Published in final edited form as:

Stat Methods Med Res. 2019 March ; 28(3): 734–748. doi:10.1177/0962280217735700.

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

Melanie L Davis¹, Brian Neelon¹, Paul J Nietert¹, Kelly J Hunt¹, Lane F. Burgette², Andrew B Lawson¹, Leonard E Egede³

¹Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA

²RAND Corporation, Arlington, VA, USA

³Division of General Internal Medicine Froedtert, The Medical College of Wisconsin, Milwaukee, WI, USA

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 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 doubly robust 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 doubly robust 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 doubly robust estimate that adjusts only for patient-level information, and a hierarchical spatial doubly robust 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.

Keywords

Diabetes control; doubly robust estimator; geographic confounding; health disparities; propensity scores; spatial data analysis

Corresponding author: Melanie L Davis, Department of Public Health Sciences, Medical University of South Carolina, Charleston 29425, SC, USA. davml@musc.edu.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

1 Introduction

Diabetes is the seventh leading cause of death in the USA and is associated with a number of adverse health outcomes, including stroke, heart disease, kidney failure, and amputation.¹ 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.^{2–4} These disparities are explained in part by individual-level factors such as age, sex, marital status, and comorbidities.^{5,6} 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.^{7,8} 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.⁹ 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 analysis¹⁰ (PSA) offers a principled approach to causal inference in observational studies, and has gained increasing traction in health disparities studies in recent years.^{11,12} 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.¹³ A particularly attractive weight-based estimator is the “doubly robust” (DR) estimator,¹⁴ 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.¹¹

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,^{15–18} these methods have been limited to non-clustered data in which the response variable is a region-level proportion. Arpino and Mealli¹⁹ and Li et al.¹¹ have 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.¹¹ 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.

2 Spatial propensity score analysis

2.1 Overview of propensity score weighting methods

We begin by briefly reviewing the inferential properties of PSA as outlined in Rosenbaum and Rubin¹⁰ and summarized more recently in Lunceford and Davidian.²⁰ 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,²¹ 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 $\tau = E(Y_1) - E(Y_0)$. Because we observe only one of (Y_1, Y_0), unbiased estimation of the ATE, τ , requires that we instead estimate the average effect conditional on *observed* treatment assignment, that is, $\tau^* = E(Y_1|Z = 1) - E(Y_0|Z = 0)$.

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, $\tau^* = \tau$, and the observed treatment difference τ^* 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 $\tau^* = \tau$. Nevertheless, it may be reasonable to assume that (Y_1, Y_0) are conditionally independent of Z given a vector of covariates $\mathbf{X}$. This is commonly referred to as the “no unmeasured confounding” assumption.¹⁰ Under this assumption, the ATE can be identified from the observed data ($Y, Z, \mathbf{X}$) through the equation

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

where the outer expectation is taken with respect to the distribution of covariates $\mathbf{X}$ in the entire population and $\mathbf{x}$ is an observed realization of the random variable $\mathbf{X}$. The third line of equation (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.¹¹ propose using the average controlled difference (ACD) as a *descriptive* estimand analogous to the ATE, where the ACD is defined as

$$\Delta = E_X[E(Y|X = \mathbf{x}, Z = 1) - E(Y|X = \mathbf{x}, Z = 0)] \tag{2}$$

Because the latter expression is identical to the last line of equation (1), we use Δ throughout to denote either ATE or 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, $e(\mathbf{x}) = \Pr(Z = 1|X = \mathbf{x})$, is the conditional probability of exposure given $\mathbf{X}$, where the so-called “overlap” condition, $0 < e(\mathbf{x}) < 1$, is assumed to hold. Rosenbaum and Rubin¹⁰ established that $e(\mathbf{x})$ functions as a balancing score such that

$$\Delta = E\left[\frac{ZY}{e(\mathbf{x})} - \frac{(1-Z)Y}{1-e(\mathbf{x})}\right] \tag{3}$$

when both the overlap and unconfoundedness assumptions hold. Hence, an unbiased, Horvitz-Thompson type²² 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

$$\text{logit}(e_i) = \text{logit}[e(\mathbf{x}_i)] = \text{logit}[\Pr(Z_i = 1 | \mathbf{X}_i = \mathbf{x}_i)] = \mathbf{x}_i^T \boldsymbol{\beta}, \quad i = 1, \dots, n \quad (4)$$

If model (4) is correctly specified, an unbiased, inverse-probability weight (IPW) estimator of the ACD is given by

$$\widehat{\Delta}^{IPW} = n^{-1} \sum_{i=1}^n \left[\frac{Z_i Y_i}{\widehat{e}_i} - \frac{(1 - Z_i) Y_i}{1 - \widehat{e}_i} \right] \quad (5)$$

where $\widehat{e}_i$ denotes the estimated propensity score for subject i . To guard against misspecification of the propensity score model, Robins et al.²³ developed a semiparametric doubly robust (DR) estimator of the form

$$\begin{aligned} \widehat{\Delta}^{DR} &= \frac{1}{n} \sum_{i=1}^n \widehat{\Delta}_i \\ \widehat{\Delta}_i &= \left[\frac{Z_i Y_i}{\widehat{e}_i} - \frac{(Z_i - \widehat{e}_i) \widehat{Y}_{i1}}{\widehat{e}_i} \right] - \left[\frac{(1 - Z_i) Y_i}{1 - \widehat{e}_i} + \frac{(Z_i - \widehat{e}_i) \widehat{Y}_{i0}}{1 - \widehat{e}_i} \right] \end{aligned} \quad (6)$$

where $\widehat{Y}_{i1}$ and $\widehat{Y}_{i0}$ are predicted outcomes obtained by regressing Y on $\mathbf{X}$ 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 (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 $\widehat{\Delta}^{DR}$ is given by

$$s^2 = \frac{1}{n^2} \sum_{i=1}^n (\widehat{\Delta}_i - \widehat{\Delta}^{DR})^2 \quad (7)$$

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

2.2 A doubly robust estimator for hierarchical spatial data

Li et al.¹¹ recently extended the DR estimator to the multilevel setting, where $(Y_{ij}, Z_{ij}, \mathbf{X}_{ij})$ denote the data for the j -th subject in cluster i . Li et al. propose the following hierarchical DR estimator of the ACD

$$\begin{aligned}\widehat{\Delta}^{DR} &= \frac{1}{N} \sum_{i=1}^n \sum_{j=1}^{n_i} \widehat{\Delta}_{ij} \\ \widehat{\Delta}_{ij} &= \left[\frac{Z_{ij} Y_{ij}}{\widehat{e}_{ij}} - \frac{(Z_{ij} - \widehat{e}_{ij}) \widehat{Y}_{ij1}}{\widehat{e}_{ij}} \right] - \left[\frac{(1 - Z_{ij}) Y_{ij}}{1 - \widehat{e}_{ij}} + \frac{(Z_{ij} - \widehat{e}_{ij}) \widehat{Y}_{ij0}}{1 - \widehat{e}_{ij}} \right]\end{aligned}\quad (8)$$

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 (7), the large-sample variance estimator of $\widehat{\Delta}^{DR}$ is given by

$$s^2 = \frac{1}{N^2} \sum_{i=1}^n \sum_{j=1}^{n_i} (\widehat{\Delta}_{ij} - \widehat{\Delta}^{DR})^2 \quad (9)$$

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 (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 $\mathbf{x}_{ij}$ represent a set of patient-level covariates. The spatial propensity score model is given by

$$\text{logit}(e_{ij}) = \text{logit}[\Pr(Z_{ij} = 1 | \mathbf{X}_{ij} = \mathbf{x}_{ij}, \phi_{1i})] = \mathbf{x}_{ij}^T \beta + \phi_{1i} \quad (10)$$

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

$$\text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, \mathbf{X}_{ij} = \mathbf{x}_{ij}, \phi_{2i})] = \mathbf{x}_{ij}^T \gamma + z_{ij} \alpha + \phi_{2i} \quad (11)$$

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) prior²⁴ that takes the conditional form

$$\phi_{ki} | \phi_{k(-i)}, \sigma_k^2 \sim N\left(\frac{1}{m_i} \sum_{h \sim i} \phi_{kh}, \sigma_k^2 / m_i\right), \quad k = 1, 2 \quad (12)$$

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, $\phi_{k(-i)}$. 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,²⁵ the joint distribution for $\phi_k = (\phi_{k1}, \dots, \phi_{kn})^T$ is given by

$$\pi(\phi_k | \sigma_k^2) \propto \exp\left(-\frac{1}{2\sigma_k^2} \phi_k^T \mathbf{Q} \phi_k\right), \quad k = 1, 2 \quad (13)$$

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

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 (10) and (11) can be easily implemented using off-the-shelf software such as SAS PROC GLIMMIX.²⁶ 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 $\mathbf{Q}$ in expression (13), and then incorporating this as part of a user-defined covariance matrix in PROC GLIMMIX. Details can be found in Rasmussen.²⁷ 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^{28,29} is an iterative

estimation procedure achieved through Taylor series expansions of the response about current estimates of the fixed and random effects.³⁰ 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 (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 a in equation (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.³¹ For our application, we adopt the efficient integrated nested Laplace approximation (INLA) proposed by Rue et al.³² 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 $N(0, 1e5)$ priors to fixed effects and $Ga(1, 5e-05)$ priors for the spatial precision (i.e. inverse variance) terms, where $Ga(a, b)$ 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.³³ for the regression coefficients and spatial variances. Alternative prior specifications are discussed in Section 4.

3 Simulation study

3.1 Data description

To examine the performance of the proposed spatial DR estimator, we conducted 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.³⁴ 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

$$\text{logit}[\Pr(Z_{ij} = 1 | X_{ij} = x_{ij}, \phi_{1i})] = \beta_0 + x_{ij}\beta_1 + \phi_{1i} \quad (14)$$

$$\text{logit}[\Pr(Y_{ij} = 1 | Z_{ij} = z_{ij}, X_{ij} = x_{ij}, \phi_{2i})] = \gamma_0 + x_{ij}\gamma_1 + z_{ij}\alpha + \phi_{2i}, \quad i = 1, \dots, 272; j = 1, \dots, n_i$$

(15)

where X_{ij} was simulated according to a $N(5, 2)$ 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 $a = 0.90$; 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 (13) with $\sigma_{\phi_1}^2$ and $\sigma_{\phi_2}^2$ 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.³⁵ 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 (14) and (15) 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 (13), with $\sigma^2 = 2$ and coefficients $\beta_2 = -0.1$ and $\beta_3 = 0.1$ for the respective spatially varying covariates in the propensity score model and $\gamma_2 = 0.3$ and $\gamma_3 = -0.3$ 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.

3.2 Results

Table 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 (14) and (15). Rows indicate the varying levels of spatial heterogeneity (σ_{ϕ}^2) and sample sizes (n_i) used to generate the data, including the case where the simulated data contained no spatial heterogeneity ($\sigma_{\phi}^2 = 0$). 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 $\sigma_\phi^2 = 1$, and 0.92 when $n_i = 100$ and $\sigma_\phi^2 = 9$. 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 $\sigma_\phi^2 = 1$ and to 0.16 when $\sigma_\phi^2 = 9$. 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 $\sigma_\phi^2 = 9$, 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 2 demonstrates the doubly robust property of the spatial DR estimator. As in Table 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. $\sigma_\phi^2 = 9$), 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.¹¹

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 (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 frequentist model predictions to construct the ACD and fully Bayesian methods for spatial prediction in subsequent analyses involving the outcome model.

Table 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 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.

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 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 elsewhere³⁶; 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 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.⁵

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 1, first column). 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 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 1, top row). Many of these same areas also exhibited

above-average uncontrolled HbA1c (Figure 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 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 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.³⁷ Standardized differences were used to compare the covariate distributions across the two race groups.¹³ We also derived county-specific weighted proportions of NHB and NHW veterans and mapped the distribution of the unweighted, 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 4 and Figure 2. As Table 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 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, a non-spatial DR estimate, and a 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 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,²⁴ and ICAR priors with $Ga(1, 1)$ and $Ga(1, 0.5)$ 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 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 (9). We therefore report the large-sample standard errors in Table 5.

Table 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 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.³⁸

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 confounding in estimating the overall race disparity, spatial random effect predictions from a well-constructed fully Bayesian outcome model alone can lend investigators valuable information in allocating resources and targeting communities. Figure 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”.³⁹ 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 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.

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 feedback³¹ 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 work³⁸ 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, X . As Hodges and Reich⁴⁰ 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 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 (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, $\hat{a}$, in outcome model (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.⁴¹

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.⁴² 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 5, investigators can disentangle the factors that contribute to racial disparities, a goal of recent disparity studies.⁴³

Future work might accommodate multiple exposure categories, taking advantage of recent methods for causal inference among multiple treatment groups.⁴⁴ 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 exposures.

Acknowledgments

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the VHA Health Services Research and Development (HSR&D) program (PI: Leonard Egede) (grant no. CIN 13–418). 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 4 and Figure 1 as part of previous work.³⁶

References

1. CDC. National diabetes statistics report, 2014: Estimates of diabetes and its burden in the United States. Technical report, 2014.
2. Cowie CC, Rust KF, Byrd-Holt DD, et al. Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population. *Diab Care* 2006; 29: 1263–1268.
3. Harris MI, Klein R, Cowie CC, et al. 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. *Diab Care* 1998; 21: 1230–1235.
4. Gu K, Cowie CC and Harris MI. Mortality in adults with and without diabetes in a national cohort of the U.S. population, 1971–1993. *Diab Care* 1998; 21: 1138–1145.
5. Egede LE, Gebregziabher M, Hunt KJ, et al. Regional, geographic, and racial/ethnic variation in glycemic control in a national sample of veterans with diabetes. *Diab Care* 2011; 34: 938–943.
6. Ali MK, Bullard KM, Imperatore G, et al. 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; 61: 32–37.
7. Zgibor JC, Gieraltowski LB, Talbott EO, et al. The association between driving distance and glycemic control in rural areas. *J Diab Sci Technol* 2011; 5: 494–500.
8. Salois MJ. Obesity and diabetes, the built environment, and the local food economy in the United States, 2007. *Econ Human Biol* 2012; 10: 35–42. [PubMed: 21561816]
9. King DK, Glasgow RE, Toobert DJ, et al. Self-efficacy, problem solving, and social-environmental support are associated with diabetes self-management behaviors. *Diab Care* 2010; 33: 751–753.
10. Rosenbaum PR and Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; 70: 41–55.
11. Li F, Zaslavsky AM and Landrum MB. Propensity score weighting with multilevel data. *Stat Med* 2013; 32: 3373–3387. [PubMed: 23526267]
12. Ye Y, Bond JC, Schmidt LA, et al. Toward a better understanding of when to apply propensity scoring: a comparison with conventional regression in ethnic disparities research. *Ann Epidemiol* 2012; 22: 691–697. [PubMed: 22902041]
13. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavior Res* 2011; 46: 399–424.
14. Bang H and Robins JM. Doubly robust estimation in missing data and causal inference models. *Biometrics* 2005; 61: 962–973. [PubMed: 16401269]
15. Chagas ALS, Toneto R and Azzoni CR. A spatial propensity score matching evaluation of the social impacts of sugarcane growing on municipalities in Brazil. *Int Regional Sci Rev* 2012; 35: 48–69.
16. Gonzales R, Aranda P and Mendizabal J. Is microfinance truly useless for poverty reduction and womens empowerment? A Bayesian spatial-propensity score matching evaluation in Bolivia. Working Paper 2016–06, Partnership for Economic Policy (PEP), 2016.
17. Keele L, Titiunik R and Zubizarreta JR. Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *J Royal Stat Soc: Ser A (Statistics in Society)* 2015; 178: 223–239.
18. Papadogeorgou G, Choirat C and Zigler C. Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching, 2016 arXiv:1610.07583.

19. Arpino B and Mealli F. The specification of the propensity score in multilevel observational studies. *Comput Stat Data Anal* 2011; 55: 1770–1780.
20. Lunceford JK and Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat Med* 2004; 23: 2937–2960. [PubMed: 15351954]
21. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Edu Psychol* 1974; 66: 688–701.
22. Horvitz D and Thompson D. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc* 1952; 47: 663–685.
23. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; 89: 846.
24. Besag J, York J and Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 1991; 43: 1–20.
25. Banerjee S, Carlin BP and Gelfand AE. *Hierarchical modeling and analysis for spatial data*, 2nd ed. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2014.
26. Institute SAS. *The SAS system for Windows: Release 9.2*. Cary, NC: SAS Institute, 2011.
27. Rasmussen S. Modelling of discrete spatial variation in epidemiology with SAS using GLIMMIX. *Comput Meth Progr Biomed* 2004; 76: 83–89.
28. Breslow NE and Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993; 88: 9–25.
29. Wolfinger R and O'Connell M. Generalized linear mixed models a pseudo-likelihood approach. *J Stat Comput Simul* 1993; 48: 233–243.
30. Fitzmaurice GM, Laird NM and Ware JH. *Applied longitudinal analysis*. Hoboken, NJ: Wiley-Interscience, 2004.
31. McCandless LC, Gustafson P and Austin PC. Bayesian propensity score analysis for observational data. *Stat Med* 2009; 28: 94–112. [PubMed: 19012268]
32. Rue H, Martino S and Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J Royal Stat Soc: Series B (Stat Methodol)* 2009; 71(2): 319–392.
33. Carroll R, Lawson A, Faes C, et al. Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial Spatio-temporal Epidemiol* 2015; 45–54.
34. US Census Bureau. 2014 TIGER/Line Shapefiles (machine-readable data files)/prepared by the U.S. Census Bureau, 2014.
35. Resnick HE, Foster GL, Bardsley J, et al. Achievement of American Diabetes Association clinical practice recommendations among U.S. adults with diabetes, 1999–2002. *Diab Care* 2006; 29: 531–537.
36. Walker RJ, Neelon B, Davis M, et al. Racial differences in spatial patterns for poor glycemic control in the southeastern United States. *Ann Epidemiol* 2017; (under review).
37. Austin PC and 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. *Stat Med* 2015; 34: 3661–3679. [PubMed: 26238958]
38. Yang D, Howard G, Coffey CS, et al. The confounding of race and geography: How much of the excess stroke mortality among African Americans is explained by geography? *Neuroepidemiology* 2004; 23: 118–122. [PubMed: 15084780]
39. Wing S, Casper M, Davis WB, et al. Stroke mortality maps. United States whites aged 35–74 years, 1962–1982. *Stroke* 1988; 19: 1507–1513. [PubMed: 3201510]
40. Hodges JS and Reich BJ. Adding spatially-correlated errors can mess up the fixed effect you love. *Am Stat* 2010; 64: 325–334.
41. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci* 2010; 25: 1–21. [PubMed: 20871802]
42. Hebert PL, Sisk JE and Howell EA. When does a difference become a disparity? Conceptualizing racial and ethnic disparities in health. *Health Affairs* 2008; 27: 374–382. [PubMed: 18332492]

43. Baicker K, Chandra A, Skinner J, et al. Who you are and where you live: How race and geography affect the treatment of Medicare beneficiaries. *Health Affairs* 2004; 33–44. [PubMed: 15318565]
44. Mccaffrey DF, Griffin BA, Almirall D, et al. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* 2013; 32: 3388–3414. [PubMed: 23508673]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 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.

Figure 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.

Figure 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).

Table 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 ^a	R ^b	C ^c	B ^a	R ^b	C ^c	B ^a	R ^b	C ^c	B ^a	R ^b	C ^c	B ^a	R ^b	C ^c	B ^a	R ^b	C ^c
<i>n_i</i> =100																		
$\sigma_{\phi}^2 = 0$	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
$\sigma_{\phi}^2 = 1$	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
$\sigma_{\phi}^2 = 4$	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
$\sigma_{\phi}^2 = 9$	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
<i>n_i</i> =50																		
$\sigma_{\phi}^2 = 0$	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
$\sigma_{\phi}^2 = 1$	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
$\sigma_{\phi}^2 = 4$	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
$\sigma_{\phi}^2 = 9$	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
<i>n_i</i> =25																		
$\sigma_{\phi}^2 = 0$	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
$\sigma_{\phi}^2 = 1$	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
$\sigma_{\phi}^2 = 4$	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
$\sigma_{\phi}^2 = 9$	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

^aMean absolute bias.

^bRoot mean squared error (RMSE).

^c95% coverage.

Table 2.

Mean bias, RMSE, and 95% coverage of the partially misspecified DR estimators under various sample sizes, spatial variances, and estimation methods.

	Maximum Likelihood						Penalized Quasi-likelihood						Bayesian					
	Propensity score ^a			Outcome ^b			Propensity score			Outcome			Propensity score			Outcome		
	B ^c	R ^d	C ^e	B ^c	R ^d	C ^e	B ^c	R ^d	C ^e	B ^c	R ^d	C ^e	B ^c	R ^d	C ^e	B ^c	R ^d	C ^e
<i>n_F</i> =100																		
$\sigma_{\phi}^2 = 1$	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
$\sigma_{\phi}^2 = 4$	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
$\sigma_{\phi}^2 = 9$	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
<i>n_F</i> =50																		
$\sigma_{\phi}^2 = 1$	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
$\sigma_{\phi}^2 = 4$	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
$\sigma_{\phi}^2 = 9$	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
<i>n_F</i> =25																		
$\sigma_{\phi}^2 = 1$	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
$\sigma_{\phi}^2 = 4$	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
$\sigma_{\phi}^2 = 9$	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

Note: Columns indicate which model was misspecified.

^aPropensity score model (14) misspecified (no random intercept), outcome model (15) correctly specified.

^bOutcome model (15) misspecified (no random intercept), propensity score model (14) correctly specified.

^cMean absolute bias.

^dRoot mean squared error (RMSE).

^e95% coverage.

Table 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.

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

Note: 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.

Table 4.

Balance of covariates between NHB and NHW veterans in unweighted, non-spatially weighted, and spatially weighted samples.

Variable	Unweighted			Non-spatially weighted			Spatially weighted		
	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

“Stand. Diff”: absolute value of the standardized difference.

Table 5.

Estimated risk differences and 95% confidence intervals (CI) in percent uncontrolled HbA1c under various models.

Model	Risk difference	95% CI	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