.e., inverse variance) terms, where $\mathrm{Ga}(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b. To investigate sensitivity to prior specification, in our case study we consider alternate priors per the recommendation of Carroll et al. [57] for the regression coefficients and spatial variances. Alternative prior specifications are discussed in Section 3.4.

3.3 Simulation Study

3.3.1 Data Description

To examine the performance of the proposed spatial DR estimator, we conduct a series of simulation studies. The goals were to 1) examine the inferential properties (e.g., bias, 95% coverage) of the proposed spatial DR estimator under varying sample sizes and degrees of spatial heterogeneity; 2) explore the impact of ignoring spatial heterogeneity during model fitting; and 3) compare the performance of the three estimation strategies described in the previous section. Additionally, we conducted a sub-study to assess the ability of the spatial DR estimator to capture the true ACD when important spatially varying covariates were ignored during model fitting. To emulate the geographic structure in our application, we used the US Census county-level adjacency matrix for South Carolina, Georgia, and Alabama [42]. This matrix contains $n = 272$ counties and 1528 pairwise adjacencies. For the primary study, we generated 100 datasets from the following propensity score and outcome models:

$$\begin{aligned} \text{logit}[\Pr(Z_{ij} = 1 | X_{ij} = x_{ij}, \phi_{1i})] &= \beta_0 + x_{ij}\beta_1 + \phi_{1i} \quad (3.14) \\ \text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, X_{ij} = x_{ij}, \phi_{2i})] &= \gamma_0 + x_{ij}\gamma_1 + z_{ij}\alpha + \phi_{2i}, \quad (3.15) \\ i = 1, \ldots, 272;\ j = 1, \ldots, n_i, & \end{aligned}$$

where X_{ij} was simulated according to a $\text{N}(5, 2)$ distribution; the fixed effect coefficients were set at $\beta_0 = 0.25$, $\beta_1 = $ -0.15, $\gamma_0 = 0.35$, $\gamma_1 = $ -0.50, and $\alpha = 0.90$; n_i was allowed to take on three values: 25, 50, and 100; and ϕ_{1i} and ϕ_{2i} were simulated

from ICAR models given in equation (4.4) with $\sigma^2_{\phi 1}$ and $\sigma^2_{\phi 2}$ each taking values 1, 4, and 9 to represent increasing degrees of spatial variation. These parameter values yielded an average risk difference of approximately 0.10, which follows the existing literature on disparities in glycemic control [58]. We also examined scenarios where both of the above models excluded spatial effects, in order to examine the behavior of the spatial DR estimator when the data exhibited no spatial heterogeneity.

To accomplish the aim of our sub-study, we augmented the models in equations (3.14) and (3.16) to include a county-level covariate generated according to a $N(10, 3)$ distribution and an additional spatially smoothed county-level covariate simulated according to the ICAR model given in equation (4.4), with $\sigma^2 = 2$ and coefficients β_2 = -0.1 and $\beta_3 = 0.1$ for the respective spatially varying covariates in the propensity score model and $\gamma_2 = 0.3$ and γ_3 = -0.3 for the respective spatially varying covariates in the outcome model. Spatial variances for ϕ_{1i} and ϕ_{2i} were each set to the intermediate level of 4.

3.3.2 Results

Table 3.1 summarizes the performance of the spatial DR estimator when the data were generated according to the random intercept propensity score and outcome models given in equations (3.14) and (3.16). Rows indicate the varying levels of spatial heterogeneity (σ^2_ϕ) and sample sizes (n_i) used to generate the data, including the case where the simulated data contained no spatial heterogeneity ($\sigma^2_\phi = 0$). Columns delineate the mean absolute bias, RMSE, and 95% coverage of the estimated ACDs under the three estimation strategies.

Several trends emerge from the simulations. First, when the generated data contained no spatial heterogeneity, the spatial DR estimator performed well, with negligible bias, low RMSE, and near nominal coverage. For example, when $n_i = 100$ and maximum likelihood estimation was used, the bias under the spatial DR estimator was 0.003, with 95% coverage equal to 0.97. These trends continued even as the sample size decreased. Under $n_i = 25$, for instance, the bias ranged from 0.007 to 0.008 across the three estimation approaches.

Second, as the spatial heterogeneity in the data increased, the spatial DR estimator continued to perform well, whereas the non-spatial estimator displayed increasingly poor performance. For example, under maximum likelihood, the 95% coverage for the spatial model was 0.96 when $n_i = 100$ and $\sigma^2_\phi = 1$, and 0.92 when $n_i = 100$ and $\sigma^2_\phi = 9$. In contrast, the non-spatial models showed poor coverage whenever spatial heterogeneity was present. For example, with $n_i = 100$, the coverage under maximum likelihood for the non-spatial estimator decreased from 0.57 when $\sigma^2_\phi = 1$ and to 0.16 when $\sigma^2_\phi = 9$. As sample size decreased, bias and RMSE of the spatial DR estimator increased but remained favorable, particularly in contrast to the non-spatial DR estimator. The coverage of the DR estimator also remained near nominal levels as n_i decreased, except in the most extreme scenario in which $n_i = 25$ and $\sigma^2_\phi = 9$, where the coverage under maximum likelihood fell to 0.85. However, this was vastly higher than the 0.24 coverage observed for the non-spatial estimator.

Table 3.2 demonstrates the doubly robust property of the spatial DR estimator. As in Table 3.1, rows delineate varying degrees of spatial heterogeneity and county sample sizes. Columns indicate which of the two models, the propensity score or

the outcome model, was misspecified by excluding a spatial random intercept. In general, correctly specifying either the spatial propensity score or outcome model resulted in low bias and RMSE, confirming the doubly robust property of the proposed estimator. Not surprisingly, as the spatial heterogeneity increased to extreme levels (e.g., $\sigma^2_\phi = 9$), misspecifying one of the models led to modest increases in bias and RMSE. These increases were more prominent when the outcome model was misspecified, a result consistent with previous work suggesting more deleterious consequences for misspecifying the outcome model rather than the propensity score model in hierarchical settings [9].

Across all scenarios, the three estimation strategies yielded similar results, suggesting that any of the three approaches can be adopted in practice. However, if a secondary aim is to explore spatial heterogeneity in the outcome model, our experience suggests that INLA yields smoother and indeed more accurate predictions of spatial effects (e.g., ϕ_{2i} in equation (3.11)) than the other two estimation methods. Thus, if a subsequent goal is spatial prediction, as in our application, we recommend working with INLA throughout, or, alternatively, using a frequentist procedure to estimate the ACD and Bayesian methods for spatial prediction in subsequent analyses involving the outcome model.

Table 3.3 presents results of the sub-study using INLA to estimate the propensity score and outcome models. The goal of the sub-study was to assess the ability of the proposed spatial doubly robust estimator to capture the true ACD when relevant county-level covariates were left out of the analysis. The non-spatial analysis ignored space entirely, whereas the intermediate spatial analysis included a spatial random

effect in both the propensity score and outcome models yet treated the spatially varying covariates as unmeasured. The benchmark analysis fit the true models that included both the spatial random effects and the spatially varying county-level covariates. As Table 3.3 indicates, the non-spatial analysis performed poorly, whereas the spatial analysis that included only random intercepts retained favorable properties across all scenarios, including low bias, low RMSE and near-nominal coverage. As expected, we observed good performance under the benchmark analysis that included the fixed county-level covariates in addition to the spatial intercepts. Overall, there does not appear to be much difference between the spatial and benchmark models. These results support the use of the proposed spatial DR estimator, as it appears to capture the true risk difference even when county-level fixed effects are ignored during model fitting. As it is not uncommon for these covariates to be unavailable to the analyst, the spatial DR estimator provides a practically useful strategy to account for unmeasured geographic confounding.

3.4 Analysis of Racial Disparities in Glycemic Control

Our work was motivated by a study examining racial disparities in glycemic control among veterans with type 2 diabetes. The goal of the study was two-fold: first, to estimate racial disparities in poor glycemic control while accounting for relevant patient information and spatial variation; and second, to identify counties with high rates of poor glycemic control across the study region. Our analysis was based on

Table 3.1: Simulation results for random intercept models: Mean bias, RMSE, and 95% coverage of the non-spatial and spatial DR estimators under various sample sizes, spatial variances, and estimation methods. "Spatial" models included random intercepts in both the propensity score and outcome models; "non-spatial" excluded spatial effects in both models.

	Maximum Likelihood						Penalized Quasi-likelihood						Bayesian					
	Non-spatial			Spatial			Non-spatial			Spatial			Non-spatial			Spatial		
	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]
$n_i = 100$																		
$\sigma^2_\phi = 0$	0.003	0.004	97	0.003	0.004	97	0.003	0.004	97	0.003	0.004	97	0.004	0.004	98	0.004	0.005	97
$\sigma^2_\phi = 1$	0.009	0.013	57	0.004	0.005	96	0.009	0.013	57	0.004	0.005	96	0.011	0.014	57	0.004	0.006	89
$\sigma^2_\phi = 4$	0.027	0.035	24	0.004	0.005	96	0.027	0.035	24	0.004	0.005	96	0.029	0.039	26	0.005	0.007	93
$\sigma^2_\phi = 9$	0.044	0.057	16	0.005	0.007	92	0.044	0.057	16	0.005	0.007	92	0.048	0.063	14	0.005	0.007	94
$n_i = 50$																		
$\sigma^2_\phi = 0$	0.005	0.007	96	0.005	0.007	97	0.005	0.007	96	0.005	0.007	97	0.006	0.007	96	0.005	0.007	92
$\sigma^2_\phi = 1$	0.011	0.014	76	0.005	0.007	96	0.010	0.014	76	0.005	0.007	96	0.010	0.013	76	0.006	0.007	92
$\sigma^2_\phi = 4$	0.029	0.040	39	0.007	0.008	94	0.029	0.040	39	0.007	0.008	94	0.027	0.034	33	0.007	0.009	87
$\sigma^2_\phi = 9$	0.049	0.067	17	0.008	0.009	91	0.049	0.067	17	0.008	0.009	91	0.044	0.055	20	0.008	0.009	93
$n_i = 25$																		
$\sigma^2_\phi = 0$	0.007	0.009	95	0.007	0.009	95	0.007	0.009	95	0.007	0.009	96	0.008	0.010	92	0.008	0.010	93
$\sigma^2_\phi = 1$	0.011	0.014	83	0.008	0.010	94	0.011	0.014	83	0.008	0.010	93	0.013	0.016	77	0.009	0.011	89
$\sigma^2_\phi = 4$	0.029	0.037	40	0.008	0.011	92	0.029	0.037	40	0.008	0.011	91	0.029	0.036	42	0.011	0.013	85
$\sigma^2_\phi = 9$	0.061	0.077	24	0.010	0.013	85	0.048	0.060	26	0.010	0.013	85	0.045	0.057	30	0.010	0.012	83

[1] Mean absolute bias
[2] Root mean squared error (RMSE)
[3] 95% coverage

Table 3.2: Mean bias, RMSE, and 95% coverage of the partially misspecified DR estimators under various sample sizes, spatial variances, and estimation methods. Columns indicate which model was misspecified.

	Maximum Likelihood						Penalized Quasi-likelihood						Bayesian					
	Propensity Score†			Outcome‡			Propensity Score			Outcome			Propensity Score			Outcome		
	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]	B[1]	R[2]	C[3]
$n_i = 100$																		
$\sigma^2_\phi = 1$	0.004	0.005	93	0.004	0.005	95	0.004	0.005	93	0.004	0.005	95	0.004	0.005	95	0.004	0.005	96
$\sigma^2_\phi = 4$	0.004	0.005	94	0.006	0.007	92	0.004	0.005	94	0.006	0.007	92	0.005	0.006	82	0.006	0.007	94
$\sigma^2_\phi = 9$	0.005	0.006	83	0.009	0.012	83	0.005	0.006	83	0.009	0.012	83	0.005	0.007	77	0.010	0.012	77
$n_i = 50$																		
$\sigma^2_\phi = 1$	0.006	0.007	91	0.006	0.007	94	0.006	0.007	91	0.006	0.007	94	0.006	0.008	91	0.006	0.008	90
$\sigma^2_\phi = 4$	0.007	0.009	85	0.009	0.011	84	0.007	0.008	84	0.009	0.011	84	0.007	0.009	88	0.008	0.009	96
$\sigma^2_\phi = 9$	0.007	0.009	83	0.014	0.019	72	0.007	0.009	83	0.014	0.020	72	0.007	0.008	86	0.012	0.015	82
$n_i = 25$																		
$\sigma^2_\phi = 1$	0.008	0.010	91	0.008	0.010	91	0.008	0.010	91	0.008	0.010	91	0.008	0.010	95	0.008	0.010	97
$\sigma^2_\phi = 4$	0.009	0.013	86	0.012	0.016	82	0.009	0.013	86	0.012	0.016	82	0.009	0.011	81	0.010	0.013	93
$\sigma^2_\phi = 9$	0.012	0.014	75	0.021	0.026	65	0.012	0.014	75	0.021	0.026	65	0.009	0.012	86	0.018	0.024	72

† Propensity score model (3.14) misspecified (no random intercept), outcome model (3.16) correctly specified.
‡ Outcome model (3.16) misspecified (no random intercept), propensity score model (3.14) correctly specified.
[1] Mean absolute bias
[2] Root mean squared error (RMSE)
[3] 95% coverage

Table 3.3: Results for the sub-study: Mean bias, RMSE, and 95% coverage of the non-spatial, spatial, and "benchmark" doubly robust estimators under various sample sizes. The spatial DR estimator included a spatial random intercept in the propensity score and outcome models but ignored the spatially varying covariates. The benchmark DR estimator incorporated the spatially varying covariates in addition to the spatial random intercept in the propensity score and outcome models.

	Non-spatial			**Spatial**			**Benchmark**		
Sample size	Bias	RMSE	Coverage	Bias	RMSE	Coverage	Bias	RMSE	Coverage
$n_i = 100$	0.032	0.044	24	0.006	0.011	90	0.006	0.013	91
$n_i = 50$	0.034	0.049	31	0.007	0.008	93	0.006	0.008	93
$n_i = 25$	0.031	0.046	47	0.010	0.020	88	0.010	0.018	85

a sample of 64 022 NHB and NHW veterans with residential addresses in Alabama, Georgia or South Carolina. Poor glycemic control was defined as having at least one hemoglobin A1c (HbA1c) measurement ≥ 8 in fiscal year 2014. Study details have been reported elsewhere [59]; here, we summarize key features of the data. Within-county sample sizes ranged from 5 to 2409, with a median of 108. Ten of the 272 counties in the study region had no NHB veterans. This does not pose a problem for estimating the county-level spatial effects, since the smoothing property of the ICAR prior provides the necessary shrinkage to ensure reliable county-specific estimates. Overall, 36.5% of individuals in the study exhibited poor glycemic control (40.8% for NHBs, 33.2% for NHWs). Table 3.4 displays the variables that were included in the propensity score and outcome models. These variables include demographic information and comorbidities that have been shown to be associated with poor glycemic control [38].

In order to visualize geographic differences in racial distribution and poor glycemic control, we aggregated the data to the county level and constructed unadjusted maps of raw percents of NHBs and poor glycemic control by county (Figure 3.1, first col-

umn). Additionally, we assembled maps of local indicators of spatial association (LISA) to identify clusters and outliers of high and low percent NHBs and uncontrolled HbA1c (Figure 3.1, second column). Using local Moran's I tests with an α level of 0.10, we classified counties into four types: "high-high" clusters, defined as counties with significantly high rates of NHBs (top row) or uncontrolled HbA1c (bottom row) surrounded by other counties with significantly elevated rates of these variables; "high-low" outliers, defined as counties with significantly high rates of NHBs or uncontrolled HbA1c surrounded by neighboring counties with significantly low rates; and "low-high" outliers and "low-low" clusters, which were defined analogously. All other counties exhibited non-significant spatial effects. The results indicate distinct geographical patterns in racial distribution and poor glycemic control. There are several clusters with high percentages of NHB veterans, primarily in South Carolina, western Georgia, and central Alabama (Figure 3.1, top row). Many of these same areas also exhibited above-average uncontrolled HbA1c (Figure 3.1, bottom row), particularly western portions of Georgia and Alabama. In contrast, counties in northern Georgia exhibited below-average percents of NHB veterans and uncontrolled HbA1c. These patterns point to potential associations between residential location, race, and poor glycemic control, suggesting that geographic confounding may be present in this study. Spearman's correlation between percent NHBs and percent uncontrolled HbA1c across the counties was 0.224 (p-value = 0.0002), further supporting this conclusion.

Next, we compared the covariate balance between NHB and NHW veterans in unweighted, non-spatial propensity score weighted, and spatial propensity score

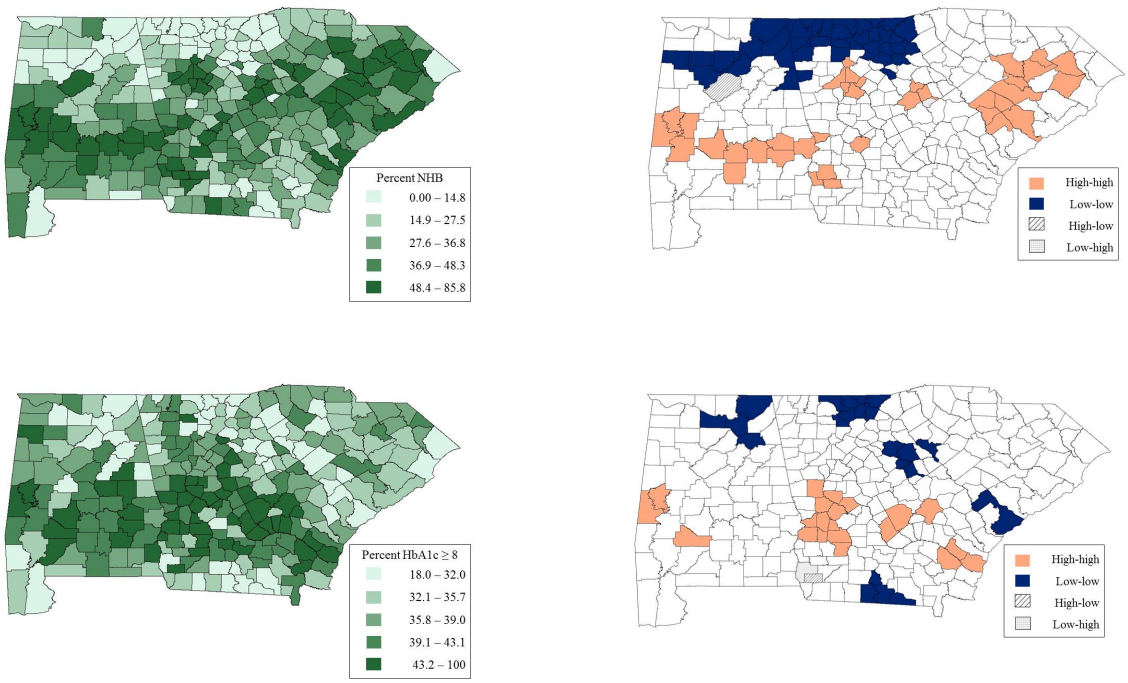

Figure 3.1: Unadjusted percents and local indicators of spatial association (LISA) for NHB and poor glycemic control. Top left: Unadjusted percent NHB; Top Right: NHB LISA; Bottom Left: Unadjusted percent poor glycemic control; Bottom Right: Poor glycemic control LISA.

weighted samples. To construct the non-spatially weighted sample, we fit a logistic propensity score model that included only the fixed patient-level covariates described in Table 3.4. To construct the spatially weighted sample, we fit a logistic propensity score model that included these same covariates as well a spatial intercept. We then used the subject-specific weights to form weighted means and proportions across the covariates [60]. Standardized differences were used to compare the covariate distributions across the two race groups [2]. We also derived county-specific weighted proportions of NHB and NHW veterans and mapped the distribution of the un-

weighted, non-spatially weighted and spatially weighted proportions. If the spatial propensity score model is adequately specified, the weighted covariate distributions and spatial patterns should be similar across race groups.

The results are presented in Table 3.4 and Figure 3.2. As Table 3.4 indicates, the weighted samples showed vastly improved balance compared to the unweighted sample, suggesting a well-specified propensity score model at the patient level. Figure 3.2 shows the spatial distribution of NHB and NHW veterans under the unweighted, non-spatially weighted and spatially weighted samples. In both the unweighted and non-spatially weighted samples, the spatial distribution of NHB and NHW veterans varied substantially. For example, a larger proportion of NHB veterans lived in central Georgia and central and western Alabama, whereas a larger proportion of NHW veterans lived in northern Alabama and Georgia. This spatial imbalance is not surprising since the spatially unweighted samples fail to account for differences in the spatial distribution of the two race groups. After spatial weighting, the spatial distribution of NHB veterans more closely resembled that of NHW veterans (unweighted Spearman correlation between race groups = 0.602, non-spatially weighted Spearman correlation = 0.629, and spatially weighted Spearman correlation = 0.996). These results highlight the need to balance on both individual- and county-level factors when groups differ with respect to both sets of characteristics.

Next, we derived three estimates of the global average controlled risk difference between NHBs and NHWs: an unadjusted estimate, and non-spatial DR estimate, and spatial DR estimate. To construct the non-spatial estimate, we fit propensity score and outcome models that included only the fixed covariates described in Table

Table 3.4: Balance of covariates between NHB and NHW veterans in unweighted, non-spatially weighted, and spatially weighted samples; "Stand. Diff." denotes the absolute value of the standardized difference.

	Unweighted			Non-spatially Weighted			Spatially Weighted		
Variable	NHB	NHW	Stand. Diff.	NHB	NHW	Stand. Diff.	NHB	NHW	Stand. Diff.
Age	64.21	69.96	0.596	68.20	67.51	0.066	68.04	67.40	0.062
Male	93.21	97.59	0.210	96.11	95.37	0.037	96.03	95.08	0.046
Service Percent ≥ 50	46.74	36.31	0.213	41.18	40.96	0.004	40.36	40.98	0.013
Married	54.86	68.49	0.283	58.00	65.93	0.164	58.02	64.69	0.137
Urban	70.31	52.15	0.379	58.57	59.47	0.018	59.95	59.50	0.009
Substance Abuse	10.52	4.43	0.233	6.83	6.80	0.001	6.85	6.84	0.000
Anemia	3.60	3.16	0.024	3.75	3.30	0.024	3.63	3.29	0.019
Cancer	2.79	2.92	0.008	3.17	2.78	0.023	2.98	2.72	0.016
Cerebrovascular Disease	4.16	3.83	0.017	4.20	3.92	0.014	4.15	3.92	0.012
Congestive Heart Failure	8.50	9.34	0.029	9.41	8.94	0.016	9.30	8.83	0.016
Cardiovascular Disease	9.14	16.26	0.215	13.89	13.23	0.019	13.82	13.26	0.016
Depression	34.68	26.24	0.184	31.11	31.04	0.002	30.76	31.14	0.008
Hypertension	87.12	82.24	0.136	84.36	83.66	0.019	83.73	83.45	0.008
Liver Disease	3.95	2.92	0.057	3.44	3.26	0.010	3.47	3.19	0.016
Lung Conditions	12.62	18.21	0.155	16.62	15.82	0.022	16.50	15.83	0.018
Electrolyte Diseases	6.15	4.47	0.075	5.46	5.19	0.012	5.18	5.18	0.000
Obesity	23.73	20.33	0.082	21.66	21.55	0.003	21.59	21.46	0.003
Psychoses	7.54	3.41	0.182	5.21	5.15	0.003	5.20	5.15	0.002
Peripheral Vascular Disease	6.78	8.80	0.075	8.62	7.90	0.026	8.68	7.77	0.033
Other Disease	21.39	16.28	0.131	18.75	18.25	0.013	18.19	18.23	0.001

3.4. Given our dual aims of estimating the ACD and conducting subsequent spatial analysis of uncontrolled HbA1c, we adopted a Bayesian approach for inference. All models were fit in `INLA`, first using the default priors discussed in Section 2.3. As a sensitivity check, we refit the models using alternative priors, such as the proper CAR, the Besag, York and Mollie (BYM) prior [32], and ICAR priors with $\text{Ga}(1, 1)$ and $\text{Ga}(1, 0.5)$ precisions. In each case, we obtained results nearly identical to our default ICAR prior. Additionally, we computed bootstrap standard errors for both the non-spatial and spatial DR estimators by resampling with replacement from the original dataset to create 100 new datasets of size 64 022. These samples provided an estimate of the sampling distribution of the DR estimators. The bootstrap standard errors were then formed by computing the standard deviation for each estimator

Figure 3.2: Balance of spatial distribution between NHB and NHW veterans in unweighted (top row), non-spatially weighted (middle row) and spatially weighted (bottom row) samples

.

across the samples. For both the non-spatial and spatial DR estimators, we found that the bootstrap standard errors were nearly identical to those for the large-sample approximation given in equation (3.9). We therefore report the large-sample standard

Table 3.5: Estimated risk differences and 95% credible intervals (CrI) in percent uncontrolled HbA1c under various models.

Model	**Risk Difference**	**95% CrI**	**% Reduction**
Unadjusted	0.076	(0.068, 0.083)	n/a
Non-spatial DR	0.020	(0.011, 0.029)	74
Spatial DR	0.016	(0.005, 0.027)	20

errors in Table 3.5.

Table 3.5 presents the three estimates of the average risk difference between NHBs and NHWs. In our sample, 40.8% of NHB veterans experienced poor HbA1c control compared to 33.2% for NHWs, for an observed sample risk difference of 0.076. When individual-level factors in Table 3.4 were included, the resulting marginal risk difference decreased from 0.076 to 0.020 (95% interval: [0.011, 0.029]), for a 74% decrease. After including a spatial random effect in both stages of the PSA, we observed a further 20% decrease in the risk difference for a final estimate of $\hat{\Delta}^{DR} = 0.016$ (95% interval: [0.005, 0.027]). Thus, failing to incorporate spatial variation would have overestimated the true risk difference in HbA1c control. These results are consistent with previous studies that have found modest reductions in race disparities after accounting for geographic factors [61].

Once a global estimate of the risk difference was established, the second goal of our analysis was to examine spatial variation in the risk of uncontrolled A1c after accounting for potential confounders including race. This secondary aim shifts our focus from estimating a global disparity to identifying hotspots of elevated risk of poor glycemic control after controlling for important patient-level covariates. While we strongly recommend using the spatial DR estimator to address geographic con-

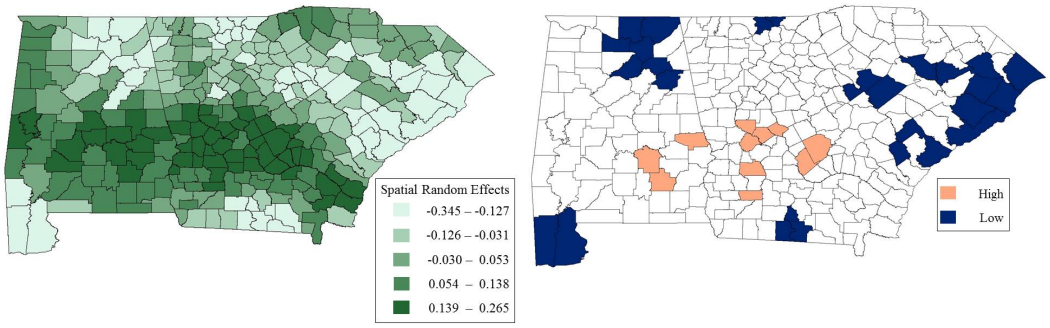

Figure 3.3: Spatial random effects by county and corresponding significance assessed via 95% credible interval (e.g., "High" if interval entirely positive, "Low" if entirely negative)

founding in estimating the overall race disparity, spatial random effect predictions from a well-constructed outcome model alone can lend investigators valuable information in allocating resources and targeting communities. Figure 3.3 displays the fitted spatial random effects and indicates significant spatial effects, assessed in terms of the 95% credible intervals of the random effect estimates. If the interval was entirely positive, the county was designated as "high significant" and if the interval was entirely negative, the county was designated as "low significant". The results indicate a cluster of counties with high effects stretching from central Georgia to Alabama, an area historically encompassed by the "stroke belt" [62]. In contrast, many counties in South Carolina were identified as "low significant", indicating adequate glycemic control. Interestingly, some counties showed significant effects only after covariate adjustment, e.g., in the southwestern corner of Alabama. As these counties are designated "low significant", they demonstrate that once patient-level factors are accounted for, they have significantly improved HbA1c control compared

to surrounding areas. This suggests that the unadjusted differences observed in Figure 3.1 can be explained in part by the demographic make-up these counties. These findings point to the need for community-based and locally-tailored interventions in areas of highest need, particularly along the stroke belt.

3.5 Discussion

We have proposed a spatial DR estimator to estimate a minimally biased average controlled risk difference among race/ethnicity groups in health disparity studies. The spatial DR estimator is an augmentation of the well-established DR estimator and extends recent work in multilevel DR estimation to the spatial setting. To construct the estimator, we introduced spatial random effects into the propensity score and outcome models to account for spatial variation due to potential unmeasured geographic confounders. The spatial effects were assigned CAR priors that promote local spatial smoothing to improve small-area estimation. For statistical inference, we considered both Bayesian and frequentist estimation methods that can be implemented in freely available software such as R or SAS. In the case of Bayesian estimation, we separated the propensity score and outcome models to avoid feedback [55] between the models. We instead used the predictions from the separate models to construct an appropriate DR estimator, which was in turn used to estimate the global ACD.

Through a series of simulation studies, we explored the performance of the spatial DR estimator under varying degrees of spatial heterogeneity and sample size.

When the true generating model incorporated geographic confounding, the spatial DR estimator consistently demonstrated lower bias, lower RMSE, and more reliable coverage than its non-spatial counterpart. Conversely, when the true generating model excluded geographic confounding, the spatial DR estimator performed on par with its correctly specified non-spatial counterpart. In our sub-study, we introduced county-level covariates that were subsequently omitted during model fitting. The results demonstrated that the spatial DR estimator provided unbiased estimates and retained near optimal coverage in the absence of the covariates. This suggests that by incorporating spatial random effects into the estimation process, the spatial DR estimator can alleviate omitted-variable bias at the cluster level. Together, these results point to the benefit of spatial DR estimator in correcting for geographic confounding in health disparities studies.

Our application explored the impact of geographic confounding in racial disparities among a sample of diabetic veterans residing in the southeastern United States. After demonstrating improvement in balance in the propensity score weighted sample, we constructed three estimates of the racial disparity in uncontrolled HbA1c: an unadjusted estimate, a DR-based estimate that adjusted only for individual-level factors, and a spatial DR estimate that adjusted for county-level effects. Our results suggest that adjustment for geographic confounding bias is essential to obtaining an accurate estimate of the global risk difference across large spatial regions. In particular, we found a 20% reduction in the health disparity after adjusting for spatial effects. This reduction is consistent with other studies that incorporate geographic information in racial disparities work [61] and may point to differences in access to

care at the community level. The secondary aim of this study identified areas of poor glycemic control in central Alabama and Georgia and relatively good control in coastal South Carolina after controlling for patient characteristics. As a whole, this information can help community stakeholders direct attention, resources, and policy efforts in a cost-effective manner to ameliorate diabetes-related disparities.

Throughout the paper, we have used the term "geographic confounding" to describe cluster-level spatial heterogeneity that is associated with both race designation and health outcomes. We have deliberately adopted this nomenclature to avoid confusion with the more commonly used term "spatial confounding," which in the spatial literature is used to describe a type of collinearity that arises between Gaussian process random effects and spatially patterned cluster-level covariates, $\boldsymbol{X}$. As Hodges and Reich [63] demonstrate, spatial collinearity can lead to biased estimates of the fixed effect parameters when the spatial effects and fixed covariates compete for overlapping information. To address this issue, they propose a restricted spatial regression that constrains the spatial effects to the orthogonal complement of $\boldsymbol{X}$. We have taken a fundamentally different approach by separating the estimation and modeling stages of spatial PSA. By adopting a two-stage PSA approach, we shift the focus from *estimation* of regression coefficients to *prediction* of potential outcomes, i.e., $\hat{Y}_{ij1}$ and $\hat{Y}_{ij0}$ in equation (3.8). We then use the DR estimator for controlled descriptive comparisons. Thus, the spatial random effects serve only to improve the propensity score and outcome predictions that feed into the DR estimator, rather than to remove bias in the race effect estimate, $\hat{\alpha}$, in outcome model (3.11). As such, we are less concerned with correctly partitioning the spatial effect into fixed and

random components than with accurately predicting propensity scores and potential outcomes using all available spatial information. This goal is supported by previous literature suggesting that collinearity itself is not a primary concern in PSA as long as the predicted propensity scores yield balanced group comparisons [13].

On a more practical note, many authors define "health disparity" as a social construct encompassing historic, geographic and system-level injustices that engender health differences between race groups [64]. Viewed in this way, it may be inappropriate to control for geographic confounding when estimating health disparities, as this would remove part of the disparity effect. Our aim has not been to re-define what constitutes a disparity, but rather to obtain a fully adjusted estimate of the risk difference in glycemic control across race groups. In other words, we wish to make comparisons between racial groups that reside in similar geographic areas. By comparing unadjusted, partially adjusted, and fully adjusted risk differences, as we did in Table 3.5, investigators can disentangle the factors that contribute to racial disparities, a goal of recent disparity studies [65].

Future work might accommodate multiple exposure categories, taking advantage of recent methods for causal inference among multiple treatment groups [66]. The proposed method could also be adapted to handle propensity score matching or stratification. More broadly, the approach could be embedded within a larger spatial causal inference framework, to investigate spatially varying treatment effects, i.e., a "space-by-race" interaction, spatially varying selection bias, or spatial mediation effects. Finally, the work presented here could be applied to other population health settings, such as studies involving telehealth or spatially varying environmental ex-

posures.

3.6 Acknowledgments

This study was supported by grant number CIN 13-418 funded by the VHA Health Services Research and Development (HSR&D) program (PI: Leonard Egede). The funding agency did not participate in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. The manuscript represents the views of the authors and not those of the VA or HSR&D. This study was also funded in part by grants from the National Center for Advancing Translational Science (award number UL1 TR000062), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (award number P60 AR062755), and the National Institute of General Medical Sciences (award number U54-GM104941). We reported a portion of the descriptive summaries in Table 3.4 and Figure 3.1 as part of previous work [59].

Chapter 4

Aim 2

Title: *Propensity Score Matching for Multilevel Spatial Data: Accounting for Geographic Confounding in Health Disparity Studies*

Authors: Melanie L. Davis, Brian Neelon, Paul J. Nietert, Lane F. Burgette, Kelly J. Hunt, Andrew B. Lawson, Leonard Egede

Status: Submitted to Journal of the Royal Statistical Society - Series A in December 2017 (Under Review)

Abstract: We introduce a spatial propensity score matching method to account for "geographic confounding", which occurs when the confounding factors, whether observed or unobserved, vary by geographic region. We augment the propensity score and outcome models with spatial random effects, which are assigned conditionally autoregressive priors to improve inferences by borrowing information across neighboring geographic regions. Through a series of simulations, we show that ignoring spatial heterogeneity results in increased absolute bias and mean squared error, whereas incorporating spatial random effects improves inferences whether the treatment effect is estimated with or without further regression adjustment in the model for the outcome. We apply this approach to a study exploring racial disparities in diabetes specialty care between non-Hispanic black and non-Hispanic white veterans. We construct multiple global estimates of the risk difference in diabetes care: a crude unadjusted estimate, an estimate based solely on patient-level matching, and an estimate that incorporates both patient and spatial information. The crude unadjusted estimate suggests that specialty care is more prevalent among non-Hispanic blacks, while patient-level matching indicates that it is less prevalent. Hierarchical spatial matching supports the latter conclusion, with a further increase in the magnitude of the disparity. These results highlight the importance of accounting for spatial heterogeneity in propensity score analysis, and suggest the need for clinical care and management strategies that are culturally sensitive and racially inclusive.

4.1 Introduction

Type 2 diabetes is the seventh leading cause of death in the United States (CDC, 2014) and disproportionately affects US military veterans [67]. Not only is diabetes more prevalent among veterans [68], but veterans also experience higher comorbidity rates and increased risk of complications than the non-veteran population [69, 70]. The Department of Veterans Affairs (VA) has recently taken steps to address access to care through improved specialty care and emerging telehealth technologies [71]. Nevertheless, veterans continue to face a number of barriers to disease management, including wait times, geographic isolation from care facilities, and insufficient information regarding available health resources [72]. Thus, there is an ongoing need for improved disease management efforts within the VA to help veterans manage their diabetes through healthy diets, regular exercise, and proper medication adherence [73].

At the same time, evidence shows that racial minorities have a higher prevalence of diabetes [35], poorer diabetes outcomes [36], and higher mortality rates compared to non-Hispanic whites [37]. These disparities are explained in part by individual demographics, such as age, sex and marital status [38, 39]. However, patient demographics may explain only one piece of the puzzle. Recent work examining diabetes care found that after accounting for both patient characteristics and facility-level factors, the disparity between non-Hispanic white and non-Hispanic black veterans in LDL cholesterol testing actually increased, with non-Hispanic blacks having lower rates of appropriate LDL management [74]. Studies have also shown that care providers may experience "clinical inertia", whereby a provider fails to respond to a

patient's need for intensified treatment [75]. Indeed, a recent VA study demonstrated widespread clinical inertia in the treatment of veterans with diabetes [76]. Just as personal barriers to disease management may disproportionately affect racial minorities [36, 77], clinical inertia is also thought to be exacerbated for racial minorities whose care providers may have misleading perceptions regarding racial and ethnic minorities' attitudes toward treatment [78]. As a result, ongoing studies are needed to accurately quantify the extent of racial disparities in diabetes care, and to identify strategies for improved disease management.

Because racial disparity studies are inherently observational, it is necessary to account for multiple sources of confounding, both at individual and community levels, in order to obtain minimally biased estimates of race disparities. In particular, it is necessary to account not only for individual-level confounding, but also geographic confounding, which occurs when confounding factors, whether observed or unobserved, vary by geographic location. Here, we use the term "confounding" somewhat broadly to denote a general distortion of the true relationship between race and diabetes-related health outcomes [79]. Depending on the problem at hand, geographic location may act as a common cause of exposure and outcome – and hence as a true confounder – or as a mediator lying on the causal pathway between exposure and outcome. From a statistical standpoint, the two can be handled similarly, as long as the goal is to estimate the adjusted or "direct" effect of exposure on outcome. This is frequently the case in health disparities studies, as policymakers often wish to quantify the direct relationship between race and health outcomes. In the special case of geographic confounding, the goal is to appropriately account for spatial

variation when estimating the extent of racial disparities.

In this paper, we seek to understand how racial minorities engage with the health care system compared to a group of individuals who differ from these patients only in racial identity. Propensity score analysis (PSA) offers a principled approach to addressing this problem. Specifically, PSA enables estimation of the average treatment effect among the treated (ATT), yielding a minimally confounded response to the question "What would the experience of a racial minority have been if the individual were not in this racial group?". The ATT is of particular interest in racial disparity studies, since interventions arising from these studies are typically designed to improve care for specific race groups rather than the population as a whole. Propensity score matching and weighting are two common approaches to PSA, and both can provide unbiased estimates of the ATT. However, propensity score weighting can often result in unstable weights and large variances under extreme propensity score estimates [13]. Matching, on the other hand, offers an intuitive approach to forming a control group that is similar to the treatment group across all factors included in the propensity score model. In fact, matching has been found to perform as well as if not marginally superior to weighting in achieving covariate balance [80].

While there is some previous work on propensity score matching in the context of multilevel data [25] and aggregate (region-level) spatial data [26, 47, 27, 48], there does not currently exist an integrative approach that allows spatial information to augment patient-level information through a hierarchical data structure. We therefore propose a hierarchical spatial propensity score matching framework to address geographic confounding when the ATT is the desired target of inference. While re-

cent work suggests benefits to within-cluster matching [81], this recommendation is not easily extended to the spatial setting. Spatial clusters such as counties may have very small sample sizes and may not function independently of one another in terms of policy and resources.

To address these limitations, we augment traditional propensity score analysis with spatial random effects to account for variation due to unobserved geographic confounders. The random effects are assigned conditionally autoregressive prior distributions that promote localized spatial smoothing by borrowing information from surrounding geographic areas. This information sharing is critical to improving small area estimation. It also reflects our intuition that neighboring areas share resources and should therefore behave similarly with respect to diabetes-related health outcomes. We explore the performance of this method in simulation studies under varying degrees of spatial heterogeneity and sample size. We also conduct simulations to assess whether the outcome variable should be modeled via unadjusted or adjusted regression, helping to shed light on the current debate on this topic. We apply our methods to an analysis of diabetes care and education visits within the Veterans Health Administration. Because the VA is the largest integrated health care system in the United States, its health care decisions and policies are far-reaching; moreover, VA patients represent a "sentinel population" in health care, signaling needs of the more general public population [34]. We show that addressing geographic confounding yields improved effect estimates of racial disparities, which can in turn help guide policy decisions by motivating clinical care teams to engage patients, monitor diabetes management, and design racially and culturally sensitive strategies to alleviate

disparities within the VA and beyond.

4.2 Spatial Propensity Score Analysis

4.2.1 Overview of Propensity Score Matching Methods

We begin by briefly reviewing the inferential properties of PSA as outlined in Rosenbaum and Rubin [8] and summarized more recently in Lunceford and Davidian [14]. Let Z denote a group indicator taking values 0 or 1. In theory, Z can represent an assigned treatment group (e.g., $Z = 1$ if treated and 0 if control) or a manipulable exposure group. As we present our work in the context of racial disparities, we acknowledge that race is not a manipulable exposure; however, health care system engagement and treatment of individuals of different racial groups is manipulable and should be the target of intervention should a disparity exist. For further discussion regarding perceptions of immutable characteristics within the causal framework, see Greiner and Rubin [82] and Davis et al. [83].

According to the causal framework outlined by Rubin (1974), each individual is assumed to have two potential outcomes (Y_1, Y_0), where Y_1 and Y_0 denote the (potentially counterfactual) outcomes under $Z = 1$ and $Z = 0$, respectively. The observed response, Y, is given by $Y = ZY_1 + (1 - Z)Y_0$, so that $Y = Y_1$ if $Z = 1$ and $Y = Y_0$ otherwise. The causal estimand of interest depends on the clinical question at hand. Common choices are the population average treatment effect (ATE), defined as $\Delta_{ATE} = E(Y_1) - E(Y_0)$, or the population average treatment effect on the treated (ATT), defined as $\Delta_{ATT} = E(Y_1 - Y_0|Z = 1)$. The former

provides a causal comparison between the treated and the entire control population, while the latter provides a causal comparison restricted to the treated population. The ATT is often desired in program evaluation or when the treatment is not likely to be targeted universally, as is the case in our motivating study.

Under unconfoundedness, propensity score methods can be used to derive unbiased estimators of the ATE or ATT in observational studies. The propensity score, $e(\boldsymbol{x}) = \Pr(Z = 1|\boldsymbol{X} = \boldsymbol{x})$, is the conditional probability of exposure given covariates $\boldsymbol{X}$, where the so-called "overlap" condition, $0 < e(\boldsymbol{x}) < 1$, is assumed to hold. Propensity score matching is a technique that forms matched pairs between exposed and unexposed subjects based on the similarity of their estimated propensity scores [8, 18, 2]. As is true across all propensity score methods, matching techniques require the analyst to first decide on the form of the propensity score model (typically a logistic regression model) and the variables to be included in the model. After propensity scores have been generated, the analyst must first make decisions on the matching strategy: greedy or optimal algorithms, matching with or without replacement, the matching variable itself (e.g., propensity score or the logit of propensity score), and the rules for designating acceptable matches. Because the focus of this work is to address geographic confounding through the use of spatial random effects, our analysis strives to incorporate well evidenced propensity score methods that lend themselves to otherwise straightforward inference.

Greedy algorithms create nearest-neighbor best pair matches by iteratively choosing an individual in the treatment group, finding the control with the most similar propensity score and removing that pair from the selection process. Thus, greedy

matching does not revisit matches once they are formed. Recent work has shown that greedy algorithms perform similarly to other matching procedures in their ability to form well-balanced groups [13]. Matching with replacement allows a control unit to be used in more than one pair match, whereas without replacement restricts a control to participation in only one matched pair. Matching with replacement can yield a suitable matched sample; however, a matched sample based on very few influential control units can lead to inflated variance estimates [7]. Therefore, some researchers recommend matching without replacement, which has been found to perform as well as matching with replacement but avoids analytic complexity and the variance pitfall [20]. In terms of acceptable match designation, Austin [2] recommends a caliper width equal to 0.2 times the standard deviation of the logit of the propensity score as a valuable compromise between preserving match quality and minimizing mean square error (MSE) of the treatment effect. Given the above recommendations, we adopt a nearest neighbor algorithm that matches individuals without replacement based on the logit of the propensity score and a caliper of 0.2 times the standard deviation. These choices yield a sample of treated individuals and a well-matched control group that is a subset of the entire control population, naturally allowing for estimation of the ATT. Finally, some authors recommend fitting an adjusted regression model to the outcome to address any residual imbalance between exposure groups [13], while others advocate for an unadjusted model [2]. Given this ongoing debate [84], we consider both approaches in our simulations studies to determine the preferred method in the context of spatial PSA.

4.2.2 Multi-level Spatial Matching

PSA has been recently extended to the hierarchical data setting in which individuals are nested within clusters such as health care plan [9]. Arpino and Mealli [25] in particular have proposed propensity score matching methods for hierarchical data that incorporate random effects into the propensity score model when within-cluster matching is not feasible. They demonstrate that random effects are capable of capturing unmeasured heterogeneity that occurs when cluster-level confounders are omitted in PSA.

The multilevel matching estimator proposed by Arpino and Mealli [25] is readily extended to the spatial setting by augmenting the propensity score model with spatial random effects. Turning to our motivating application, let Y_{ij} be an indicator variable taking the value 1 if the j-th patient residing in the i-th county receives a specialty care visit, let Z_{ij} be an indicator taking the value 1 if the patient is non-Hispanic black (NHB) and 0 if non-Hispanic white (NHW), and let $\boldsymbol{x}_{ij}$ represent a set of observed patient- and county-level covariates. The spatial propensity score model is given by

$$\text{logit}(e_{ij}) = \text{logit}[\Pr(Z_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{1i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \phi_{1i}, \tag{4.1}$$

where ϕ_{1i} is the spatial random effect for county i. The spatial effect ϕ_{1i} accounts for unmeasured county-level factors associated with race, and circumvents the need to match within county, which is infeasible in the case of small cluster sizes.

Once the propensity scores are estimated, we match each NHB patient to a corresponding NHW patient to form a matched sample. The R package `Matching` [85]

allows for direct input of the desired matching variable, is flexible enough to accommodate various strategies, and has been used in multilevel matching [81]. After matching, we can estimate the ATT by performing unadjusted regression analysis. Alternatively, Stuart [13] recommends fitting an adjusted outcome model that can address residual imbalance across the groups with respect to important covariates and space. To fit the adjusted outcome model, we again incorporate a spatial random effect into our binary outcome model

$$\text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{2i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + z_{ij}\alpha + \phi_{2i}, \tag{4.2}$$

where ϕ_{2i} denotes the spatial random effect for county i in the outcome model. The spatial random effects can represent geographic variability in health care access, availability of community outreach and medical education programs, or access to other resources associated with diabetes management. To investigate the impact of adjusted regression on the outcome in spatial settings, we pursue both unadjusted and adjusted estimates as part of our simulation study.

To encourage maximal spatial smoothing, we assign the random effects ϕ_{1i} and ϕ_{2i} independent intrinsic conditional autoregressive (ICAR) priors [32]. Let $k = 1$ denote the propensity score model and $k = 2$ denote the outcome model. The ICAR prior for ϕ_{ki} takes the conditional form

$$\phi_{ki} \mid \boldsymbol{\phi}_{k(-i)}, \sigma_k^2 \sim \text{N}\left(\frac{1}{m_i}\sum_{h \sim i} \phi_{kh}, \sigma_k^2/m_i\right), \quad k = 1, 2, \tag{4.3}$$

where $h \sim i$ indicates that county h is a geographic neighbor of county i, m_i is

the number of neighbors, and, for model k, σ_k^2 is the conditional variance of ϕ_{ki} given the remaining spatial effects, $\boldsymbol{\phi}_{k(-i)}$. Following Brook's Lemma [33], the joint distribution for $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kn})^T$ is given by

$$\pi(\boldsymbol{\phi}_k \mid \sigma_k^2) \propto \exp\left(-\frac{1}{2\sigma_k^2}\boldsymbol{\phi}_k^T \boldsymbol{Q} \boldsymbol{\phi}_k\right), \quad k = 1, 2, \tag{4.4}$$

where $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{A}$ is a spatial structure matrix of rank $n-1$, with $\boldsymbol{M} = \text{diag}(m_1, \ldots, m_n)$ and $\boldsymbol{A}$ representing an $n \times n$ adjacency matrix with $a_{ii} = 0$, $a_{ih} = 1$ if $i \sim h$, and $a_{ih} = 0$ otherwise. When a fixed intercept is included in the model, a sum-to-zero constraint must be applied to $\boldsymbol{\phi}_k$ to ensure an identifiable model.

The ICAR prior is appealing because it imposes spatial smoothing, reflecting the intuition that adjacent spatial units are more similar in terms of access to health care, resources and policies than non-neighbors. Moreover, by promoting localized spatial smoothing and information sharing from surrounding geographic areas, the ICAR prior reduces uncertainty in estimating the propensity scores and, in turn, the ATT.

4.2.3 Model Fitting and Inference

For our case study, we adopt a Bayesian model fitting approach and assign prior distributions to all model parameters. As a default, we assign weakly informative $\text{N}(0, 1\text{e}5)$ priors to fixed effects and $\text{Ga}(1, 5\text{e-}05)$ priors for the spatial precision terms, where $\text{Ga}(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b. We fit the propensity score and outcome models separately, thus avoiding

the so-called "feedback" issue that can arise when the models are fit jointly under a fully Bayesian approach [31]. We use approximate Bayesian methods for posterior inference. Specifically, we adopt the efficient integrated nested Laplace approximation (INLA) proposed by Rue et al. [56]. INLA uses a Laplace approximation to estimate the joint posterior of the model parameters, yielding improved computational capabilities over standard Markov chain Monte Carlo routines. This method can be readily implemented in the R package `INLA` (www.r-inla.org), where the `Besag` option is used to specify the ICAR prior. The posterior means of the propensity scores are then used to match individuals.

In our application, we match individuals without replacement using the logit of the estimated propensity score with a caliper of 0.2 times the standard deviation as recommended by Austin [2]. We consider both unadjusted and adjusted outcome models when estimating the ATT. For both the unadjusted and the adjusted outcome models, we compute a "standardized" risk difference first by assuming each patient is NHB and, second, by assuming each patient is NHW. The difference provides an estimate of the ATT. In order to construct a credible interval (CrI) around this estimate, we used the `inla.posterior.sample` function within R-INLA to obtain 1000 Monte Carlo draws from the approximate posterior distribution. The mean of the risk difference across the 1000 samples is reported as the estimated ATT, and the corresponding the 95% CrI is derived from the 2.5 and 97.5 percentiles.

4.3 Simulation Study

4.3.1 Data Description

In order to assess the properties of hierarchical spatial matching, we conducted two simulation studies. The goal of the first study was to quantify the impact of ignoring true spatial heterogeneity when estimating propensity scores, to measure the performance of the proposed spatial PSA methodology under various reasonable true-data scenarios, and to suggest options within readily available software to achieve a minimally biased estimate of the ATT. The goal of the second study was to assess how well the proposed method accounts for omitted variable bias.

To mirror the spatial structure of our application, for both studies we generated patient-level data clustered at the county level across the southeastern United States. To emulate the geographic structure in our application, we used the US Census county-level adjacency matrix for South Carolina, Georgia, and Alabama [42]. This matrix contains $n = 272$ counties and 1528 pairwise adjacencies. For the first study, we generated 100 datasets with treatment assignment and outcome according to the following propensity score and outcome models:

$$\text{logit}[\Pr(Z_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{1i})] = \beta_0 + \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \phi_{1i} \quad (4.5)$$

$$\text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{2i})] = \gamma_0 + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma} + z_{ij}\alpha + \phi_{2i}, \quad (4.6)$$

where $i = 1, \ldots, 272,\ \ j = 1, \ldots, n_i$ and $\boldsymbol{x}_{ij}$ is a 5×1 vector comprising patient-level covariates generated from the following distributions: N(5,2), N(0,1), Bernoulli(0.4),

Bernoulli(0.2), Bernoulli(0.05). The fixed effect coefficients were set at $\beta_0 = 0.25$ and $\boldsymbol{\beta} = \{-0.15, -0.2, 0.5, 0.6, -0.3\}$, $\gamma_0 = 0.25$, $\boldsymbol{\gamma} = \{-0.75, 0.1, .5, .15, -0.40\}$, and $\alpha = 0.60$; n_i was allowed to vary ($n_i = 25$, 50) to reflect sample sizes similar to those in the application; ϕ_{1i} and ϕ_{2i} were simulated from ICAR models given in equation (4.4) with $\sigma^2 = \{0, 3, 6\}$ representing varying degrees of spatial variation. The case where $\sigma^2 = 0$ corresponded to the scenario in which geographic confounding was not present.

For the second simulation, we evaluated the performance of spatial PSA when relevant county-level predictors were omitted from the analysis. We augmented the models in equations (4.5) and (4.6) with two county-level covariates, the first generated from a N(10,3) distribution and the second simulated according to the ICAR model given in equation (4.4) with $\sigma^2 = 2$. The regression coefficients for the two covariates were set to -0.10 and 0.1 for the propensity score model and to 0.3 and -0.3 for the outcome model. The spatial variances for ϕ_{1i} and ϕ_{2i} were set to 3 to mimic the estimates in our application. We then fit the propensity score and outcome models ignoring these covariates in order to assess the impact of the omitted variables on the ATT estimate.

4.3.2 Results

Table 4.1 summarizes the results of the first simulation study. This table presents measures of performance of spatial propensity score matching under varying degrees of spatial variation and sample sizes. Rows indicate the sample size and spatial variance values; columns indicate whether the estimate was derived from an unadjusted

Table 4.1: Results for simulation study 1: Mean bias, RMSE, and 95% coverage of the risk difference under various sample sizes and spatial variances (rows) and estimation methods (columns). $\sigma^2_\phi = 0$ represents no spatial variation

	Unadjusted						Adjusted					
	Non-spatial			Spatial			Non-spatial			Spatial		
$\sigma^2_\phi = 0$												
$n_i = 25$	0.007	0.008	96	0.007	0.009	88	0.007	0.008	94	0.007	0.009	90
$n_i = 50$	0.005	0.007	88	0.006	0.008	87	0.006	0.007	87	0.007	0.008	85
$\sigma^2_\phi = 3$												
$n_i = 25$	0.014	0.018	65	0.011	0.013	89	0.014	0.018	62	0.009	0.011	90
$n_i = 50$	0.013	0.017	53	0.008	0.009	87	0.014	0.018	49	0.006	0.007	92
$\sigma^2_\phi = 6$												
$n_i = 25$	0.021	0.027	51	0.011	0.014	92	0.022	0.028	46	0.009	0.011	92
$n_i = 50$	0.021	0.029	42	0.010	0.013	77	0.021	0.029	37	0.006	0.008	92

model that included an indicator for race only or an adjusted model that included patient covariates with and without an additional spatial random effect. Within each strategy, columns further indicate whether spatial random effects were incorporated in the analysis. Explicitly, "Unadjusted, Non-Spatial" implies that the propensity score model included only individual-level covariates, while the outcome model included only a indicator for race; "Unadjusted, Spatial" implies that the propensity score model included both individual-level covariates and a spatial random effect, while the outcome model included only race; "Adjusted, Non-spatial" implies that the propensity score and outcome models ignored space but included individual-level covariates; and "Adjusted, Spatial" implies that with fit a fully adjusted spatial model for both the propensity score and outcome models.

Several trends are apparent in Table 4.1. First, ignoring geographic confounding and utilizing only patient-level measures is detrimental. We observe poor perfor-

mance of the non-spatial analyses as the bias and RMSE are increased while the coverage is decreased. Secondly, regression adjustment appears to yield smaller bias and RMSE than unadjusted analysis and typically better coverage when spatial analyses are performed. For example, when $\sigma^2_\phi = 3$ and $n_i = 50$ and space is ignored, coverage is 53% and 49% in the unadjusted and adjusted analyses, respectively. However, when spatial PSA is conducted for the same data, we observe near-nominal coverage, with the additional adjustment in the outcome model yielding a slightly better result than the unadjusted outcome analysis (92% versus 90%). Lastly, in the case of no true spatial heterogeneity, conducting spatial analysis does not appear to be highly detrimental, as it contributes no additional bias. For instance, when $\sigma^2_\phi = 0$ and $n_i = 50$, non-spatial and spatial analyses yielded nearly identical coverage probabilities. We observe similar trends in measures of bias and RMSE. These results suggest that incorporating spatial random effects into the propensity score model and the adjusted outcome model yields favorable results when spatial variation is present and does not yield negative consequences when the data exhibit no spatial heterogeneity.

Table 4.2 presents results of the second study using INLA to estimate the propensity score and outcome models. The goal of the second study was to assess the ability of the proposed spatial matching framework to capture the true ATT when relevant county-level covariates were left out of the analysis. The non-spatial analyses ignored space entirely and fit either an unadjusted model for the outcome or a non-spatial, covariate adjusted outcome model. The spatial analyses incorporated a spatial random effect in the propensity score model and fit either an unadjusted model for the outcome or a spatial and patient-level adjusted model for the outcome. As Table

Table 4.2: Results for simulation study 2: Mean bias, RMSE, and 95% coverage of the risk difference under various sample sizes (rows) and estimation methods (columns) with a fixed spatial variation ($\sigma^2_\phi = 3$)

	Unadjusted						Adjusted					
	Non-spatial			Spatial			Non-spatial			Spatial		
$n_i = 25$	0.018	0.027	76	0.012	0.016	94	0.018	0.027	71	0.010	0.013	94
$n_i = 50$	0.019	0.025	77	0.009	0.011	94	0.015	0.021	67	0.008	0.010	94

4.2 indicates, the non-spatial analysis performed poorly, whereas the spatial analyses retained favorable properties across all scenarios, including low bias, low RMSE and near-nominal coverage. These results support the use of the proposed spatial matching framework, as it appears to capture the true risk difference even when county-level fixed effects are ignored during model fitting. As it is not uncommon for these covariates to be unavailable to the analyst, spatial matching provides a practical strategy to account for unmeasured geographic confounding. Together, the results of the simulation studies demonstrate that spatial propensity score matching through the inclusion of spatial random effects addresses geographic confounding and outperforms analyses that include only patient-level covariates.

4.4 Analysis of Racial Disparities in Diabetes Care and Management

We conducted an analysis to examine the direct association between race and the likelihood of a diabetes care visit in 2014. Our sample consisted of 20,636 NHB ($n = 9,277$) and NHW ($n = 11,359$) veterans with uncontrolled type 2 diabetes

living in Georgia, Alabama and South Carolina. Uncontrolled type 2 diabetes was defined as HbA1c $\geq$ 8 at the start of 2014. Table 4.3 displays the patient-level variables that were included in the propensity score and adjusted outcome models. Approximately 13% of the patients had a diabetes care visit following indication of poor control (15.0% for NHBs, 11.2% for NHWs).

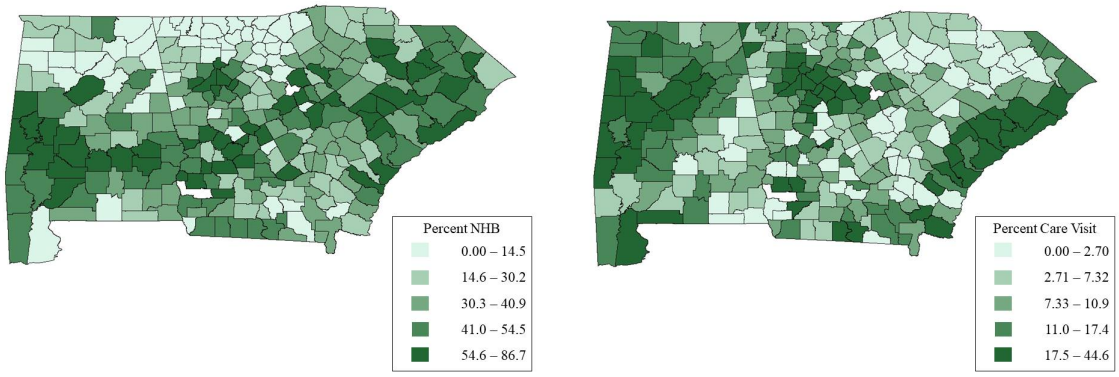

Figure 4.1: Unadjusted percent of veterans with uncontrolled diabetes who are NHB (left) and unadjusted percent of veterans with uncontrolled diabetes who received a diabetes care education visit (right)

Figure 4.1 displays the per-county percents of NHB veterans and veterans with diabetes care visits. The maps suggest that the percent of NHB veterans and the percent of veterans with diabetes care visits exhibit spatial variation, with clustering around areas in western Alabama, Atlanta, Georgia and coastal South Carolina.

In order to assess the covariate balance between NHB and NHW veterans in the original and spatial propensity score matched samples, we estimated the difference in means or proportions in each of the samples. To construct the spatially matched sample, we fit a logistic propensity score model that included the patient-level co-

Table 4.3: Balance of covariates between NHB and NHW veterans in pre-matched and post-matched samples; "Stand. Diff." denotes the absolute value of the standardized difference

	Pre-match			Post-match		
Variable	NHB	NHW	Stand. Diff.	NHB	NHW	Stand. Diff.
Age	62.00	67.49	0.573	64.52	64.49	0.003
Female	7.43	2.67	0.219	7.38	6.83	0.021
Service Percent ≥ 50	45.97	39.41	0.133	45.95	48.20	0.045
Married	52.83	65.79	0.266	52.88	57.84	0.099
Substance Abuse	11.04	4.46	0.248	10.99	10.51	0.015
Cerebrovascular Disease	3.58	3.24	0.019	3.57	3.69	0.006
Congestive Heart Failure	8.58	11.07	0.084	8.58	8.69	0.004
Cardiovascular Disease	8.46	15.64	0.222	8.47	8.89	0.015
Depression	35.94	31.55	0.093	35.92	36.11	0.004
Hypertension	88.41	84.44	0.116	88.42	86.65	0.054
Obesity	27.23	25.87	0.031	27.23	27.71	0.011
Psychoses	6.83	3.81	0.135	6.82	7.09	0.011
Homeless	0.91	0.18	0.099	0.86	0.99	0.014

variates described in Table 4.3 and a spatial intercept term. Matching was based on the logit of the estimated propensity score and a caliper of 0.2 times the standard deviation of the logit is imposed in order to ensure a well-matched sample. We observed a decrease in the standardized difference across the patient-level covariates in the spatially matched sample. However, some standardized differences were still sizable; for example, the standardized difference for "married" is close to 0.10, which would be considered the threshold for negligible difference. We were therefore motivated to consider regression adjustment in the outcome model to address any residual imbalance.

Figure 4.2 displays the spatial distribution of NHB and NHW veterans in the unmatched and spatially matched samples. The spatial distribution of NHB and

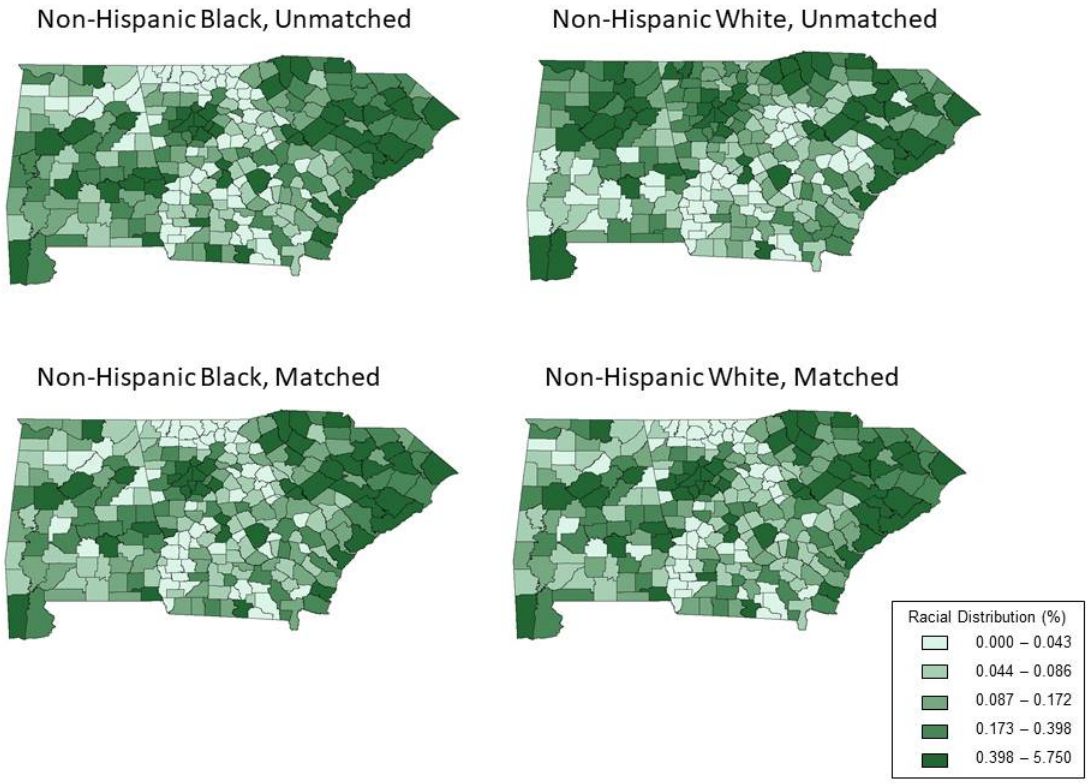

Figure 4.2: Balance of spatial distribution between NHB and NHW veterans in unmatched (top row), and spatially matched (bottom row) samples
.

NHW veterans varied in the unmatched sample, implying that NHBs and NHWs were concentrated in different areas. While a high percent of both NHB and NHW veterans live in urban areas such as Atlanta, NHW veterans alone appear to be concentrated in northern Georgia, where only 0.00% to 0.043% of NHB veterans reside (lightest shade on the map). This spatial imbalance is ameliorated once a spatially matched sample is created. In the spatially matched sample, the distribution of NHW veterans (the "controls") more closely mimics the nearly unchanged distribution of NHB veterans

(the "treated"), indicating that we have selected geographically well-matched controls.

To assess the performance of the proposed spatial PSA, we fit five models. We first examined the observed sample risk difference in diabetes care between NHB and NHW ("unadjusted" analysis). Next, we performed a non-spatial analysis that included patient covariates in the propensity score model, but not in the outcome model ("Patient I" analysis). Third, we replicated the patient covariates in the propensity score model and additionally fit a covariate adjusted logistic regression model for the outcome ("Patient II" analysis). Fourth, we fit a model that included an additional spatial random effect in the propensity score model, while the outcome model was left unadjusted ("Spatial I" analysis). Finally, we conducted a fully adjusted spatial PSA that included patient-level covariates and spatial random effects in both models ("Spatial II" analysis). We used the estimated coefficients from the model to form a standardized estimate of the risk difference. The reported 95% CrI was constructed using the 2.5 and 97.5 percentiles of the sample distribution of risk differences.

Results of this stepwise analysis are presented in Table 4.4. The unadjusted risk difference indicates that NHB veterans with uncontrolled diabetes have a greater probability of receiving diabetes care and education (risk difference = 0.038, 95% CrI = [0.029, 0.047]). This result is somewhat counterintuitive in light of recent studies on care management, which have found that NHBs are less likely to receive intensified treatment [78]. However, NHB veterans included in this analysis were more likely to be obese, female, and have a higher rate of service connected disability (Table 4.3). The imbalance in these factors may explain the positive direction of the disparity. Once we matched on patient level factors ("Patient I"), the risk difference

reversed direction (-0.021, 95% CrI = [-0.031, -0.011]), indicating that NHB veterans with uncontrolled diabetes have a lower probability of receiving specialized care and education. With further covariate adjustment in the outcome model ("Patient II"), the risk difference decreased slightly (-0.027, 95% CrI = [-0.038, -0.016]) but was similar to the estimate from the matched sample unadjusted model. Because the percent of NHB veterans and the percent of veterans receiving care visits by county appear to exhibit spatial variation, it is likely that when geography is ignored, the true disparity is not fully revealed, as NHBs may be more likely to live in areas with high rates of care visits. When spatial random intercepts were included in the analysis and the matched sample was geographically balanced, we observed an increase in the magnitude of the disparity. In the unadjusted spatial analysis ("Spatial I"), the estimated risk difference was -0.057 (95% CrI = [-0.068, -0.046]). In the adjusted spatial analysis ("Spatial II"), the estimated risk difference was -0.071 (95% CrI = [-0.085, -0.057]), suggesting a 7 percentage point difference in the receipt of diabetes care between NHBs and NHWs. While in general agreement with the effect estimate from the unadjusted spatial analysis in the matched sample, the effect estimate from further regression adjustment indicates a more marked racial disparity, providing strong evidence for the incorporation of spatial random effects in both the propensity score and outcome models.

Table 4.4: Estimated risk differences in the racial disparity of diabetic care visits under modeling strategies. Negative values indicate that NHBs have a lower estimated risk of receiving a diabetic care visit. Unadjusted: observed sample risk difference. Patient I: covariate-adjusted propensity model, unadjusted outcome model. Patient II: covariate-adjusted propensity score and outcome models. Spatial I: spatial propensity score model, unadjusted outcome model. Spatial II: spatial propensity score and outcome models.

Model	**Risk Difference**	**95% CrI**
Unadjusted	0.038	(0.029, 0.047)
Patient I	-0.021	(-0.031, -0.011)
Patient II	-0.027	(-0.038, -0.016)
Spatial I	-0.057	(-0.068, -0.046)
Spatial II	-0.071	(-0.085, -0.057)

4.5 Discussion

We have proposed a spatial propensity score matching framework to estimate the ATT among racial groups in studies examining disparities in health management and system engagement. To account for unmeasured geographic confounding, we incorporated spatial random effects into the propensity score model and, in the case of further regression adjustment, into the outcome model as well. These spatial effects can represent geographic confounders such as proximity to health care facility, access to resources, and community support and education. The spatial effects were assigned CAR priors that promote local spatial smoothing and are able to improve estimation in areas with sparse data. We adopted a Bayesian inferential approach, but fit the propensity score and outcome models separately to avoid potential feedback concerns that arise from joint estimation [31]. By implementing Bayesian estimation within R-INLA, we used readily available, free software that can be utilized in a multitude of studies across many health care data platforms.

In simulation, we examined the performance of the proposed spatial propensity score matching framework under varying degrees of spatial variation and sample size. Under true geographic confounding, spatial matching outperformed matching that failed to incorporate spatial information. Spatial matching demonstrated decreased bias and RMSE and improved coverage compared to non-spatial matching. This result was true whether the ATT was estimated by unadjusted regression in the matched sample or further covariate and spatial adjustment was employed. In general, regression adjustment to address residual imbalance led to lower bias and RMSE. When true geographic confounders were ignored in the analysis, and only a spatial random intercept was included in the modeling, spatial matching offered reasonably low bias and RMSE and nearly nominal coverage, suggesting that the proposed method can alleviate bias due to omitted spatially varying confounders. In contrast, the non-spatial analysis performed very poorly. This supports the need to address geographic confounding in studies of racial disparities.

Our application explored the impact of geographic confounding in racial disparities among veterans with uncontrolled diabetes in the southeastern United States. We reported an unadjusted estimate of the ATT, a patient-level matched estimate, a spatially matched estimate, and a spatially matched estimate that further addressed imbalance through an adjusted regression model. The crude unadjusted estimate suggested that NHB veterans with uncontrolled diabetes may have a higher risk of receiving a specialty care visit; however, once patient-level factors were balanced, we saw this association reverse. Furthermore, once we additionally balanced on space, the disparity in diabetes care visits became more pronounced, with NHB veterans

having a lower probability of receiving a specialty care visit to address their uncontrolled diabetes.

These findings have important policy implications for mitigating disparities in diabetes management and for improving patient engagement with the health care system. First, policymakers can target intervention to identify, engage and maintain patients who are in need of intensified treatment. Vulnerable populations who are less likely to seek specialized care may need to be recruited in local, well-trusted community settings [86, 87, 88]. These patients may also benefit from care navigators or patient advocates in a complex care setting [89]. Clinician training can also be tailored to address issues such as "clinical inertia" and the conduct of culturally sensitive consultations [90]. Lastly, disease management media and instruction pamphlets can encourage patients to seek guidance and agency of their clinical care. These policy efforts can help the VA achieve its stated mission to "champion advancement of health equity and reduction of health disparities for disadvantaged Veterans" as outlined in its recent Health Equity Action Plan [91].

Future work could adapt spatial propensity score methodology to stratification or a combination of propensity score methods. Furthermore, the proposed methods could be extended to accommodate time-varying treatments or broader types of outcomes, such as count, survival or multivariate outcomes. Lastly, the work presented here could be applied to numerous other public health applications, such as studies addressing the implementation of telemedicine or spatially varying outreach programs.

Acknowledgements

This study was supported by grants CIN 13-418 (PI: Leonard Egede) and HX002299-01A2 (Co-PIs: Brian Neelon and Kelly J. Hunt) funded by the VHA Health Services Research and Development (HSR&D) program. The funding agency did not participate in the design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript. The manuscript represents the views of the authors and not those of the VA or HSR&D.

Chapter 5

Aim 3

Title: *Analysis of Racial Differences in Hospital Stays in the Presence of Geographic Confounding*

Authors: Melanie L. Davis, Brian Neelon, Paul J. Nietert, Kelly J. Hunt, Andrew B. Lawson, Lane F. Burgette, Leonard E. Egede

Status: In preparation for submission to Spatial and Spatiotemporal Epidemiology

5.1 Introduction

It is estimated that between 12 and 14 percent of Americans have type 2 diabetes, with a heavier burden among non-white racial minorities [92]. Patients with type 2 diabetes experience more common and lengthier inpatient hospital stays and are more likely to die in the hospital than their non-diabetic counterparts [93]. In fact, inpatient stays are the highest medical expenditure among type 2 diabetics, accounting for 43% of the 256 billion dollars spent annually [94]. Promisingly, the number

of days Americans spend in the hospital has been decreasing over time. However, for ethnic and racial minorities, this trend may actually represent short but frequent health care system encounters of poor-quality, insufficient care [95]. It is still unclear how observed decreases in inpatient days differentially affect racial minorities even after disease management initiatives to reduce inpatient encounters have been enacted [96].

There are a number of factors that may contribute to disparities in inpatient hospital stays. Barriers to access of inpatient services may vary at both individual and facility levels. Differences in comorbidity burden or financial obligation in payment for services may account for some of the racial differences that have been reported [97]. Patient-provider relationships and patient advocacy can also influence hospitalizations and the number of days patients spend in the hospital [96]. Furthermore, the use of inpatient services has been shown to vary geographically. For example, recent studies have shown that after socioeconomic status and disease burden is controlled for, areas with higher hospital capacity tend to have higher hospitalization rates [98]. Even in a relatively homogeneous patient population such as veterans receiving care within the Veterans Health Administration (VHA), geographic variation in inpatient utilization persists [99].

In this paper, we are interested in understanding racial differences in the risk of hospitalization and the number of inpatient days among veterans with type 2 diabetes after accounting for potential confounding factors. Propensity score matching offers a principled approach to addressing the issue of confounding and enables estimation of the average treatment effect among the treated (ATT), yielding a minimally con-

founded response to the question "What would the experience of a racial minority have been if the individual were not in this racial group?". The ATT is often of interest in health disparities studies as interventions typically target the at-risk group rather than the population as a whole. Because community factors such as accessibility to health care facilities and availability of disease management resources can exacerbate health disparities, it is critical to account for not only patient-level confounding but also geographic confounding, which occurs when confounding factors vary spatially. In previous work, we have shown that incorporating spatial random effects into the propensity score model can yield a matched sample balanced on the distribution of racial groups across geographical regions. This in turn minimizes the bias in estimating the ATT among the matched sample [100].

Once a matched sample is generated, fitting a model for the outcome can address any residual imbalance and may yield improved effect estimates [100, 13]. The specification of the outcome model is flexible and allows analysts to tailor it to the specific research question at hand. Two-part hurdle models [101] have been utilized in studies examining racial disparities in inpatient stays [96], as they allow researchers to address questions regarding both the risk of hospitalization and the number of inpatient days while accounting for zero-inflation in the count response. Hurdle models are two-part mixture models comprising a binary component that models (in our case) the probability of hospitalization, and a truncated count component that models the number of days in the hospital among those who are hospitalized. The truncated negative binomial distribution is an attractive choice for the count model as it allows for overdispersion relative to the Poisson assumption that the variance is equal to

the mean. The negative binomial hurdle model can easily be extended to the spatial setting by incorporating spatial effects into both the binary and count components of the model. The resulting model can be fit within a Bayesian framework using standard software such as R-INLA [56].

In this work, we combine methods in spatial propensity score matching and hierarchical spatial hurdle models to achieve minimally biased estimates of the racial disparity in the risk of hospitalization and the mean number of inpatient days. We conduct a simulation study to assess the performance of spatial propensity score analysis in combination with the spatial hurdle model under unknown geographic confounders. We apply these methods to an analysis of the effect of race on hospitalization and inpatient days among type 2 diabetic veterans receiving care within the VHA in the southeastern United States in the 2014 fiscal year. As the VHA is concerned with reducing hospitalizations and inpatient days while simultaneously maintaining quality care, it is important to understand differences in health care services between non-Hispanic whites and racial minorities. Furthermore, as patients receiving care within the VHA are often considered a "sentinel" population, the results of this study may indicate areas in need of attention in the general population [34].

5.2 Methods

5.2.1 Spatial Propensity Score Analysis

We begin by first addressing the inferential properties and model specification of the propensity score and extensions that incorporate a spatial random effect to address geographic confounding. Let Z denote a race group indicator taking the value 1 if a patient is non-Hispanic black (NHB) and 0 if non-Hispanic white (NHW). While race itself is not a manipulable characteristic, the experience of different racial groups within the health care system *is* manipulable and should be the target of intervention if an analysis demonstrates a disparity.

In the context of the causal framework outlined by Rubin [6], each individual is assumed to have two potential outcomes $(Y^{(1)}, Y^{(0)})$, where $Y^{(1)}$ and $Y^{(0)}$ denote the (potentially counterfactual) outcomes under $Z = 1$ and $Z = 0$, respectively. The observed response, Y, is given by $Y = ZY^{(1)} + (1-Z)Y^{(0)}$, so that $Y = Y^{(1)}$ if $Z = 1$ and $Y = Y^{(0)}$ otherwise. The population ATT is the average difference between the potential outcomes among "treated" patients, formally defined as $\Delta_{ATT} = E(Y^{(1)} - Y^{(0)}|Z = 1)$. The ATT is often desired in program evaluation or when the treatment is not likely to be targeted universally, as is the case in our motivating study.

Under unconfoundedness, propensity score methods can be used to derive unbiased estimators of the ATT in observational studies. The propensity score, $e(\boldsymbol{x}) = \Pr(Z = 1|\boldsymbol{X} = \boldsymbol{x})$, is the conditional probability of exposure given covariates $\boldsymbol{X}$, where the so-called "overlap" condition, $0 < e(\boldsymbol{x}) < 1$, is assumed to hold. Propensity score matching is a technique that forms matched pairs between exposed and unex-

posed subjects based on the similarity of their estimated propensity scores [2, 8, 18].

Recent work has explored the use of propensity score analysis in the hierarchical setting, where patients are nested within clusters such as health care plans or hospitals [25, 9, 81]. Arpino and Mealli [25] proposed propensity score matching methods for hierarchical data that incorporate random effects into the propensity score model. In the hierarchical spatial setting, we augment the propensity score model with spatial random effects. The spatial effects are then assigned distributions that account for spatial correlation and promote smoothing across spatial units. More concretely, let Z_{ij} denote an indicator variable taking the value 1 if the j^{th} patient in the i^{th} county is non-Hispanic black (NHB) and 0 if non-Hispanic white (NHW), and let $\boldsymbol{x}_{ij}$ represent a set of observed patient-level covariates. The spatial propensity score model is given by

$$\text{logit}(e_{ij}) = \text{logit}[\Pr(Z_{ij} = 1 | \boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}, \phi_{1i})] = \boldsymbol{x}_{ij}^T \boldsymbol{\alpha} + \phi_{1i}, \tag{5.1}$$

where ϕ_{1i} is the spatial random effect for county i. The spatial effect ϕ_{1i} accounts for unmeasured county-level factors associated with race, and circumvents the need to match within county, which may be infeasible in the case of small cluster sizes [81].

By matching on the spatial propensity scores, investigators can achieve balance across patient factors and geographical distribution between NHB and NHW patients. Once the matched sample is constructed, a variety of outcome models specific to the research question at hand can be employed. Adjusted outcome models can be used to address any residual imbalance between exposure groups [13] and have been shown in simulation to be beneficial to unbiased estimation of the ATT [100].

5.2.2 Two Part Spatial Hurdle Models

Our motivating study of inpatient hospitalization practices poses a set of unique analytic challenges. First, approximately 71% of the patients in the sample were not hospitalized in 2014, resulting in substantial zero-inflation. Furthermore, among those who were hospitalized, there was a wide range of counts of inpatient days. In order to capture both zero-inflation for patients who do not experience a hospitalization and overdispersion of inpatient days among patients who do experience a hospitalization, we propose a negative binomial hurdle outcome model. A hurdle model [101] is a two-part mixture model consisting of a point mass at zero followed by a zero-truncated count distribution for the positive observations. The choice of count distribution can vary, but the negative binomial distribution is attractive because it accounts for overdispersion in the counts.

Let Y represent the number of inpatient days in a fiscal year. The probability of experiencing a hospitalization (i.e., any positive number of inpatient days) is expressed as $\Pr(Y > 0) = \pi$ where $0 < \pi < 1$. For $y = 1, 2, ...,$ the probability that $Y = y$ is given by $\Pr(Y = y) = \frac{\pi p(y;\mu,r)}{1-p(0;\mu,r)}$, where $p(y;\mu,r)$ denotes the probability distribution function of a negative binomial distribution with mean μ and overdispersion parameter r, and $p(0;\mu,r)$ denotes the negative binomial distribution evaluated at 0. The mean count among hospitalized patients is given by $E(Y|Y > 0) = \nu = \frac{\mu}{1-p(0;\mu,r)}$, while the overall all mean among hospitalized and non-hospitalized patients is $E(Y) = \frac{\pi\mu}{1-p(0;\mu,r)}$. The variance of the negative binomial hurdle model is $V(Y) = \nu(\nu - \mu) + \frac{\pi\tau^2}{1-p(0;\mu,r)}$ where $\tau^2 = \mu(1 + \mu/r)$ is the variance of the negative binomial distribution.

Turning to our case study, let Y_{ij} denote the number of inpatient days for the j^{th} patient in the i_th county. In this context, the negative binomial hurdle model is expressed as

$$\begin{aligned}
\Pr(Y_{ij} = y_{ij} \mid x_{ij}, z_{ij}, \phi_i, \mu_{ij}, r) &= (1-\pi_{ij})1_{(y_{ij}=0)} + \pi_{ij}\text{TNegBin}(y_{ij};\mu_{ij},r)1_{(y_{ij}>0)} \\
&= (1-\pi_{ij})1_{(y_{ij}=0)} + \left[\frac{\pi_{ij}}{1-(\frac{r}{\mu_{ij}+r})^r}\frac{\Gamma(y_{ij}+r)}{\Gamma(r)y_{ij}!}\right. \\
&\quad \left.\times\left(\frac{\mu_{ij}}{\mu_{ij}+r}\right)^{y_{ij}}\left(\frac{r}{\mu_{ij}+r}\right)^{r}\right]1_{(y_{ij}>0)}, \quad r>0 \qquad (5.2)
\end{aligned}$$

where $\pi_{ij} = \Pr(Y_{ij} > 0)$ is the probability of hospitalization and $\text{TNegBin}(y_{ij};\mu_{ij},r)$ is a truncated negative binomial distribution with parameters μ_{ij} and r. We model π_{ij} and μ_{ij} as

$$\begin{aligned}
\text{logit}(\pi_{ij}) &= \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + z_{ij}\gamma + \phi_{2i} \\
\ln(\mu_{ij}) &= \boldsymbol{x}_{ij}^T\boldsymbol{\gamma} + z_{ij}\theta + \phi_{3i}, \qquad (5.3)
\end{aligned}$$

where z_{ij} is a binary indicator for race and, as in Equation 5.1, $\boldsymbol{x}_{ij}$ represent a set of patient-level covariates and ϕ_{2i} and ϕ_{3i} are spatial random effects.

To encourage maximal spatial smoothing, we assign the random effects ϕ_{1i}, ϕ_{2i}, and ϕ_{3i} independent intrinsic conditional autoregressive (ICAR) priors [32]. Let $k = 1$ denote the propensity score model, $k = 2$ denote the outcome model for the risk of hospitalization, and $k = 3$ denote the outcome model for the mean number

of inpatient days. The ICAR prior for ϕ_{ki} takes the conditional form

$$\phi_{ki} \mid \boldsymbol{\phi}_{k(-i)}, \sigma_k^2 \sim \mathrm{N}\left(\frac{1}{m_i}\sum_{h\sim i}\phi_{kh}, \sigma_k^2/m_i\right), \quad k = 1, 2, 3, \tag{5.4}$$

where $h \sim i$ indicates that county h is a geographic neighbor of county i, m_i is the number of neighbors, and, for model k, σ_k^2 is the conditional variance of ϕ_{ki} given the remaining spatial effects, $\boldsymbol{\phi}_{k(-i)}$. Following Brook's Lemma [33], the joint distribution for $\boldsymbol{\phi}_k = (\phi_{k1}, \ldots, \phi_{kn})^T$ is given by

$$\pi(\boldsymbol{\phi}_k \mid \sigma_k^2) \propto \exp\left(-\frac{1}{2\sigma_k^2}\boldsymbol{\phi}_k^T \boldsymbol{Q} \boldsymbol{\phi}_k\right), \quad k = 1, 2, 3, \tag{5.5}$$

where $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{A}$ is a spatial structure matrix of rank $n-1$, with $\boldsymbol{M} = \mathrm{diag}(m_1, \ldots, m_n)$ and $\boldsymbol{A}$ representing an $n \times n$ adjacency matrix with $a_{ii} = 0$, $a_{ih} = 1$ if $i \sim h$, and $a_{ih} = 0$ otherwise. When a fixed intercept is included in the model, a sum-to-zero constraint must be applied to $\boldsymbol{\phi}_k$ to ensure an identifiable model.

The ICAR prior is appealing because it imposes spatial smoothing, reflecting the intuition that adjacent spatial units are more similar in terms of access to health care, resources and policies than non-neighbors. Moreover, by promoting localized spatial smoothing and information sharing from surrounding geographic areas, the ICAR prior reduces uncertainty in estimating the propensity scores and, in turn, the ATT.

5.2.3 Treatment Effect Estimation

Because our outcome analysis involves a two-part model, we can estimate an ATT for each part of the model. Furthermore, it is possible to combine results from both models in order to estimate an ATT across the entire population of hospitalized and non-hospitalized patients. The three ATTs are more formally defined as

$$\Delta_1 = E(Y_1^{(1)} - Y_1^{(0)} \mid Z = 1) \tag{5.6}$$

$$\Delta_2 = E(Y^{(1)} - Y^{(0)} \mid Z = 1, Y > 0) \tag{5.7}$$

$$\Delta_3 = E(Y^{(1)} - Y^{(0)} \mid Z = 1). \tag{5.8}$$

where

$$Y_1 = \begin{cases} 1, & \text{if } Y > 0 \\ 0, & \text{otherwise} \end{cases}$$

is a binary indicator of hospital admission (i.e., at least one inpatient day). Equation (5.6) yields the ATT of the racial disparity in the risk of hospitalization. Equation (5.7) yields the ATT of racial differences in the mean number of inpatient days among those who were hospitalized. Lastly, equation (5.8) yields the ATT of the disparity in the mean number of inpatient days among the entire susceptible population (those who had a hospitalization and those who did not). Each of the ATTs can be of practical interest depending on the clinical or public health question at hand.

5.3 Model Fitting

In this work, we adopt a Bayesian model fitting approach and assign prior distributions to all model parameters. As a default, we assign weakly informative $\mathrm{N}(0, 1e5)$ priors to fixed effects and, to ensure robust precision estimates, specify $\mathrm{Ga}(1, 0.5)$ priors for the spatial precision terms, where $\mathrm{Ga}(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b. We fit the propensity score and outcome models separately, thus avoiding the so-called "feedback" issue that can arise when the models are fit jointly under a fully Bayesian approach [55]. We use approximate Bayesian methods for posterior inference. Specifically, we adopt the efficient integrated nested Laplace approximation (INLA) proposed by Rue et al. [56]. INLA uses a Laplace approximation to estimate the joint posterior of the model parameters, yielding improved computational capabilities over standard Markov chain Monte Carlo routines. This method can be readily implemented in the R package `INLA` (www.r-inla.org), where the `Besag` option is used to specify the ICAR prior. The posterior means of the propensity scores are then used to match individuals.

We match individuals without replacement using the logit of the estimated propensity score with a caliper of 0.2 times the standard deviation as recommended by Austin [2] using the R package `Matching` [85]. Once a matched sample is constructed, we fit spatial hurdle models. Because INLA does not have a built-in option for fitting hurdle models, we adapt the work of Quiroz et al. [102] by constructing an $N \times 2$ matrix of values indicating a hospitalization and the number of visits for each patient where N is the total number of observations. We then jointly fit a binomial model to the binary portion, i.e. any inpatient days, and a zero-inflated negative binomial

model to the number of visits, where missing values are imposed for those who were not admitted. The zero-inflated negative binomial with missing values forces INLA to fit a zero-truncated negative binomial which, combined with the binomial model for any inpatient days, yields a negative binomial hurdle model. We construct an estimate of each of the three ATTs using "standardization", a technique that allows us to marginalize across the population by estimating the predicted responses, $Y_{ij}^{(0)}$ and $Y_{ij}^{(1)}$, under the observed and counterfactual racial group. In order to construct a credible interval (CrI) around this estimate, we use the `inla.posterior.sample` function within R-INLA to obtain 1000 Monte Carlo draws from the approximate posterior distribution. The mean of the three estimates across the 1000 samples is reported as the estimated ATT, and the corresponding 95% CrI is derived from the 2.5 and 97.5 percentiles.

5.4 Simulation Study

5.4.1 Data Description

In order to assess the properties of spatial propensity score matching with spatial hurdle outcome modeling, we conducted a simulation study. The goal of this study was to assess how well the proposed method accounts for omitted variable bias, namely when geographic confounders related to both the exposure and outcome of interest are unknown or unmeasured and thus are not included in the propensity score or outcome model.

To mirror the spatial structure of our application, we generated patient-level

data clustered at the county level across the southeastern United States. To emulate the geographic structure in our application, we used the US Census county-level adjacency matrix for South Carolina, Georgia, and Alabama (U.S. Census Bureau, 2014) This matrix contains $n = 272$ counties and $1,528$ pairwise adjacencies. We generated 100 datasets with treatment assignment and outcome according to the following propensity score (Equation 5.9) and outcome models (Equations 5.10 and 5.11):

$$\text{logit}(e_{ij}) = \text{logit}[\Pr(Z_{ij} = 1 | X_{ij} = x_{ij}, V_i = v_i, \phi_{1i})] = \alpha_0 + x_{ij}\alpha_1 + v_i\alpha_2 + \phi_{1i} \quad (5.9)$$

$$\begin{aligned}\text{logit}(\pi_{ij}) = \text{logit}[\Pr(Y_{1ij} = 1 | Z_{ij} = z_{ij}, X_{ij} = x_{ij}, V_i = v_i, \phi_{2i})] = \beta_0 + x_{ij}\beta_1 \\ + z_{ij}\beta_2 + v_i\beta_3 + \phi_{2i}\end{aligned} \quad (5.10)$$

$$\begin{aligned}\ln(\mu_{ij}) = \ln[E(Y | Z_{ij} = z_{ij}, X_{ij} = x_{ij}, V_i = v_i, \phi_{3i}, Y_{ij} > 0)] = \quad \gamma_0 + x_{ij}\gamma_1 \\ + z_{ij}\gamma_2 + v_i\gamma_3 + \phi_{3i}\end{aligned} \quad (5.11)$$

where $i = 1, \ldots, 272$, $j = 1, \ldots, n_i$, x_{ij} is a patient-level covariate generated from a Bernoulli(0.05) distribution and V_i is a county-level covariate generated from a N(10,3) distribution. The fixed effect coefficients were set at $\alpha_0 = 0.2$, α_1 -1.5, $\alpha_2 =$ -0.1, $\beta_0 =$ -1.0, $\beta_1 = 0.5$, $\beta_2 =$ -0.3, $\beta_3 = 0.1$, $\gamma_0 = 1$, $\gamma_1 =$ -0.5, $\gamma_2 = 0.3$, and $\gamma_3 =$ 0.1; n_i was set to 100; ϕ_{1i}, ϕ_{2i}, and ϕ_{3i} were simulated from ICAR models given in Equation (5.5) with $\sigma^2 = 1$ to mimic the spatial variation observed in the case study.

5.4.2 Results

Table 5.1 presents results of the simulation study. The goal of this study was to assess the ability of the proposed spatial propensity score matching and spatial hurdle model to capture the true ATTs compared to fitting only a spatial hurdle outcome model without matching when relevant county-level covariates were left out of the analysis. As Table 5.1 indicates, fitting only an outcome model resulted in poor performance, whereas fitting the propensity score and outcome model yielded lower bias and RMSE and reasonable coverage. For example, when estimating Δ_2, the mean difference in inpatient days among those who were hospitalized, fitting a propensity score model in addition to the outcome model resulted in 94% coverage compared to 54% coverage when only an outcome model was fit. Misspecification of the model for the mean count is especially detrimental, as both the coefficients and the overdispersion parameter may be affected, potentially leading to extreme counts. The results of this simulation study support the use of spatial propensity score matching prior to fitting the spatial hurdle model as it appears to capture the true risk difference even when county-level fixed effects are ignored during model fitting. As it is not uncommon for these covariates to be unavailable to the analyst, spatial matching provides a practical strategy to account for unmeasured geographic confounding.

Table 5.1: Results of the simulation study: Mean absolute bias, RMSE, and 95% coverage of the three ATTs (Equations (5.6) - (5.8)) when (left) only an outcome model is fit and (right) a propensity score (PS) and outcome are fit under an omitted spatial covariate scenario

	Outcome Model Only			**PS + Outcome Model**		
ATT	Bias	RMSE	Coverage	Bias	RMSE	Coverage
Δ_1	0.007	0.010	90	0.007	0.009	91
Δ_2	0.311	0.396	54	0.159	0.304	94
Δ_3	0.118	0.165	80	0.111	0.149	87

5.5 Analysis of Racial Disparities in Hospitalization and Inpatient Days

5.5.1 Data Description

The data consist of 23,533 veterans (9,695 NHB; 13,838 NHW) with type 2 diabetes living in Georgia, Alabama and South Carolina in 2014. This geographic region of 272 counties had a mean county sample size of $n = 86.5$ (range: 1 to 1,385). Figure 5.1 displays the per-county percent of NHB veterans, percent of veterans who experience a hospitalization, the mean number of inpatient days among those with a hospitalization, and the mean number of inpatient days across all patients. These maps suggest spatial variation in racial clustering and hospitalization patterns. Approximately 29.0% of the patients experienced a hospitalization in 2014 (31.3% for NHBs, 28.3% for NHWs). Among those who experienced a hospitalization, the mean number of inpatient days was 3.9 (4.9 for NHBs, 3.6 for NHWs). Across all patients – i.e., those with and without a hospitalization – the mean number of inpatient days was 1.1 (1.3 for NHBs, 1.0 for NHWs).

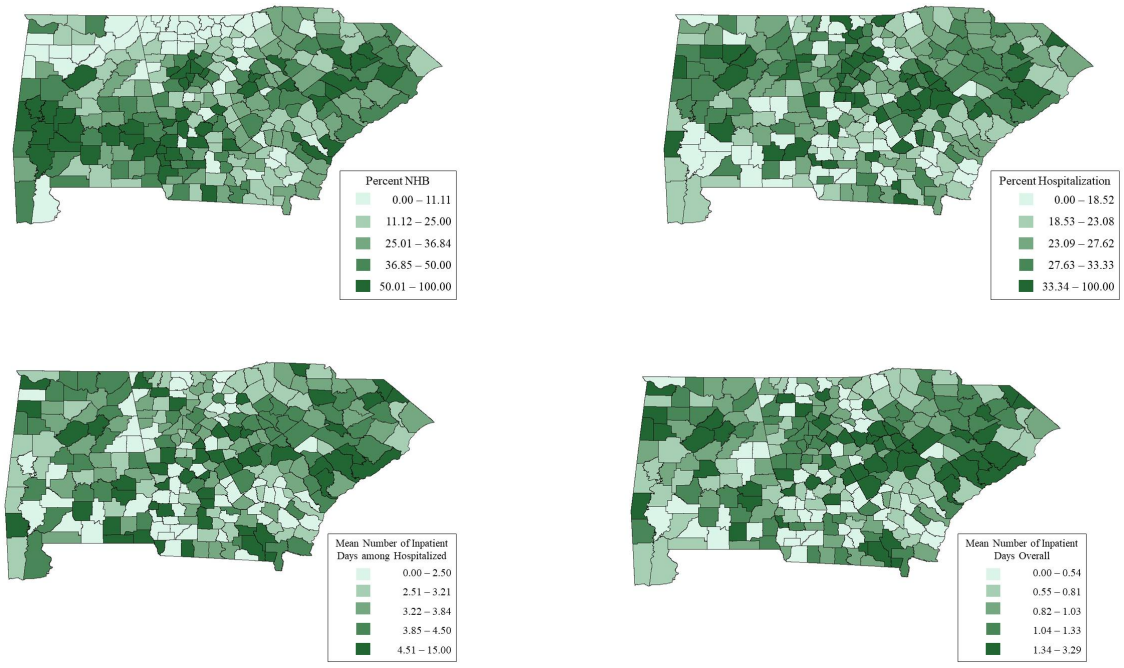

Figure 5.1: Percent of patients who are NHB (top left), percent hospitalization (top right), mean number of inpatient days among those who were hospitalized (bottom left), and mean number of inpatients days in the entire sample comprising hospitalized and non-hospitalized patients (bottom right)

5.5.2 Analysis and Results

We first fit a logistic propensity score model that included patient-level covariates for age, sex, service connected disability and comorbidity burden as well as a county-level spatial random effect. Accounting for comorbidity burden is critical, as this allows us to compare hospitalization patterns among patients with similar disease profile. We then matched patients based on the logit of the estimated propensity score and a caliper of 0.2 times the standard deviation of the logit. The caliper discarded $n = 2,687$ NHB patient observations due to poor matches, thus ensuring

a well-balanced sample. In the matched sample, 29.71% of patients (NHB: 30.5%, NHW: 28.9%) experienced a hospitalization. Among those hospitalized, the mean number of inpatient days was 4.0; across the entire matched sample, the mean number of inpatient days was 1.2 days (range: 0 to 86 days). While NHB patients in the original sample were, on average, younger, had a greater service connected disability, were more likely to be female, and experienced a greater comorbidity burden, we saw balance across these patient features in the matched sample.

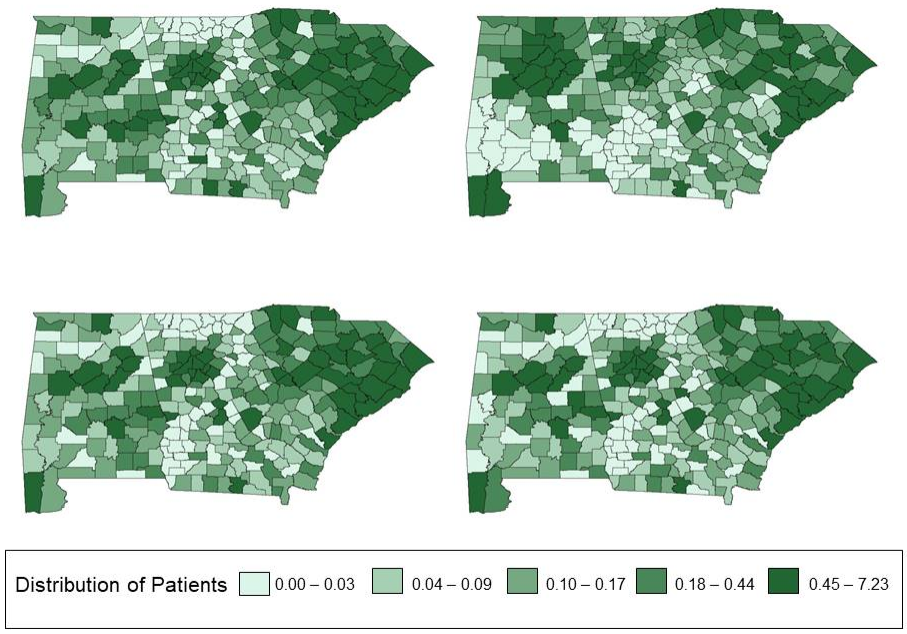

Figure 5.2: Spatial distribution (percent of racial group living in the county) between NHB and NHW veterans in unmatched (top row), and spatially matched (bottom row) samples

.

Figure 5.2 displays the spatial distribution of NHB and NHW veterans in the

unmatched and spatially matched samples. The spatial distribution of NHB and NHW veterans varied in the unmatched sample, implying that NHBs and NHWs were concentrated in different areas. While a high percent of both NHB and NHW veterans live in urban areas such as Atlanta, NHW veterans alone appear to be concentrated in northern Georgia, where only 0.00% to 0.03% of NHB veterans reside (lightest shade on the map). This spatial imbalance was eliminated once a spatially matched sample is created. In the spatially matched sample, the distribution of NHW veterans (the "controls") more closely mimics the nearly unchanged distribution of NHB veterans (the "treated"), indicating that we have selected geographically well-matched controls.

Once a well-balanced sample was constructed, we fit the spatial negative binomial hurdle model with the same patient-level covariates and a spatial random effect. We then used the estimated coefficients from the two parts of the model to form a standardized estimate of the difference in the risk of hospitalization, the mean number of inpatient days among patients with a hospitalization, and the mean number of inpatient days across all patients. The reported 95% CrI was constructed using the 2.5 and 97.5 percentiles of the sample distribution of risk and mean differences. Estimates of the three ATTs are reported in Table 5.2. Negative estimates indicate that NHB veterans have a lower probability of hospitalization or mean number of inpatient days while positive estimates indicate the opposite.

The results in Table 5.2 indicate that the difference in the risk of hospitalization between NHB and NHW patients is 1.5 percentage points, with NHB patients having a lower risk of experiencing a hospitalization (-0.015, 95% CrI = [-0.028, -0.001]).

Table 5.2: Estimated ATTs in the racial disparity of the risk of hospitalization (Δ_1), mean number of inpatient days among those hospitalized (Δ_2), and mean number of inpatient days across the entire patient population (hospitalized and non-hospitalized patients) (Δ_3)

ATT	**Estimate**	**95% CrI**
Δ_1	-0.015	(-0.028, -0.001)
Δ_2	0.431	(-0.214, 1.138)
Δ_3	0.112	(-0.219, 0.476)

Conversely, NHB patients who are hospitalized spend on average approximately one half day longer in the hospital than NHW patients (0.431 days, 95% CrI = [-0.214, 1.138]). While the 95% posterior interval does include 0, the posterior probability that $\Delta_2>0$ was 0.84, providing moderate evidence of an increase in the mean number of inpatient days for NHB veterans. Lastly, we observe a slight increase in the mean number of inpatient days among NHBs compared to NHWs across the entire population, i.e. those who were and were not hospitalized (0.112 days, 95% CrI = [-0.219, 0.476]).

In a similar analysis that excluded spatial random effects, we observed a notable difference in the estimate of the mean number of inpatient days among those who had been hospitalized: the estimated ATT was 0.678 days (95% CrI = [0.187, 1.200]. Thus, ignoring geographic confounding would result in a potentially misleading estimate suggesting that hospitalized NHB veterans have a large, highly significant increase in length of stay compared to their NHW counterparts. This result emphasizes the need to control for geographic confounding and indicates that part of the racial disparity in the number of inpatient days observed in the literature can be attributed to racial minorities living in areas with facilities that have differential

hospitalization patterns in comparison to facilities in predominantly white areas.

5.6 Discussion

We have combined recent work in spatial propensity score matching and spatial hurdle models for hierarchical data to understand racial differences in hospitalization and inpatient days. To conduct this type of analysis, we utilized spatial random effects in the propensity score and outcome models to account for spatial variation due to potential unmeasured geographic confounders. The spatial effects were assigned CAR priors that promote local spatial smoothing to improve small-area estimation. We performed this work within the Bayesian modeling framework in R-INLA, software that is free and readily accessible to researchers.

In simulation, we explored the impact of fitting only the outcome portion of the analysis versus two-stage propensity score and outcome modeling in the presence of unknown geographic confounding. We observed favorable performance of the two-stage modeling across the estimation of the three ATTs: the risk difference in hospitalization, the mean number of inpatient days among those hospitalized, and the mean number of inpatient days overall. The analysis that included only an outcome model performed particularly poorly in the estimation of the mean number of inpatient days. These results indicate that first using spatial propensity score matching to create a balanced sample and then fitting a spatial hurdle model for the outcome is a reasonable approach when true geographic confounders may be unmeasured or unknown.

Our application study explored racial differences in hospitalization and inpatient days among a sample of diabetic veterans residing in the southeastern United States. We achieved both patient-level covariate balance and spatial balance in the matched sample and proceeded with the spatial hurdle outcome model to estimate the three ATTs of interest. The small but statistically significant estimate for the risk difference in hospitalization between NHBs and NHWs indicated that after accounting for patients characteristics and geographic residence, NHB patients may be less likely to be hospitalized. While the estimate for the mean difference in inpatient days among those who were hospitalized was not statistically significant, potentially related to the small sample size of hospitalized individuals, it did suggest that hospitalized NHB patients on average spend a half day longer in the hospital compared to hospitalized NHW patients. However, ignoring geographic confounding would have led to a potential overestimate of the disparity. These results could have implications in cost and patient wellness. Future interventions may target barriers to patient-provider communication concerning hospitalizations or aim to identify regions with issues in inpatient resource access due to limited facilities or workforce. Clinical programs that target disease management and outpatient treatment should be inclusive of racial minorities in order to achieve goals of reducing hospitalizations and extended inpatient stays among a chronic disease population such as type 2 diabetics.

Future work might explore longitudinal trends in hospitalization and inpatient stays between NHB and NHW patients. Additionally, as this work is restricted to hospitalizations within the VHA, additional database resources may be explored to understand racial differences among the broader patient population who may be

receiving inpatient care, namely emergency or trauma care, at local or specialized hospitals outside of the VHA.

Chapter 6

Conclusions of Research

6.1 Summary

In summary, this dissertation has proposed methodology to address geographic confounding, which occurs when measured or unmeasured confounding factors vary spatially, through the augmentation of the propensity score model with a spatial random effect. This work has proposed methods in both propensity score weighting and propensity score matching. In simulation, it has been shown that ignoring space in the presence of true geographic confounding has detrimental effects on bias, RMSE and coverage. Furthermore, in the case of propensity score matching, fitting a spatial outcome model in addition to a spatial propensity score model improves estimation of the ATT while fitting a spatial propensity score model in addition to a spatial outcome model achieves better estimates and coverage in the presence of unknown spatial confounders compared to a spatial outcome model alone. This work has been

applied to a diabetic cohort of veterans receiving care within the VHA in fiscal year 2014. In the application studies, we have analyzed the effect of race on glycemic control, diabetes education and care, and hospitalization practices. We have seen that addressing geographic confounding can result in a reduction or amplification of the estimate of a disparity and that conducting spatial propensity score analysis can elucidate information about disparities that traditional propensity score analysis would miss.

6.2 Implications

This research provides a meaningful intersection of recent work in multilevel propensity score analysis and spatial propensity score analysis. It allows researchers to address geographic confounding while preserving patient-level data. Additionally, it complements existing methods in exploratory spatial data analysis and allows an analyst to construct a global, minimally biased effect estimate of interest. With proper elucidation regarding the complexities and nuances, the proposed methodology can be applied to studies across a spectrum of public health issues, namely: racial disparities, spatially varying exposures, and program evaluation.

6.3 Limitations and Extensions

While this research aims to address confounding that occurs due to geographic differences, it is not intended to promote geographic relocation as a solution to narrow differences between exposed and unexposed groups of individuals; rather, it provides

indication that targeting geographic regions of need may be beneficial and can complement efforts to target vulnerable exposure groups. It is fair to recognize that geographic boundaries are often arbitrary and may not represent ideal or homogeneous regions of comparison. Lastly, it has been noted that differences between individuals reporting a county of residence and those for whom this information is missing may be present and should be explored, especially when making generalizations from the results of studies that require geographic information as inclusion criteria. This work can be extended to space-time outcome models or space-by-exposure interactions.

Bibliography

[1] Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: What it is and what it isn't. BMJ. 1996;312(7023):71–72.

[2] Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research. 2011;46(3):399–424.

[3] Grossman J, Mackenzie FJ. The Randomized Controlled Trial: Gold standard, or merely standard? Perspectives in Biology and Medicine. 2005;48(4):516–534.

[4] Kunz R, Oxman AD. The unpredictability paradox: Review of empirical comparisons of randomised and non-randomised clinical trials. British Medical Journal. 1998;317(7167):1185–1190.

[5] Kwiatkowski T, Libman R, Tilley B, Lewandowski C, Grotta J, Lyden P, et al. The Impact of Imbalances in Baseline Stroke Severity on Outcome in the National Institute of Neurological Disorders and Stroke Recombinant

Tissue Plasminogen Activator Stroke Study. Annals of Emergency Medicine. 2005;45(4):377 –384.

[6] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974 Oct;66(5):688–701.

[7] Caliendo M, Kopeinig S. Some Practical Guidance for the Implementation of Propensity Score Matching. Journal of Economic Surveys. 2008;22(1):31–72.

[8] Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983;70(1):41–55.

[9] Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. Stat Med. 2013 Aug;32(19):3373–3387.

[10] Rubin D. Randomization analysis of experimental data: the Fisher randomization test. Journal of the American Statistical Association. 1980;75(371):591–593.

[11] Schwartz S, Gatto N, Campbell U. What Would Have Been is Not What Would Be: Counterfactuals of the Past and Potential Outcomes of the Future. In: Shrout P, Keyes K, Ornstein K, editors. Causality and Psychotherapy: Finding the Determinants of Disorders and their Cures. Oxford: Oxford University Press; 2011. p. 25–46.

[12] Little R, Rubin D. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. Annual Review of Public Health. 2000;21:121–145.

[13] Stuart EA. Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science. 2010 Feb;25(1):1–21.

[14] Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. Statistics in Medicine. 2004;23(19):2937–2960.

[15] Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. Epidemiology. 2000;11(5):550–560.

[16] Davidian M. Double Robustness in Estimation of Causal Treatment Effects; 2007. NC State University Lecture.

[17] Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. Journal of the American Statistical Association. 1994;89(427):846.

[18] Rosenbaum PR, Rubin DB. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. The American Statistician. 1985;39(1):33–38.

[19] Rubin DB. Matching to remove bias in observational studies. Biometrics. 1973;29:159–184.

[20] Austin PC. A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine. 2014;33(6):1057–1069.

[21] Rosenbaum PR. A Characterization of Optimal Designs for Observational Studies. Journal of the Royal Statistical Society Series B (Methodological). 1991;53(3):597âĂŞ610.

[22] Hansen BB. Full Matching in an Observational Study of Coaching for the SAT. Journal of the American Statistical Association. 2004;99(467):609–618.

[23] S B, Park S, Won E, Park Y, Kim H. Propensity Score Matching:A Conceptual Review for Radiology Researchers. Korean Journal of Radiology. 2015;16(2):286–296.

[24] Firebaugh G. A Rule for Inferring Individual-Level Relationships from Aggregate Data. American Sociological Review. 1978;43(4):557–572.

[25] Arpino B, Mealli F. The specification of the propensity score in multilevel observational studies. Computational Statistics and Data Analysis. 2011;55(4):1770–1780.

[26] Chagas ALS, Toneto R, Azzoni CR. A Spatial Propensity Score Matching Evaluation of the Social Impacts of Sugarcane Growing on Municipalities in Brazil. International Regional Science Review. 2012 Jan;35(1).

[27] Gonzales R, Aranda P, Mendizabal J. Is microfinance truly useless for poverty reduction and womens empowerment? A Bayesian spatial-propensity score matching evaluation in Bolivia. Partnership for Economic Policy (PEP); 2016. 2016-06.

[28] Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. New York, NY: Cambridge University Press; 2007.

[29] Abadie A, Imbens GW. Matching on the Estimated Propensity Score. National Bureau of Economic Research, Inc; 2009.

[30] An W. Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. Sociological Methodology. 2010;40(1):151–189.

[31] McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. Statistics in Medicine. 2009;28(1):94–112.

[32] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics. 1991;43(1):1–20.

[33] Banerjee S, Carlin BP, Gelfand AE. Hierarchical modeling and analysis for spatial data. 2nd ed. Boca Raton, Fla: CRC Press, Taylor and Francis Group; 2014.

[34] Institute of Medicine (US). How Far Have We Come in Reducing Health Disparities? Progress Since 2000: Workshop Summary. 2012;.

[35] Cowie CC, Rust KF, Byrd-Holt DD, Eberhardt MS, Flegal KM, Engelgau MM, et al. Prevalence of Diabetes and Impaired Fasting Glucose in Adults in the U.S. Population. Diabetes Care. 2006;29(6):1263–1268.

[36] Harris MI, Klein R, Cowie CC, Rowland M, Byrd-Holt DD. Is the Risk of Diabetic Retinopathy Greater in Non-Hispanic Blacks and Mexican Americans Than in Non-Hispanic Whites With Type 2 Diabetes?: A U.S. population study. Diabetes Care. 1998;21(8):1230–1235.

[37] Gu K, Cowie CC, Harris MI. Mortality in Adults With and Without Diabetes in a National Cohort of the U.S. Population, 1971–1993. Diabetes Care. 1998;21(7):1138–1145.

[38] Egede LE, Gebregziabher M, Hunt KJ, Axon RN, Echols C, Gilbert GE, et al. Regional, Geographic, and Racial/Ethnic Variation in Glycemic Control in a National Sample of Veterans With Diabetes. Diabetes Care. 2011;34(4):938–943.

[39] Ali MK, Bullard KM, Imperatore G, Barker L, Gregg EW. Characteristics Associated with Poor Glycemic Control Among Adults with Self-Reported Diagnosed Diabetes - National Health and Nutrition Examination Survey, United States, 2007-2010. Morbidity and Mortality Weekly Report. 2012 Jun;61(02):32–37.

[40] Zgibor JC, Gieraltowski LB, Talbott EO, Fabio A, Sharma RK, Karimi H. The Association between Driving Distance and Glycemic Control in Rural Areas. Journal of Diabetes Science and Technology. 2011 May;5(3):494–500.

[41] Salois MJ. Obesity and diabetes, the built environment, and the local food economy in the United States, 2007. Economics and Human Biology. 2012;10(1):35–42.

[42] U S Census Bureau. TIGER/Line Shapefiles; 2014.

[43] CDC. National Diabetes Statistics Report, 2014: Estimates of Diabetes and Its Burden in the United States; 2014.

[44] King DK, Glasgow RE, Toobert DJ, Strycker LA, Estabrooks PA, Osuna D, et al. Self-Efficacy, Problem Solving, and Social-Environmental Support Are Associated With Diabetes Self-Management Behaviors. Diabetes Care. 2010;33(4):751–753.

[45] Ye Y, Bond JC, Schmidt LA, Mulia N, Tam TW. Toward a better understanding of when to apply propensity scoring: a comparison with conventional regression in ethnic disparities research. Annals of Epidemiology. 2012 Oct;22(10):691–697.

[46] Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Causal Inference Models. Biometrics. 2005 Dec;61(4):962–973.

[47] Keele L, Titiunik R, Zubizarreta JR. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2015;178(1):223–239.

[48] Papadogeorgou G, Choirat C, Zigler C. Adjusting for Unmeasured Spatial Confounding with Distance Adjusted Propensity Score Matching; 2016.

[49] Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association. 1952;47(260):663–685.

[50] SAS Institute. The SAS system for Windows: Release 9.2. Cary, NC: SAS Institute; 2011.

[51] Rasmussen S. Modelling of discrete spatial variation in epidemiology with SAS using GLIMMIX. Computer Methods and Programs in Biomedicine. 2004;76(1):83–89.

[52] Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association. 1993 Mar;88(421):9–25.

[53] Wolfinger R, O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation. 1993;48(3-4):233–243.

[54] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken, NJ: Wiley-Interscience; 2004.

[55] McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. Statistics in Medicine. 2009;28(1):94–112.

[56] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the

Royal Statistical Society: Series B (Statistical Methodology). 2009;71(2):319–392.

[57] Carroll R, Lawson AB, Faes C, Kirby RS, Aregay M, Watjou K. Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. Spatial and Spatio-temporal Epidemiology. 2015;p. 45–54.

[58] Resnick HE, Foster GL, Bardsley J, Ratner RE. Achievement of American Diabetes Association Clinical Practice Recommendations Among U.S. Adults With Diabetes, 1999–2002. Diabetes Care. 2006;29(3):531–537.

[59] Walker RJ, Neelon B, Davis M, Egede L. Racial Differences in Spatial Patterns for Poor Glycemic Control in the Southeastern United States. Medical Care. Under review;.

[60] Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine. 2015;34(28):3661–3679.

[61] Yang D, Howard G, Coffey CS, Roseman J. The Confounding of Race and Geography: How Much of the Excess Stroke Mortality among African Americans Is Explained by Geography? Neuroepidemiology. 2004;23(3):118–122.

[62] Wing S, Casper M, Davis WB, Pellom A, Riggan W, Tyroler HA. Stroke mortality maps. United States whites aged 35-74 years, 1962-1982. Stroke. 1988;19(12):1507–1513.

[63] Hodges JS, Reich BJ. Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. The American Statistician. 2010;64(4):325–334.

[64] Hebert PL, Sisk JE, Howell EA. When does a difference become a disparity? Conceptualizing racial and ethnic disparities in health. Health Affairs. 2008 3;27(2):374–382.

[65] Baicker K, Chandra A, Skinner J, Wennberg J. Who You Are And Where You Live: How Race And Geography Affect The Treatment Of Medicare Beneficiaries. Health Affairs. 2004 Jul;.

[66] Mccaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A Tutorial on Propensity Score Estimation for Multiple Treatments Using Generalized Boosted Models. Statistics in Medicine. 2013 Mar;32(19):3388–3414.

[67] Eibner C, Krull H, Brown KM, Cefalu M, Mulcahy AW. Current and Projected Characteristics and Unique Health Care Needs of the Patient Population Served by the Department of Veterans Affairs; 2016.

[68] Miller DR, Safford MM, Pogach LM. Who Has Diabetes? Best Estimates of Diabetes Prevalence in the Department of Veterans Affairs Based on Computerized Patient Data. Diabetes Care. 2004;27(suppl 2):b10–b21.

[69] King KB, Findley TW, Williams AE, Bucknell AL. Veterans With Diabetes Receive Athroplasty More Frequently and at a Younger Age. Clinical Orthopaedics and Related Research. 2013;417(9):3049–3054.

[70] Assari S. Veterans and Risk of Heart Disease in the United States: A Cohort with 20 Years of Follow Up. International Journal of Preventive Medicine. 2014;5(6):703–709.

[71] Kehle S, Greer N, Rutks I. Interventions to Improve VeteransâĂŹ Access to Care: A Systematic Review of the Literature. Department of Veterans Affairs Health Services Research and Development; 2011.

[72] Elnitsky CA, Andresen EM, Clark ME, McGarity S, Hall CG, Kerns RD. Access to the US Department of Veterans Affairs health system: self-reported barriers to care among returnees of Operations Enduring Freedom and Iraqi Freedom. BMC Health Services Research. 2013;13(1).

[73] García-Pérez LE, Álvarez M, Dilla T, Gil-Guillén V, Orozco-Beltrán D. Adherence to Therapies in Patients with Type 2 Diabetes. Diabetes Therapy. 2013 Dec;4(2):175–194.

[74] Heisler M, Smith DM, Hayward RA, Krein SL, Kerr EA. Racial Disparities in Diabetes Care Processes, Outcomes, and Treatment Intensity. Medical Care. 2003;41(11):1221–1232.

[75] Ziemer DC, Miller CD, Rhee MK, Doyle JP, Watkins C, Cook CB, et al. Clinical Inertia Contributes to Poor Diabetes Control in a Primary Care Setting. The Diabetes Educator. 2005;31(4):564–571.

[76] Griffith ML, Boord JB, Eden SK, Matheny ME. Clinical Inertia of Discharge Planning among Patients with Poorly Controlled Diabetes Mellitus. The Journal of Clinical Endocrinology and Metabolism. 2012;97(6):2019–2026.

[77] Chou AF, Brown AF, Jensen RE, Shih S, Pawlson G, Scholle SH. Gender and Racial Disparities in the Management of Diabetes Mellitus Among Medicare Patients. Women's Health Issues. 2007;17(3):150–161.

[78] White RO, Beech BM, Miller S. Health Care Disparities and Diabetes Care: Practical Considerations for Primary Care Providers. Clinical Diabetes. 2009;27(3):105–112.

[79] Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken, NJ: Wiley-Interscience; 2011.

[80] Austin PC. The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. Medical Decision Making. 2009;29(6):661–677.

[81] Arpino B, Cannas M. Propensity score matching with clustered data. An application to the estimation of the impact of caesarean section on the Apgar score. Statistics in Medicine. 2016;35.

[82] Greiner DJ, Rubin DB. Causal Effects of Perceived Immutable Characteristics. The Review of Economics and Statistics. 2011;93(3):775–785.

[83] Davis ML, Neelon B, Neitert PJ, Hunt KJ, Burgette LF, Lawson AB, et al. Addressing geographic confounding through spatial propensity scores: a study of racial disparities in diabetes. Statistical Methods in Medical Research. 2017;.

[84] Nguyen TL, Collins GS, Spence J, Daures JP, Devereaux PJ, Landais P, et al. Double-adjustment in propensity score matching analysis: choosing a threshold for considering residual imbalance. BMC Medical Research Methodology. 2017;17(78).

[85] Sekhon JS. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R. Journal of Statistical Software. 2011;42(7).

[86] Shu L, Stevens GD. Vulnerability and Unmet Health Care Needs. Journal of General Internal Medicine. 2005 Feb;20(2):148–154.

[87] Kilbourne AM, Switzer G, Hyman K, Crowley-Matoka M, Fine MJ. Advancing Health Disparities Research Within the Health Care System: A Conceptual Framework. American Journal of Public Health. 2006;96:2113–2121.

[88] Ackermann RT, Finch EA, Brizendine E, Zhou H, Marrero DG. Translating the Diabetes Prevention Program into the Community: The DEPLOY Pilot Study. American Journal of Preventive Medicine. 2008;35(4):357 – 363.

[89] Natale-Pereira A, Enard K, Nevarez L, Jones LA. The Role of Patient Navigators in Eliminating Health Disparities. Cancer. 2011;117(15):3543–3552.

[90] Maillet NA, Melkus GD, Spollett G. Using Focus Groups to Characterize the Health Beliefs and Practices of Black Women with Non-Insulin-Dependent Diabetes. The Diabetes Educator. 1996 Feb;22(1).

[91] Veterans Health Administration Office of Health Equity. Health Equity Action Plan (HEAP); 2016.

[92] Menke A, Casagrande S, Geiss L, Cowie C. Prevalence of and trends in diabetes among adults in the united states, 1988-2012. JAMA. 2015;314(10):1021–1029.

[93] Currie CJ, Morgan CL, Peters JR. The Epidemiology and Cost of Inpatient Care for Peripheral Vascular Disease, Infection, Neuropathy, and Ulceration in Diabetes. Diabetes Care. 1998;21(1):42–48.

[94] Economic Costs of Diabetes in the U.S. in 2012. Diabetes Care. 2013;.

[95] Kalra AD, Fisher RS, Axelrod P. Decreasd Length of Stay and Cumulative Hospitalized Days Despite Increased Patient Admissions and Readmissions in an Area of Urban Poverty. Journal of General Internal Medicine. 2010;25(9):930–935.

[96] Kominski GF, Morisky DE, Afifi AA, Kotlerman JB. The Effect of Disease Management on Utilization of Services by Race/Ethnicity: Evidence from the Florida Medicaid Program. The American Journal of Managed Care. 2008;14(3):168–172.

[97] Schneider E, Zaslavsky A, Epstein A. Racial disparities in the quality of care for enrollees in medicare managed care. JAMA. 2002;287(10):1288–1294.

[98] Fisher ES, Wennberg JE, Stukel TA, Skinner JS, Sharp SM, Freeman JL, et al. Associations among hospital capacity, utilization, and mortality of US Medicare beneficiaries, controlling for sociodemographic factors. 2000;34(6):1351–1362.

[99] Ashton CM, Petersen NJ, Souchek J, Menke TJ, Yu HJ, Pietz K, et al. Geographic Variations in Utilization Rates in Veterans Affairs Hospitals and Clinics. New England Journal of Medicine. 1999;340(1):32–39.

[100] Davis ML, Neelon B, Nietert PJ, Hunt KJ, Burgette LF, Lawson AB, et al. Propensity Score Matching for Multilevel Spatial Data: Accounting for Geographic Confounding in Health Disparity Studies;Submitted December 2017.

[101] Mullahy J. Specification and testing of some modified count data models. Journal of Econometrics. 1986;33(3):341 – 365.

[102] Quiroz ZC, Prates MO, Rue H. A Bayesian approach to estimate the biomass of anchovies off the coast of Peru. Biometrics. 2015;71(1):208–217.