A Bayesian Hierarchical Modeling Framework for Geospatial Analysis of Adverse Pregnancy Outcomes *

Cecilia Balocchi†‡, Ray Bai§¶, Jessica Liu†, Silvia P. Canelón∗∗, Edward I. George†, Yong Chen∗∗, Mary R. Boland∗∗††

Abstract

Studying the determinants of adverse pregnancy outcomes like stillbirth and preterm birth is of considerable interest in epidemiology. Understanding the role of both individual and community risk factors for these outcomes is crucial for planning appropriate clinical and public health interventions. With this goal, we develop geospatial mixed effects logistic regression models for adverse pregnancy outcomes. Our models account for both spatial autocorrelation and heterogeneity between neighborhoods. To mitigate the low incidence of stillbirth and preterm births in our data, we explore using class rebalancing techniques to improve predictive power. To assess the informative value of the covariates in our models, we use posterior distributions of their coefficients to gauge how well they can be distinguished from zero. As a case study, we model stillbirth and preterm birth in the city of Philadelphia, incorporating both patient-level data from electronic health records (EHR) data and publicly available neighborhood data at the census tract level. We find that patient-level features like self-identified race and ethnicity were highly informative for both outcomes. Neighborhood-level factors were also informative, with poverty important for stillbirth and crime important for preterm birth. Finally, we identify the neighborhoods in Philadelphia at highest risk of stillbirth and preterm birth.

1 Introduction

1.1 Background

Although there have been rapid advancements in maternal health over the past several decades, adverse pregnancy outcomes such as stillbirth, preterm birth, neonatal death, post neonatal death, and low birth weight continue to be major public health problems. While there has been a 90 percent reduction in infant mortality over the past century [28], other adverse outcomes such as low birth weight (i.e. birth weight of 2500 grams or less) and stillbirth (i.e. fetal death at or after 20 weeks of pregnancy) have remained largely unchanged [19, 42]. For example, in 2013, a total of

---

*Keywords and phrases: conditional autoregressive model, preterm birth, spatial statistics, stillbirth, random effects model
†Department of Economics and Statistics, University of Torino, Torino, Italy
‡Co-first author. Email: cecilia.balocchi@unito.it
§Department of Statistics, University of South Carolina, Columbia, SC, USA
¶Co-first author. Email: RBAI@mailbox.sc.edu
†The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
∗∗Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, PA, USA
††Email: bolandm@upenn.edu
23,595 stillbirths were reported in the United States [24]. This represents a country-wide statistic and therefore the number of stillbirths in any individual city for a given year is often very sparse. Preterm birth (i.e. birth before 37 weeks of pregnancy) is another adverse pregnancy outcome that has gradually increased over time, in spite of improved treatments. Preterm birth is far more common than stillbirth. The Centers of Disease Control and Prevention (CDC) estimated that in 2014, preterm birth affected one of every 10 infants in the United States [11]. According to the CDC, the U.S. preterm birth rate also rose to 9.93% in 2017, a 1% rise from 2016 (9.85%) and the third straight year of increases in this rate (9.57% in 2014) [27].

Adverse pregnancy outcomes are significantly associated with increased risk of neonatal mortality and morbidity, adverse neuro-developmental and cognitive outcomes, and increased health care costs [33, 48]. To reduce the risk of adverse pregnancy outcomes, it is crucial to understand the factors that may contribute to them. Gaining insight into these risk factors can help to guide clinical interventions. For example, if an obstetrician knows that a patient’s stillbirth risk is likely to be higher because of certain preexisting conditions, they can provide targeted clinical care to minimize the risk of stillbirth [42]. This care might entail increased monitoring and more frequent prenatal visits, while more aggressive steps include early induction of labor or an elective Cesarean section [42].

Understanding the risk factors for adverse pregnancy outcomes can also guide public health policy. Research on adverse prenatal outcomes can be used to expand access to clinical services, particularly for underserved communities and subpopulations at highest risk [40]. For example, it has been estimated that 5.0-7.3 percent of U.S. preterm-related deaths can be attributable to smoking during pregnancy [39]. Based on these findings, state Medicaid programs have been required to cover tobacco-cessation counseling and drug therapy for pregnant women without cost sharing since 2010 [39].

1.2 Related work and our contributions

The massive growth of electronic health records (EHRs) since the passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009 has greatly enabled the study of individual (or patient-level) risk factors and facilitated targeted clinical interventions for adverse pregnancy outcomes [31, 29]. Researchers have identified maternal race, periodontal disease, socioeconomic status, immigration status, family history of low birth weight babies, and the presence of preexisting conditions such as diabetes or chronic hypertension as being significantly associated with stillbirth or preterm birth [14, 26, 48, 24, 10].

Researchers have also identified environmental (or neighborhood-level) risk factors for adverse pregnancy outcomes. There is often considerable geographic variation in overall occurrence and risk for preterm birth, stillbirth, and low weight birth [50, 18, 40]. Numerous studies have suggested that neighborhood characteristics such as local exposure to pollutants, residential segregation, crime levels, and income inequality may also contribute to adverse pregnancy outcomes and other maternal outcomes such as severe maternal morbidity [44, 18, 17, 15, 22, 29, 30]. Studying the neighborhood distribution of adverse pregnancy outcomes also provides opportunities for targeted intervention strategies of the subpopulations at highest risk [40].

In this paper, we introduce a Bayesian hierarchical modeling framework for geospatial analysis of adverse pregnancy outcomes. Specifically, we propose Bayesian conditional autoregressive (CAR) and multilevel logistic regression models that incorporate both patient-specific attributes of the patients and geographic features of the neighborhoods in which they reside. Our CAR model
induces spatial autocorrelation between geographic neighbors and thus facilitates principled sharing of neighborhood information. In the event that there is insignificant spatial autocorrelation, we also propose an independent random effects model that can account for between-neighborhood heterogeneity. Our models allow us to study the geographic variation of adverse pregnancy outcomes and identify high-risk neighborhoods, in addition to quantifying potential risk factors for these outcomes. A model selection criterion is used to select the model with which to conduct the final analysis.

Several other researchers have also employed Bayesian spatial models to model adverse pregnancy outcomes, including spatially-varying regression models [44] and Bayesian point processes [50]. However, it does not appear as though CAR models have been used before to address this problem. Our approach differs from previous works because we employ mixed effects logistic regression models with a spatially dependent CAR prior for the random effect. Furthermore, Zahrieh et al. [50] focused only on stillbirth, while Warren et al. [44] focused only on preterm birth. In this paper, we use our models to analyze each of these two outcomes.

Since stillbirth is an extremely rare event and preterm birth is a relatively rare event, we are faced with the issue of class imbalance. As a result, our models may not adequately weight the characteristics of patients who experienced stillbirth or preterm birth, and we may ultimately produce biased estimates of the regression coefficients [8]. To mollify this issue, we explore the use of Synthetic Minority Over-sampling Technique (SMOTE) [8] with our mixed effects models. SMOTE is a class rebalancing technique which generates new synthetic cases of the rare outcome. SMOTE is routinely used for prediction of rare health events with simple logistic regression [51]. However, to the best of our knowledge, it has not been used for Bayesian mixed effects logistic regression models before. The main challenge with deploying SMOTE in this context is preserving neighborhood information from the U.S. Census Bureau when we create new cases of the rare event. We elaborate on this in Section 4.1.

Finally, we advocate the use of a measure we call Bayes-p (originally introduced by Makowski et al. [25] as the Bayesian Probability of Direction) as a quantitative measure for variable informativeness to assess directly how well a variable effect can be distinguished from zero. Simply taking point estimates of the posterior odds ratios may not offer a complete picture of the exposure-outcome associations, and examining the posterior credible intervals of the odds ratio can be complex and difficult to interpret. Moreover, posterior credible intervals of the log-odds ratio do not necessarily quantify the extent to which an effect size is different from zero, especially if zero is very close to one of the endpoints of the interval. Finally, instead of testing for “significant” vs. “insignificant” effects, it may be of greater interest to study which factors are the most informative for stillbirth or preterm birth, together with point estimates of their effects. The Bayes-p measure allows health practitioners and policymakers to quantify how different from zero is the effect of each individual or neighborhood covariate.

We apply our methodology to a case study of stillbirth and preterm birth in the city of Philadelphia, Pennsylvania. Using patient data from the University of Pennsylvania Health System (also known as Penn Medicine) and publicly available data from the United States (U.S.) Census Bureau from 2010 to 2017, we conduct a geospatial analysis of stillbirth and preterm birth in Philadelphia. The city of Philadelphia is a particularly compelling case study because it is a large urban area that has been experiencing population growth and rapid changes to its built environment for the first time in decades [2]. In addition, there is a vast repository of electronic health records (EHRs) at Penn Medicine, which greatly facilitates the study of adverse pregnancy outcomes within
Figure 1: Maps of neighborhood proportions of adverse pregnancy outcomes, either stillbirth (Left panel) or preterm birth (Right panel), measured as the number of adverse outcomes in a neighborhood divided by the number of deliveries in a neighborhood (over all years). The data is displayed only for census tracts included in our study (with more than 10 deliveries and with at least a neighboring region with more than 10 deliveries), while regions colored in gray are the ones excluded from our study.

The rest of the paper is organized as follows. In Section 2, we describe the dataset that motivated our study. In Section 3, we introduce our Bayesian mixed effects logistic regression models. In Section 4, we describe the SMOTE rebalancing procedure for handling class imbalance and the Bayes-$p$ measure for assessing the relative informativeness of different predictors. In Section 5, we present the main findings from our case study of stillbirth and preterm birth in Philadelphia. Section 6 concludes the paper.

2 Motivating data

In order to conduct our analysis of stillbirth and preterm birth, we first obtained data on patients who had deliveries at Penn Medicine hospitals from 2010 to 2017. We used a previously developed and validated algorithm by Canelón et al. [7] to identify deliveries from Penn Medicine EHRs. After removing 17,305 observations with either missing values in the patient address or with patients living outside of the City of Philadelphia, our dataset contained a total of 46,029 deliveries at hospitals within the University of Pennsylvania Health System between 2010 and 2017. Within each EHR, the following delivery outcomes were annotated: Caesarean section delivery, stillbirth, and preterm birth (each coded as either “1” or “0”). Among these 46,029 records, there were 387 stillbirths and 2902 preterm births. In addition, the EHRs also reported each patient’s residential address, the age of the patient at the time of delivery, binary variables for the patient’s self-reported race/ethnicity
(Hispanic, non-Hispanic White, non-Hispanic Black or non-Hispanic Asian) and a binary variable for multiple birth (e.g. twins, triplets). The vast majority of subjects in our data belonged to only one race, except for two patients who identified as both Asian and White, one who identified as both Asian and Black, and five who identified as both Black and White.

The city of Philadelphia is made up of 384 census tracts determined by the U.S. Census Bureau. We first matched the patients’ residential addresses to their specific longitude and latitude coordinates. Based on these coordinates, we then mapped each of the patients in our dataset to one of 384 census tracts. Because some census tracts were very sparsely populated (e.g. the areas around several rivers, such as the Schuylkill, the Wissahickon and the Pennipack), we removed tracts that had fewer than 10 deliveries in each of the eight years. Because our CAR model (described in Section 3.1) also requires all geographic areas to be connected, we also removed neighborhoods that did not share a border with any other tracts with at least 10 deliveries. This preprocessing procedure ultimately removed 24 census tracts and 110 observations from our dataset, leaving us with $N = 45,919$ deliveries in $n = 363$ census tracts. Among these 45,919 deliveries, we had 385 stillbirths and 2897 preterm births. Figure 1 shows the neighborhood proportions of these two adverse pregnancy outcomes after preprocessing. The dark gray tracts are the tracts that were removed from our analysis.

Next, we augmented our dataset with neighborhood covariates based on the census tracts to which each of the 45,919 patients belonged. Most of the neighborhood-level data was downloaded in June 2020 from https://data.census.gov/cedsci/, the United Sates Census Bureau’s online data dissemination platform. We downloaded data sets documenting racial makeup, poverty status, education level, and number of housing units for each census tract in Pennsylvania for the years 2010 through 2017. Detailed information on data sources is reported in Appendix A. Intrigued by the possibility that crime could be significantly associated with stillbirth or preterm birth, we also included neighborhood-level data on crime. Our crime data came from https://opendataphilly.org, where the Philadelphia Police Department publicly releases the location, time, and type of each reported crime in the city. This crime data was previously analyzed by Balocchi and Jensen [2]. For our study, we included the total number of violent crimes and nonviolent crimes from 2010 to 2017 in each of the census tracts as neighborhood-level predictors. Violent crimes included homicides, rapes, robberies and aggravated assaults, whereas nonviolent crimes included burglaries, thefts and motor vehicle thefts.

Table 1 reports all 19 of the patient-level and neighborhood-level covariates that we considered in our study. A few of the continuous covariates were heavily skewed right – namely, total number of occupied housing units, total number of housing violations, the total number of violent and nonviolent crimes. Consequently, we considered their log transformed values.

We initially included an indicator variable for whether the patient self-identified as White (White) and the neighborhood proportion of those identifying as White (prop_White) as covariates in our model. However, when we included White and prop_White, the variance inflation factors (VIF) for White, prop_White, and the individual and neighborhood-level race factors in Table 1 (Black, Hispanic, Asian, prop_Asian, prop_Hispanic, and prop_Black) were greater than 9.5, indicating the presence of significant multicollinearity [34]. This was because in the majority of cases, we could deduce the values for White and prop_White if the other variables were known. For example, if we observed Black = Hispanic = Asian = 0, then the patient typically identified as White (i.e. White = 1); on the other hand, if one of the non-White patient-level race variables in Table 1 was equal to one, then we could infer that typically the patient’s individual race was
Table 1: The 19 covariates we used in our analysis of adverse pregnancy outcomes in Philadelphia.

| Patient-Level Description | Neighborhood-Level Description |
|---------------------------|-------------------------------|
| age                      | prop_Asian                   |
| Black                    | prop_Hispanic                |
| Hispanic                 | prop_Black                   |
| Asian                    | prop_women_15_to_50          |
| multiple_birth           | prop_women_below_poverty     |
|                          | prop_women_public_assistance |
|                          | prop_women_labor_force       |
|                          | prop_birth_last_12_months    |
|                          | prop_women_HS_grad           |
|                          | prop_women_college_grad      |
|                          | log_occupied_housing         |
|                          | log_housing_violations       |
|                          | log_violent_crime            |
|                          | log_nonviolent_crime         |

Table 3.1: The 19 covariates we used in our analysis of adverse pregnancy outcomes in Philadelphia.

not White \((\text{White} = 0)\). The neighborhood proportion of White residents could also often be determined by taking \(\text{prop}_\text{White} \approx 1 - (\text{prop}_\text{Asian} + \text{prop}_\text{Hispanic} + \text{prop}_\text{Black})\), since the proportions of Philadelphia residents identifying as Native American, Pacific Islander, Alaskan Native or Native Hawaiian were very small and therefore excluded from our analyses. Once we removed \text{White} and \text{prop}_\text{White}, the VIFs for all patient-level and neighborhood-level covariates were below 9.5, indicating much less severe multicollinearity among the remaining covariates.

3 Bayesian hierarchical modeling framework

Let \(y_{ij}\) be a binary response variable, with “1” indicating if an adverse pregnancy outcome occurred for patient \(j\) in the \(i\)th neighborhood. Suppose that we have \(n\) neighborhoods and a total of \(N = \sum_{i=1}^{n} m_i\) observations, where \(m_i\) denotes the number of patients in the \(i\)th neighborhood. We assume that each \(y_{ij}\) follows a conditionally independent Bernoulli distribution, that is,

\[
y_{ij} | p_{ij} \sim \text{Bernoulli}(p_{ij}), \quad i = 1, \ldots, n, j = 1, \ldots, m_i.
\]

A standard model for estimating the \(p_{ij}\)'s in (3.1) is the mixed effects logistic regression model,

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \alpha_i + \mathbf{x}_{ij}^\top \beta.
\]
\( \alpha_i \) is a random effect that accounts for the variation in neighborhood \( i \) that cannot be explained by the \( p \) covariates. Note that in our dataset, the neighborhood covariates in each \( x_{ij} \) in (3.2) also depended on the year in which patient \( j \) in neighborhood \( i \) gave birth. However, we found that the effect of time was fairly negligible, and many of the neighborhood variables did not vary significantly from year to year. For notational simplicity, we do not include a subscript for year in (3.1) or (3.2). In Section 6, we discuss how our model can be extended to account for temporal effects.

A Bayesian modeling approach to (3.2) places prior distributions on the random and fixed effects. The Bayesian paradigm is appealing because it allows us to coherently quantify uncertainty for both random and fixed effects through their posterior distributions. Bayesian modeling also affords us the opportunity to incorporate prior spatial information in (3.2), which encourages sharing of information between contiguous neighborhoods. Finally, the posterior distribution for \( \beta \) will be used to compute the Bayes-\( p \) measure as a measure for relative variable informativeness in Section 4.3.

In this paper, we present two Bayesian mixed effects logistic regression models for conducting geospatial analysis of adverse pregnancy outcomes. If we are interested in modeling spatial dependence between neighborhoods, we can induce spatial autocorrelation in our model (3.2) by placing a CAR prior (described in the next section) on \( \alpha = (\alpha_1, \ldots, \alpha_n)^\top \). On the other hand, if there is no significant spatial autocorrelation, we can still model between-neighborhood heterogeneity by placing independent priors on the \( \alpha_i \)'s in (3.2).

### 3.1 CAR model for modeling spatial dependence

We first introduce the Bayesian CAR modeling approach. In order to incorporate spatial information in our analysis, we employ a CAR prior on the neighborhood-specific random effects [4, 20, 21]. The CAR model is a Gaussian Markov random field which induces spatial dependence through an adjacency matrix for the areal units, which in our case study, are the census tracts of Philadelphia.

While there are many variations of the CAR prior (see, e.g. Lee [20]), we use the proper CAR formulation of Leroux et al. [21]. Leroux et al. [21] define the distribution of each \( \alpha_i \), given the other entries \( \alpha_{-i} \), as a normal distribution centered at a weighted average of a global mean \( \alpha_0 \) and the \( \alpha_j \)'s from neighborhoods that share a border with \( \alpha_i \). That is,

\[
\alpha_i \mid \alpha_{-i}, \alpha_0, \tau_\alpha^2 \sim \mathcal{N} \left( \frac{\rho \sum_j w_{ij} \alpha_j + (1 - \rho) \alpha_0}{\rho \sum_j w_{ij} + (1 - \rho)} , \frac{\tau_\alpha^2}{\rho \sum_j w_{ij} + (1 - \rho)} \right),
\]

where the \( w_{ij} \)'s are adjacency weights that are equal to 1 if the neighborhoods \( i \) and \( j \) share a border and equal to 0 otherwise. The parameter \( \tau_\alpha^2 \) is a global variance parameter. Meanwhile, \( \rho \in [0, 1] \) represents the strength of spatial correlation between the components of \( \alpha \), with larger values of \( \rho \) corresponding to stronger influence of bordering neighborhoods.

We collect the adjacency weights \( w_{ij} \) into an adjacency matrix \( \mathbf{W} \), which we can assume known, because we can easily use the shape files from the U.S. Census Bureau to determine which of the neighborhoods (here, census tracts) share a border. As noted in Balocchi and Jensen [2] and proven in Chapter 3 of Banerjee et al. [3], the joint distribution of \( \alpha \) is uniquely determined by the set of conditional distributions defined in (3.3) and can be written more compactly as

\[
\alpha \mid \alpha_0, \tau_\alpha^2 \sim \mathcal{N} \left( \alpha_0 \cdot \mathbf{1}, \tau_\alpha^2 \mathbf{\Sigma}_{\text{CAR}} \right),
\]

(3.4)
where \( \mathbf{1} \) is a \( n \)-dimensional vector of all 1’s, \( \mathbf{\Sigma}_{\text{CAR}}^{-1} = \rho (\mathbf{D} \mathbf{W} - \mathbf{W}) + (1 - \rho) \mathbf{I}_n \), and \( \mathbf{D} \mathbf{W} - \mathbf{W} \) is the Laplacian matrix based on the neighborhood adjacency matrix \( \mathbf{W} \). Thus, to model the spatial dependence for adverse pregnancy outcomes under model (3.2), we ultimately place the CAR model (3.4) as a the prior on \( \alpha \). To model the uncertainty in the grand mean \( \alpha_0 \) in (3.4), we place a relatively noninformative prior on \( \alpha_0 \),

\[
\alpha_0 \sim \mathcal{N}(0, 100).
\]  

(3.5)

We also place a weakly informative inverse gamma prior on the global variance parameter \( \tau^2_\alpha \),

\[
\tau^2_\alpha \sim \mathcal{IG}(0.1, 0.1),
\]  

(3.6)

and we endow the spatial autocorrelation parameter \( \rho \) with a noninformative uniform prior,

\[
\rho \sim \mathcal{U}(0, 1).
\]  

(3.7)

To complete our prior specification, we need to place a prior on the fixed effects \( \beta \) in (3.2). To this end, we endow \( \beta \) with the prior,

\[
\beta \mid \mathbf{b}_0, \tau^2_\beta \sim \mathcal{N}(\mathbf{b}_0, \tau^2_\beta \mathbf{I}_p),
\]  

(3.8)

where we place a relatively noninformative prior on \( \mathbf{b}_0 \),

\[
\mathbf{b}_0 \sim \mathcal{N}(\mathbf{0}_p, 100 \cdot \mathbf{I}_p),
\]  

(3.9)

and a weakly informative prior on \( \tau^2_\beta \),

\[
\tau^2_\beta \sim \mathcal{IG}(0.1, 0.1).
\]  

(3.10)

In short, our Bayesian CAR model can be summarized as follows. We place the hierarchical CAR prior (3.4)-(3.7) on the random effects \( \alpha \) and the hierarchical prior (3.8)-(3.10) on the fixed effects \( \beta \).

### 3.2 Independent random effects model

The chief benefit of the using the CAR model described in Section 3.1 is that it induces spatial dependence and promotes sharing of information between proximal neighborhoods. However, it may be the case that the census tracts do not have significant spatial autocorrelation. In this case, it is sufficient to model the between-neighborhood heterogeneity with independent priors on the random effects \( \alpha \) in (3.2).

Instead of placing the CAR prior on \( \alpha \), the (non-spatial) Bayesian independent random effects model instead places diffuse independent normal priors on each of the \( \alpha_i \)'s in (3.2) as follows:

\[
\alpha \mid \alpha_0, \tau^2_\alpha \sim \mathcal{N}(\alpha_0 \cdot \mathbf{1}, \tau^2_\alpha \mathbf{I}_n),
\]  

(3.11)

where we place the further priors on \( (\alpha_0, \tau^2_\alpha) \),

\[
\alpha_0 \sim \mathcal{N}(0, 100),
\]  

(3.12)

and

\[
\tau^2_\alpha \sim \mathcal{IG}(0.1, 0.1).
\]  

(3.13)
Note that (3.11) is a special case of (3.4) when the spatial autocorrelation parameter $\rho$ is set equal to 0. The prior on the fixed effects $\beta$ remains the same as the prior for $\beta$ (3.8)-(3.10) in the CAR model.

Both the CAR model of Section 3.1 and the independent random effects model above can be implemented using Markov chain Monte Carlo (MCMC). Thanks to the Pólya-Gamma data augmentation strategy of Polson et al. [36], nearly all updates in our algorithm are available in closed form. Complete details are provided in Section B of the Appendix.

### 3.3 Determining which model to use

Having presented two different Bayesian mixed effects logistic regression models, we now address the issue of model selection. There are various hypothesis tests to determine whether a spatial regression model or a multilevel model with independent random effects is appropriate. For example, Moran’s $I$ statistic [32] is widely used in practice to test for the presence of significant spatial autocorrelation. However, some researchers have suggested that Moran’s $I$ may have limitations, such as being very sensitive to the choice of neighborhood matrix $W$ or being a potentially poor estimator of the spatial dependence parameter $\rho$ when the true $\rho$ is far away from zero [23].

To circumvent these issues, we instead recommend fitting both the CAR model of Section 3.1 and the independent random effects model of Section 3.2 and then using either the Deviance Information Criterion (DIC) [41] or the Watanabe-Akaike information criterion (WAIC) [45] as a model selection criterion. DIC and WAIC allow us to directly compare the model fit for our two models.

For an unknown parameter $\theta$, the deviance is $D(\theta) = -2 \log p(y|\theta)$, where $p(y|\theta)$ is the likelihood for the respective model. The DIC is given by $\text{DIC} = D(\bar{\theta}) + 2p_D$ where the first term is the deviance evaluated at the posterior mean of $\theta$, and $p_D = D(\bar{\theta}) - D(\theta)$ is the effective number of model parameters where $\bar{D}(\theta) = E_{\theta|y}[D(\theta)]$ is the posterior mean deviance. The DIC rewards better fitting models through the first term and penalizes more complex models through the second term. The model with the smallest overall DIC value is preferred. As an alternative to DIC, one can also consider the WAIC. The WAIC is given by $\text{WAIC} = -2 \log p(y|\bar{\theta}) + 2p_W$ where $p(y|\bar{\theta})$ is the posterior predictive density of the observed data, and $p_W = \sum_{i,j} \text{Var}[\log p(y_{ij}|\theta)|y]$, as recommended in [12]. Compared to DIC, WAIC considers the predictive density averaged over the posterior distribution, instead of conditioning on a point estimate, resulting in an approach more in line with the Bayesian framework.

Let $\theta = (\alpha, \beta)$. In our present framework with the mixed effects logistic regression model (3.2), the likelihood is defined as $p(y|\theta) = \prod_{i,j} p(y_{ij}|\theta) = \prod_{i,j} [p_{ij}(\theta)_{y_{ij}} (1 - p_{ij}(\theta))^{1-y_{ij}}]$, where $p_{ij}(\theta) = \exp(\alpha_i + \mathbf{x}_{ij}^T \beta)/\{\exp(\alpha_i + \mathbf{x}_{ij}^T \beta) + 1\}$. In practice, we estimate $\bar{\theta}$, $\bar{D}(\theta)$, $\bar{p}(\theta)$, and $\text{Var}[\log p(y_{ij}|\theta)|y]$ using MCMC samples from the algorithm described in Section B of the Appendix and then compute the DIC as

$$\text{DIC} = 2\bar{D}(\theta) - D(\bar{\theta})$$

and WAIC as

$$\text{WAIC} = -2 \log p(y|\bar{\theta}) + 2 \sum_{i,j} \text{Var}[\log p(y_{ij}|\theta)|y].$$

The model with the smallest DIC or WAIC is chosen as the model with which to perform the final analysis.
4 SMOTE rebalancing and the Bayes-$p$ measure

4.1 Dealing with class imbalance

As discussed in Section 2, we had 385 stillbirths and 2897 preterm births in our final dataset of 45,919 deliveries, or a rate of approximately approximately 0.84\% for stillbirth and 6\% for preterm birth. Since stillbirth and preterm birth are rare events, our model has an underrepresented sample of characteristics of the patients who did experience these outcomes. This could ultimately result in biased estimation, wherein the characteristics of the majority class (or patients who did not experience these outcomes) are overrepresented in the final estimates of regression coefficients and odds ratios [8, 51].

To address this issue of class imbalance, we propose using SMOTE [8] to balance the dataset. SMOTE creates a synthetic sample using a random sample and its $k$-nearest neighbors in the covariate space. For each continuous covariate, SMOTE randomly finds two nearest neighbors in the minority class, produces a new point midway between the two existing points, and adds that new covariate value to the dataset. For binary or categorical data, the SMOTE method takes the most commonly occurring category of nearest neighbors to create new covariate values [43]. Meanwhile, SMOTE also simultaneously undersamples the majority class until the number of cases and controls is roughly the same. After SMOTE rebalancing, a statistical model can then be fit to the rebalanced dataset. In a variety of applications, SMOTE has been shown to reduce bias and improve predictive power, sometimes drastically [8, 51].

In order to use SMOTE with our mixed effects models in Sections 3.1 and 3.2, some care has to be taken to ensure that we preserve the neighborhood information. While we want to create new observations of patients who experienced stillbirth or preterm birth, we do not want to create new neighborhood covariate values. Instead, we want to use the same neighborhood values as those reported by the U.S. Census for each census tract. Accordingly, we used only the patient-level covariates in Table 1, as well as the the year of delivery and the longitude and latitude coordinates of the addresses for the $N$ original subjects to create synthetic data. The new SMOTE-generated longitude and latitude coordinates were then matched to their census tracts, and the neighborhood covariate values for those tracts were augmented to the synthetic data. Finally, we fit the two models in Sections 3.1 and 3.2 to the rebalanced dataset.

To determine if SMOTE provides any advantages, we can use a variety of metrics to assess if SMOTE has improved our model’s predictive power. For example, with a well-chosen threshold $t \in (0, 1)$, we can classify observations in a given test set as either case (“1”) or control (“0”) with the following rule:

\[
\text{Classify } y_{ij} \text{ as } \begin{cases} 
1 & \text{if } \hat{p}_{ij} \geq t, \\
0 & \text{otherwise},
\end{cases}
\]  \hfill (4.1)

where $\hat{p}_{ij}$ is the predicted probability for patient $j$ in neighborhood $i$ experiencing an adverse pregnancy outcome under model (3.2). In our analysis in Section 5, we choose $t$ which maximizes Youden’s $J$ index [49], or the point on the receiver operating characteristic (ROC) curve that is the farthest distance from the identity line (i.e. the farthest away from the diagonal line for a completely random classifier). Based on the classification rule (4.1), we can calculate the true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). The misclassification rate (MR), sensitivity (sens), and specificity (spec) can then be computed respectively as

\[
\begin{align*}
\text{MR} &= \frac{FP + FN}{TP + FP + TN + FN}, \\
\text{sens} &= \frac{TP}{TP + FN}, \\
\text{spec} &= \frac{TN}{TN + FP}.
\end{align*}
\]  \hfill (4.2)
A lower MR tells us that our classification rule (4.1) has higher predictive accuracy for the chosen threshold $t$, while sensitivity is the true positive rate and specificity is the true negative rate. Because the use of a single threshold $t$ may be misleading, we also consider the area under the curve (AUC) of the ROC curve, which is computed using a fine grid of many different thresholds. In general, models with higher AUC are preferred.

We note that the DIC and WAIC criteria described in Section 3.3 cannot be used to compare the models for SMOTE-rebalanced vs. non-rebalanced data. This is because SMOTE creates an entirely new dataset, and DIC and WAIC is only comparable for models that use exactly the same data. For this reason, we recommend the following two-step procedure: 1) determine which model to use (e.g. using DIC or WAIC), and then 2) fit the selected model with SMOTE and see if this improves predictive accuracy.

4.2 The Bayes-$p$ measure

In epidemiological studies, it is customary to use either $p$-values or confidence intervals of the regression coefficients (or odds ratios) to determine the statistical significance of potential risk factors. However, our conclusions can be highly sensitive to the choice of $p$-value significance cutoff or confidence interval percentiles, and it may be preferable to use a quantitative measure of variable informativeness that is more agnostic.

To this end, we propose using the Bayes-$p$ measure, also known as the Probability of Direction [25]. The Bayes-$p$ measure for the $j$th variable is the posterior probability that a coefficient has the sign of its posterior median value, which can be also computed as the maximum posterior probability that the $j$th regression coefficient $\beta_j$ in $\beta$ is either less than zero or greater than zero. That is,

$$\text{Bayes-}p = \max\{\Pr(\beta_j > 0 \mid y), \Pr(\beta_j < 0 \mid y)\}.$$  (4.3)

Bayes-$p$ (4.3) is a posterior probability, which can be estimated straightforwardly from the MCMC samples for each $\beta_j$ in our model, quantifying how much of a coefficient’s distribution is non-overlapping with zero. This measures directly how well the effect of a predictor can be distinguished from zero and thus its informative value. For example, a predictor whose posterior distribution is centered around zero will be less informative than a predictor which with 90% probability is less than zero.

Note that the Bayes-$p$ does not pertain to the hypothesis testing framework and is quite different from a frequentist $p$-value. The Bayes-$p$ pertains to the amount of posterior probability “away” from zero, while the $p$-value is the probability of observing values more extreme than a test statistic under a null hypothesis $H_0: \beta_j = 0$. Moreover, while a smaller $p$-value indicates a higher level of statistical significance, a larger Bayes-$p$ corresponds to a more informative predictor.

Figure 2 illustrates two examples of the Bayes-$p$ measure. In the left panel of Figure 2, the 95% credible interval is $(0.02, 1.98)$ (suggesting statistical significance), while the 99% credible interval is $(-0.29, 2.29)$ (suggesting an insignificant association between exposure and outcome). Whereas the choice of $\alpha \in (0, 1)$ in our $100(1 - \alpha)$% posterior credible intervals can strongly influence our conclusions, the Bayes-$p$ does not rely on a specific choice of $\alpha$. Unlike posterior credible intervals, Bayes-$p$ (4.3) also allows us to directly compare the effects of different potential risk factors. A higher Bayes-$p$ indicates that the variable’s association with an adverse pregnancy outcome is likely to be more influential. In Section 5, we illustrate the utility of Bayes-$p$ in our case study on preterm birth and stillbirth in Philadelphia.
Figure 2: Graphical illustration of the Bayes-$p$ measure, computed on two Gaussian distributions $N(\mu, \sigma^2)$, with the area shaded in gray corresponding to Bayes-$p$ value for each distribution: in the left panel $\mu = 1, \sigma = 0.5$ and in the right panel $\mu = -0.5, \sigma = 0.8$.

We note that the Bayes-$p$ can equivalently be defined in terms of odds ratio $\exp(\beta_j)$, in which case, we have $\text{Bayes-p} = \max\{\Pr(\exp(\beta_j) > 1|y), \Pr(\exp(\beta_j) < 1|y)\}$. While epidemiologists typically use odds ratios to interpret the associations between outcomes and exposures, the $p$-values for the regression coefficients $\beta_j$’s are typically calculated using the estimates of the original $\beta_j$’s. Therefore, we analogously define Bayes-$p$ (4.3) in terms of the original regression coefficients.

5 Analysis of Philadelphia Data

5.1 Methods

We now apply the methodology described in Sections 3 and 4 to the study of stillbirth and preterm birth in Philadelphia. In our dataset (described in detail in Section 2), we had $N = 45,919$ deliveries in $n = 363$ census tracts and $p = 19$ covariates, as shown in Table 1. We fit the Bayesian CAR model with priors (3.4)-(3.10) on $(\alpha, \beta)$ in (3.2), and the Bayesian independent random effects model with the priors (3.11)-(3.13) on $\alpha$ and (3.8)-(3.10) on $\beta$.

For all our models, we ran two MCMC chains using the algorithm in Appendix B, for a total of 5500 iterations each. For each chain, we removed the first 500 iterations as burn-in. We also thinned our chains every 10 iterations, leaving us with a total of 1000 samples from the two chains. In Appendix C.3, we provide trace plots and autocorrelation plots for six model parameters to confirm that the number of iterations we used was sufficient to achieve convergence and that our thinned samples were not too correlated. Before fitting our models, we also standardized all the continuous variables to lie on the same scale. Our model’s estimates of the regression coefficients were then transformed back to their original scale for our final analysis.
Our two models, which we denote as CAR and indRE respectively, were then compared using DIC and WAIC. The model with the lowest DIC and WAIC was selected as the one to conduct the final analysis. For our chosen model (CAR or indRE), we also implemented that model with the SMOTE rebalancing technique described in Section 4.1. We used the R package DMwR to implement SMOTE rebalancing. The SMOTE-rebalanced model was used for our quantitative analysis of individual- and neighborhood-level risk factors if it produced a higher AUC (or higher predictive accuracy).

5.2 Results

Table 2 shows that the model achieving the smallest values of both DIC and WAIC is CAR for stillbirth and indRE for preterm birth. This indicates that the unexplained variation for preterm birth is better described by non-spatial models, suggesting that there might be unmeasured predictors of preterm birth that are not spatially correlated. For stillbirth, the unexplained variation is instead spatially autocorrelated, hinting that some unmeasured covariates that are spatially correlated might help explain the variation in stillbirth. We thus proceeded with our analyses using the CAR model with priors (3.4)-(3.10) on \((\alpha, \beta)\) in (3.2) for stillbirth, and using the independent random effect model with priors (3.11)-(3.13) and (3.8)-(3.10) on \((\alpha, \beta)\) in (3.2) for preterm birth.

In Table 3, we report the predictive performance of the CAR model for stillbirth and indRE model for preterm birth, fit on both the original dataset (denoted with Original) and the dataset rebalanced with SMOTE (denoted with SMOTE). To assess predictive performance, we computed the performance metrics described in Section 4.1 using both a Leave-One-Out (LOO) approach and an “All data” approach. For LOO, we consecutively fit the model on all except one year of data and used our estimated regression coefficients to obtain out-of-sample predictions on the one year of left-out data. We then took the average of our performance metrics across all eight test sets. For the “All data” approach, we computed these performance measures on the entire original dataset. While some observations were used to train the model because they were included in the synthetic SMOTE dataset, this is not an entirely in-sample prediction. However, the model fit on the original dataset (Original) using the entire dataset (“All data”) is performing in-sample predictions, so particular care needs to be used in interpreting these results.

As shown in Table 3, the AUC typically ranged from 0.58 to 0.70. In preliminary work, we also fit a standard logistic regression model, a (non-Bayesian) generalized linear mixed model with neighborhood random effects, and a random forest model to our dataset. We were not able to obtain AUC higher than 0.70 for any of these approaches, either without or with SMOTE. This suggests that higher-resolution covariates than the ones that were available to us are needed to greatly improve predictions for stillbirth and preterm birth.

We now give a glimpse of the results of our analysis with the CAR and indRE models. Table 4 reports the odds ratios and Bayes-\(p\) for both stillbirth and preterm birth under the Original and SMOTE settings. Figure 3 allows us to visualize the differences between these two settings.
Table 3: Predictive performance of the model fit on the rebalanced (SMOTE) and on the original (Original) dataset. Measures are computed both using a Leave-One-Out (LOO) approach, and using the entire dataset (“All data”).

|          | Stillbirth LOO | Stillbirth All data | Preterm birth LOO | Preterm birth All data |
|----------|---------------|---------------------|-------------------|------------------------|
|          | SMOTE | Original | SMOTE | Original | SMOTE | Original | SMOTE | Original |
| AUC      | 0.588 | 0.638 | 0.703 | 0.686 | 0.631 | 0.643 | 0.671 | 0.668 |
| MR       | 0.007 | 0.008 | 0.006 | 0.008 | 0.050 | 0.061 | 0.051 | 0.063 |
| sens     | 0.704 | 0.738 | 0.725 | 0.647 | 0.629 | 0.649 | 0.638 | 0.678 |
| spec     | 0.505 | 0.517 | 0.587 | 0.628 | 0.569 | 0.567 | 0.594 | 0.560 |

(Original vs. SMOTE) by comparing the point estimates and posterior credible intervals of the log-odds ratios for the 19 covariates in our models. Finally, Figure 4 illustrates the spatial distribution of the predicted risk of stillbirth or preterm birth for the neighborhoods in our models. These neighborhood risk probabilities are calculated as \( \hat{p}_i = \exp(\bar{\theta}_i)/(1 + \exp(\bar{\theta}_i)) \), where \( \bar{\theta}_i = \bar{\alpha}_i + \frac{1}{m_i} \sum_{j=1}^{m_i} x_{ij}^\top \hat{\beta} \) is the posterior mean log-odds averaged over all \( m_i \) patients in neighborhood \( i \), and \( \hat{\alpha}_i \) and \( \hat{\beta} \) denote the posterior means of these parameters under the Original setting. The variable names used in these tables and figures are described in Table 1. In the next two sections, we summarize in detail our main findings for stillbirth and preterm birth in Philadelphia. See Appendix C for additional maps displaying the spatial distribution in the city of the random effects for the two outcomes and a comparison of the risks for several pairs of neighborhoods with different characteristics.

5.3 Results for stillbirth

As shown in the first two columns of Table 3, under the LOO approach, SMOTE achieves lower misclassification rate compared to the model fit on the non-rebalanced dataset (Original), while Original has higher AUC, sensitivity and specificity. However, under the “All data” approach, SMOTE has the highest AUC overall, reaching a value of 0.70. Note that under this approach, we used the SMOTE-created synthetic dataset to estimate the model parameters \((\alpha, \beta)\). However, the predicted \( \hat{p}_{ij} \)'s are computed for all the observations in the original dataset, which is not the same as the training data created by SMOTE (although some observations invariably belong to both the training and the test sets). Because of this, we believe that it is valuable to examine the results produced by SMOTE as well.

The first four columns of Table 4 report the odds ratios (OR) and Bayes-\( p \) for the CAR model fit on the original and SMOTE-rebalanced datasets. Under the original data set, eight covariates have a Bayes-\( p \) of at least 0.9, suggesting that these are the most informative factors associated with stillbirth. Black and Hispanic patients have a similarly higher risk of stillbirth compared to the baseline (White), while patients who experience multiple birth also have higher risk of stillbirth. Patients living in neighborhoods with higher proportions of women below the poverty level and higher proportions of women with a high school degree also have higher risk of experiencing stillbirth. On the other hand, patients living in neighborhoods with higher proportions of women receiving public assistance and higher numbers of housing violations appear to be at lower risk of stillbirth.

Several differences can be noted when the model is fit with SMOTE. Under SMOTE, the patient-level
race factor Asian is considered to be highly informative, with Asian patients having a lower risk of stillbirth compared to White patients. Among the neighborhood-level predictors, the proportion of Black inhabitants, the number of nonviolent crimes, and the proportion of women in the labor force are also considered to be fairly informative under SMOTE. The proportion of Black inhabitants and number of nonviolent crimes increase the risk of stillbirth, while the proportion of women in the labor force decreases the risk. For all the aforementioned variables, the Bayes-$p$ was greater than 0.9. Under SMOTE, the proportion of women receiving public assistance, the proportion of women with a high school degree, and the number of housing violations are no longer informative predictors for stillbirth, with low Bayes-$p$ values.

Table 4: Odds ratio (OR) and Bayes-$p$ results for stillbirth and preterm birth, computed fitting the model on the entire (Original) and the SMOTE rebalanced (SMOTE) datasets.

| Patient-level | Original OR | Bayes-$p$ | SMOTE OR | Bayes-$p$ | Original OR | Bayes-$p$ | Original OR | Bayes-$p$ |
|---------------|-------------|-----------|----------|-----------|-------------|-----------|-------------|-----------|
| age           | 1.00        | 0.63      | 1.00     | 0.68      | 1.00        | 0.84      | 1.00        | 0.84      |
| Black         | 1.86        | 1.00      | 1.79     | 1.00      | 1.68        | 1.00      | 1.69        | 1.00      |
| Hispanic      | 1.48        | 1.00      | 1.50     | 1.00      | 1.03        | 0.74      | 0.91        | 1.00      |
| Asian         | 0.72        | 0.79      | 0.39     | 1.00      | 1.18        | 1.00      | 1.27        | 1.00      |
| multiple_birth| 4.47        | 1.00      | 4.90     | 1.00      | 10.45       | 1.00      | 11.97       | 1.00      |
| Neighborhood-level | prop_Asian  | 0.51     | 0.84     | 3.49   | 0.74        | 0.89      | 0.68        | 0.43      | 1.00      |
| prop_Hispanic | 0.60        | 0.84      | 0.90     | 0.53      | 0.61        | 1.00      | 0.73        | 1.00      |
| prop_Black    | 0.94        | 0.53      | 3.14     | 1.00      | 0.85        | 1.00      | 0.88        | 0.95      |
| prop_women_15_to_50 | 0.24 | 0.79 | 2.27 | 0.58 | 3.05 | 1.00 | 1.34 | 0.74 | |
| prop_women_below_poverty | 4.02 | 1.00 | 3.04 | 0.95 | 0.78 | 0.84 | 0.92 | 0.68 | |
| prop_women_public_assistance | 0.24 | 0.95 | 0.81 | 0.58 | 1.13 | 0.58 | 0.50 | 1.00 | |
| prop_women_labor_force | 0.72 | 0.63 | 0.32 | 0.95 | 0.57 | 0.89 | 0.74 | 0.95 | |
| prop_birth_last_12_months | 5.89 | 0.89 | 0.27 | 0.79 | 2.22 | 0.95 | 0.97 | 0.53 | |
| prop_women_HS_grad | 3.18 | 1.00 | 1.89 | 0.79 | 2.38 | 1.00 | 1.73 | 1.00 | |
| prop_women_college_grad | 0.44 | 0.84 | 1.29 | 0.58 | 0.64 | 0.79 | 0.50 | 1.00 | |
| log_occupied_housing | 1.21 | 0.79 | 0.84 | 0.84 | 0.93 | 0.74 | 0.69 | 1.00 | |
| log_housing_violation | 0.83 | 1.00 | 0.99 | 0.63 | 1.07 | 0.95 | 1.12 | 1.00 | |
| log_violent_crime | 1.11 | 0.74 | 1.18 | 0.68 | 1.15 | 0.89 | 1.17 | 1.00 | |
| log_nonviolent_crime | 0.94 | 0.63 | 1.71 | 1.00 | 0.92 | 0.89 | 1.14 | 0.95 | |

Table 4: Odds ratio (OR) and Bayes-$p$ results for stillbirth and preterm birth, computed fitting the model on the entire (Original) and the SMOTE rebalanced (SMOTE) datasets.

Our results for stillbirth in Table 4 are depicted graphically in the left panel of Figure 3. We plot the posterior means of the log odds-ratios, together with the 95% credible intervals. The posterior means and credible intervals under the Original setting are shown in black and the ones for SMOTE are shown in red. Figure 3 shows generally similar estimates found under these two settings, but it also detects several differences. For example, under SMOTE, the credible intervals for several predictors (patient-level race indicator for Asian, proportion of Black inhabitants, proportion of women in the labor force, and number of nonviolent crimes) are shifted farther away from zero. Meanwhile, the posterior credible intervals for the number of housing violations and the proportions of women living below the poverty level, receiving public assistance, or who have a high school degree are shifted towards zero, suggesting that these predictors are less informative under SMOTE.

Particular caution should be used when interpreting the effects of predictors for which the
Figure 3: Posterior mean and 95% credible intervals of log odd-ratios for method fit without and with SMOTE. **Left panel**: Results for stillbirth. **Right panel**: Results for preterm birth.

Log (log) odds ratio changes significantly under SMOTE. Predictors that are more “robust” and that are informative for predicting the risk of stillbirth under both the Original and SMOTE settings are the individual-level race factors of Black and Hispanic, multiple birth, and the neighborhood proportion of women below the poverty level.

Finally, the left panel of Figure 4 visualizes the predicted risk of stillbirth for each neighborhood in our model. Census tracts shaded in blue colors are those with lower predicted risk, and they concentrate in the neighborhoods of Center City, those immediately adjacent both north (such as Fishtown, Fairmount) and south (such as Bella Vista, Southwest Center City) and in Northwest Philadelphia. Additionally, pockets of census tracts with low predicted risk can be found in West Philadelphia and Northeast Philadelphia. The census tracts with highest risk of stillbirth are identified by red shades, and pockets of such areas can be found in North Philadelphia (such as North Philadelphia West), West Philadelphia (such as Cathedral Park, West Parkside, Mantua and Kingsessing) and South Philadelphia (such as Grays Ferry and West Passyunk).

5.4 Results for preterm birth

For preterm birth, SMOTE achieves lower misclassification rate and slightly higher specificity under the LOO approach, while Original has higher AUC and sensitivity, as shown by the fifth and sixth columns of Table 3. Similar to stillbirth, SMOTE achieves the highest AUC value (0.67) under the “All data” approach, motivating us to analyze the results under both approaches for preterm birth.

In the last four columns of Table 4, we report the posterior odds ratios for preterm birth when the model was fit with the original entire dataset (Original) and when it was fit with the SMOTE-rebalanced dataset (SMOTE). Under the Original setting, nine covariates have a Bayes-p of at least 0.9. Meanwhile, under SMOTE, fifteen covariates display a Bayes-p value of at least 0.9.

Of the informative patient-level variables in the Original setting, being Asian or Black leads to
higher risk of preterm birth compared to being White, and multiple birth has the strongest effect in increasing the risk of preterm birth. Among the neighborhood-level covariates, higher proportions of Black and Hispanic inhabitants lead to lower risk of preterm birth, while higher proportions of women aged 15 to 50, women who gave birth within the last 12 months, and women who graduated high school increase the risk of preterm birth. Higher numbers of housing violations are also found to increase the risk of preterm birth.

Under SMOTE, many of the aforementioned predictors remain informative, with the exception of proportion of women aged 15 to 50 and proportion of women who gave birth within the last 12 months. On the other hand, many additional predictors are found to be informative. The SMOTE model additionally finds that living in neighborhoods with higher percentages of Asian inhabitants decreases the risk of preterm birth. Living in neighborhoods with a higher number of occupied housing units and higher proportions of women receiving public assistance, women in the labor force, and women with a college education also decreases the risk of preterm birth. Finally, both violent and non-violent crime are found to be informative in increasing the risk of preterm birth.

The right panel of Figure 3 shows that the Original and SMOTE settings give mostly similar results, although though there are several differences in the significance level of some predictors. In particular, the proportion of Asian inhabitants, the proportion of women receiving public assistance, and the number of occupied housing units have a stronger negative effect and are informative under SMOTE. On the other hand, proportion of women aged 15 to 50 and proportion of women who gave birth within the last 12 months are not found to be informative under SMOTE.

As before, we advise using caution when interpreting results for predictors whose effect changes when fitting the model with the original vs. the SMOTE-rebalanced dataset. We highlight that the more “robust” and informative predictors for preterm birth are the individual-level predictors of Asian, Black, and multiple birth, and the neighborhood-level predictors of proportion of Black and
of Hispanic inhabitants, proportion of women with a high school degree, and number of housing violations.

As for the spatial distribution of the risk of preterm birth, the right panel of Figure 4 shows areas with lower risk shaded with green tones and higher risk areas shaded with red tones. Similar to stillbirth, we found that most of the areas with low risk of preterm birth are concentrated in Center City and its surrounding neighborhoods and in Northwest Philadelphia, with some pockets in West and Northeast Philadelphia. Moreover, we also found several census tracts with high risk of preterm birth in North Philadelphia, especially in North Philadelphia West, Upper North Philadelphia and East Germantown, and in West Philadelphia, especially in Kingsessing, Cathedral Park and West Parkside.

6 Discussion

In this paper, we have introduced a Bayesian hierarchical modeling framework for geospatial analysis of adverse pregnancy outcomes. Specifically, we employed Bayesian mixed effects logistic regression models to model spatial dependence or neighborhood-level heterogeneity. We proposed using the model selection criteria DIC and WAIC to determine which model provided a better model fit. To further improve our model’s predictive power, we combined our models with SMOTE rebalancing. Finally, instead of assessing statistical significance using posterior credible intervals, we used the Bayes-$p$ measure to assess the relative informativeness of different covariates on adverse pregnancy outcomes.

We applied our methodology to a case study of preterm birth and stillbirth in Philadelphia. We found that the spatial random effects model using a CAR prior better describes the risk of stillbirth, while the independent random effects model better describes the risk of preterm birth. We also found a much larger number of neighborhood-level covariates to be informative predictors of preterm birth risk compared to stillbirth. This, together with the spatial autocorrelation of the random effects of the model for stillbirth risk, suggests that there might be unmeasured covariates that are spatially correlated and that can explain stillbirth. We also found that individual-level covariates were often very informative in explaining the risk of an adverse pregnancy outcome, with different races having different risk levels and multiple birth being among the most informative predictors of both stillbirth and preterm birth risk. Finally, we also found that the low-risk and high-risk neighborhoods were often similar for both stillbirth and preterm birth. Our results may be useful for developing targeted public health interventions in areas that are at highest risk of these outcomes [40].

In our study, we did not account for any possible temporal variation across our eight years of data. In preliminary analysis, we included a random effect for time in our model, so our model was

$$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha_i + \gamma_t + x_{ij}^T \beta,$$

where the additional subscript $t$ indexes the year that patient $j$ in neighborhood $i$ gave birth. However, we found that the effect of time was fairly negligible and that including a temporal random effect did not improve our model fit. This was likely because many of the covariates (such as the proportions of different races or educational attainment) did not vary significantly within neighborhoods across time. However, if there is reason to believe that the temporal effect is significant, we could extend the models introduced in Sections 3.1 and 3.2 to explicitly account
for temporal variation by using the model (6.1) and placing independent normal priors on the $\gamma_i$'s in (6.1).

Another option for accounting for temporal variation is to employ the semiparametric (or partially linear) mixed model framework where time is included as a covariate and its effect is modeled nonparametrically. Letting $t_{ij}$ be the year that patient $j$ in neighborhood $i$ gave birth, the partially linear model is

$$y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \alpha_i + \mathbf{x}_{ij}^\top \beta + f(t_{ij}).$$

(6.2)

In addition to placing priors on $(\alpha, \beta)$ under (6.2), we could also place a Gaussian process prior on the function $f(t) = (f(t_{11}), \ldots, f(t_{1m_1}), \ldots, f(t_{n1}), \ldots, f(t_{nm_n}))^\top$. We could also approximate $f(t_{ij})$ as a linear combination of $d$ B-spline basis functions, i.e. $f(t_{ij}) \approx \sum_{k=1}^{d} \gamma_k B_k(t_{ij})$, and place independent priors on the basis coefficients $\gamma_1, \ldots, \gamma_d$.

### Acknowledgments

This work was initiated when the first listed author was a PhD student at the Wharton School, University of Pennsylvania, under the mentorship of the fifth listed author and the second listed author was a postdoc at the Perelman School of Medicine, University of Pennsylvania, under the mentorship of the sixth and seventh listed authors. We would also like to thank Phiwinlanhla Ndebele-Ngwenya for her early work on a summer project (mentored by Dr. Boland) involving zip code level analysis that helped to motivate this expanded work.

### Funding

The first listed author was supported by the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 817257. The second listed author was partially supported by generous funding from the College of Arts and Sciences at the University of South Carolina and NSF grant OIA-1655740. The fourth and seventh listed authors were supported by generous funding from the University of Pennsylvania Perelman School of Medicine and the Penn Injury Science Center, an Injury Control Research Center funded by the Centers for Disease Control and Prevention (R49CE003083). The fifth listed author was supported by NSF Grant DMS-1916245. The sixth listed author was supported by NIH grants 1R01AI130460 and 1R01LM012607.

### Software and data availability

All methods were implemented in the statistical environment R [37], using packages [36, 13, 9, 35, 6, 46, 47, 16, 5, 38]. Due to patient privacy concerns, we do not have permission to release or make the Penn Medicine data publicly available. However, code to implement our models and simulated synthetic data to run these scripts are available at https://github.com/cecilia-balocchi/geospatial-adverse-pregnancy.
Appendices

A Additional data information

In Section 2 of the main manuscript we describe the data used for our analysis of adverse pregnancy outcomes in Philadelphia. While the patient-level data from Penn Medicine EHRs is not publicly available, the neighborhood-level data were downloaded from https://data.census.gov/cedsci/ and from https://opendataphilly.org. In Table A1 we describe the precise data source for each covariate.

| Neighborhood-level covariate                                | Source | Data File |
|-------------------------------------------------------------|--------|-----------|
| Proportion of each census tract that identifies as Asian Alone | ACS    | B01001D   |
| Proportion of each census tract that identifies as Black or African-American | ACS    | B01001B   |
| Proportion of each census tract that identifies as Hispanic or Latinx | ACS    | B01001I   |
| Proportion of each census tract that identifies as White Alone | ACS    | B01001A   |
| Proportion of women aged 15-50 years in each census tract | ACS    | S1301     |
| Proportion of women aged 15-50 years in each census tract below 100 percent poverty level | ACS    | S1301     |
| Proportion of women aged 15-50 years in each census tract that received public assistance income in the past 12 months | ACS    | S1301     |
| Proportion of women aged 16-50 years in each census tract that are in the labor force | ACS    | S1301     |
| Proportion of women aged 15-50 years in each census tract who had a birth in the past 12 months | ACS    | B13016    |
| Proportion of women aged 15-50 years in each census tract that graduated high school (including equivalency) | ACS    | S1301     |
| Proportion of women aged 15-50 years in each census tract that have a Bachelor’s degree | ACS    | S1301     |
| Number of occupied housing units in each census tract | ACS    | S2502     |
| Housing Violations | OpenDataPhilly |          |
| Violent Crime Rate | OpenDataPhilly |          |
| Non-Violent Crime Rate | OpenDataPhilly |          |

Table A1: Sources and data files for neighborhood-level covariates used in our analysis. ACS refers to the American Community Survey.

B Computational details

To fit our model, we sample from the posterior distribution with MCMC. We use a Gibbs sampler to iteratively sample from the full conditional posteriors, together with the data augmentation strategy using Pólya-Gamma latent variables, proposed by Polson et al. [36]. This strategy is the equivalent in logistic regression to the data augmentation strategy using latent Gaussian random variables in probit regression [1].

Using the notation from Polson et al. [36], we say that a random variable $X$ is distributed from a Pólya-Gamma with parameters $b > 0$ and $c \in \mathcal{R}$ if

$$X \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}$$
where \( g_k \sim \text{Ga}(b, 1) \), and we denote \( X \sim \text{PG}(b, c) \).

As shown in Polson et al. [36], if \( y_i \sim \text{Bernoulli}(p_i) \) and \( \log(\frac{p_i}{1-p_i}) = x_i^\top \beta \) with \( \beta \sim \mathcal{N}(b, B) \), we can sample

\[
\omega_i | \beta \sim \text{PG}(1, x_i^\top \beta) \\
\beta | y_i, \omega \sim \mathcal{N}(m_\omega, V_\omega)
\]

with \( V_\omega = (X^\top \Omega X + B^{-1})^{-1} \) and \( m_\omega = V_\omega (X^\top \Omega \tilde{y} + B^{-1}b) \), where \( \Omega = \text{diag}(\omega_i) \) and \( \tilde{y}_i = \frac{y_i - \frac{1}{\omega_i}}{\omega_i} \).

To adapt this sampling scheme to our model, we can write \( \alpha_i + x_i^\top \beta = z_{ij} \alpha + x_{i,j}^\top \beta \), where \( z_{ij} \) is an \( n \)-dimensional vector with all entries equal to 0, except for the \( j \)-th entry which is equal to 1. Then the full conditional for our model can be written as

\[
\omega_{ij} | \alpha, \beta \sim \text{PG}(1, \alpha_i + x_{i,j}^\top \beta) \\
\alpha | y_i, \omega, \tau_\alpha \sim \mathcal{N} \left( V_\alpha \left( Z^\top \Omega (\tilde{y} - X \beta) + \tau_\alpha^{-2} \Sigma_\alpha^{-1} 1_{\alpha_0} \right), V_\alpha^{-1} \right) \\
\beta | y_i, \alpha, \omega, \tau_\beta \sim \mathcal{N} \left( V_\beta \left( X^\top \Omega (\tilde{y} - Z \alpha) + \tau_\beta^{-2} b_0 \right), V_\beta^{-1} \right)
\]

where \( Z \) is the \( N \times n \) matrix of \( z_{ij} \), \( V_\alpha = Z^\top \Omega Z + \tau_\alpha^{-2} \Sigma_\alpha^{-1} \), \( V_\beta = X^\top \Omega X + \tau_\beta^{-2} I \). Moreover, \( \Sigma_\alpha^{-1} \) denotes the prior covariance matrix of \( \alpha \), which is equal to \( \Sigma_{\text{CAR}}^{-1} \) under model (3.4) and to \( I_n \) under model (3.11).

Under the CAR prior for \( \alpha \), we also need to sample the correlation parameter \( \rho \). Note that this parameter affects the precision matrix of \( \alpha \), and its conditional distribution is given by

\[
p(\rho | e.c.) = p(\rho | \alpha, \alpha_0, \tau_\alpha) \propto | \Sigma_\alpha^{-1}(\rho) |^{1/2} \exp \left( -\frac{1}{2\tau_\alpha^2} (\alpha - \alpha_0) \Sigma_\alpha^{-1}(\rho) (\alpha - \alpha_0) \right).
\]

We sample from this distribution using a Metropolis Hastings-within Gibbs-step, with proposal density \( g(\rho^* | \rho_i) = \text{Beta}(\xi \cdot \rho_i / (1 - \rho_i), \xi) \). This ensures that the mean is equal to \( \rho_i \), and we choose \( \xi = 5 \) so that \( g(\rho^* | \rho_i) \) has small variance.

Moreover, the full conditional for the remaining parameters are standard:

\[
\alpha_0 | \alpha, \tau_\alpha \sim \mathcal{N} \left( \frac{\sum_{i=1}^n \alpha_i / \tau_\alpha^2}{1 / \tau_\alpha^2 + 1 / 100}, \frac{1}{1 / \tau_\alpha^2 + 1 / 100} \right) \\
b_0 | \beta, \tau_\beta \sim \mathcal{N} \left( \frac{\beta / \tau_\beta^2}{1 / \tau_\beta^2 + 1 / 100}, \frac{1}{1 / \tau_\beta^2 + 1 / 100} \right) \\
\tau_\alpha^2 | \alpha_0 \sim I_\mathcal{G} \left( 0.1 + (n - 1)/2, 0.1 + (\alpha - \alpha_0)^\top \Sigma_\alpha^{-1}(\alpha - \alpha_0) \right) \\
\tau_\beta^2 | \beta, b_0 \sim I_\mathcal{G} \left( 0.1 + p/2, 0.1 + \sum_{k=1}^p (\beta_k - b_{0,k})^2 / 2 \right).
\]

### C Additional results

In Section 5 of the main manuscript we describe the results for our analysis of adverse pregnancy outcomes, in particular reporting the posterior odds ratio for each individual-level and neighborhood-level covariate. Moreover, we report the predicted probabilities of an adverse outcome for each
Table A2: Comparison Odds ratio and Relative Odds ratio for pairs of neighborhoods that differ for a specific neighborhood level characteristic: Prop_Black for neighborhoods B1 and B2, prop_women_labor_force for neighborhoods L1 and L2, prop_women_college_grad for neighborhoods C1 and C2 and log_violent_crime for neighborhoods V1 and V2.

neighborhood in Figure 4. However, to better compare how the collective set of covariates affects the neighborhood level risk, in Section C.1 we report a comparison of several pairs of neighborhoods which differ for one or more covariates. Moreover, in Section C.2 we report the maps for the spatial distribution of the neighborhood-level random effects. Finally, in Section C.3 we report the MCMC trace plots.

C.1 Neighborhoods risk comparison

Table A2 compares several neighborhoods, displaying their covariates and their risk of an adverse pregnancy outcome. In fact, it is important not to consider the predictors in isolation (as they are often correlated), but to analyze the differences between two neighborhoods for all their covariates jointly.

So far we have presented results in terms of odds ratio for each predictor, but this can be misleading since the covariates cannot be artificially changed keeping the others fixed. Thus, we now visualize how covariates jointly change by considering some representative pairs of neighborhoods.

For a given predictor of interest, consider a pair of neighborhoods that display opposite values for such predictors. One can then visualize how the other covariates change and how all of them affect the change in risk of negative outcome.

For example, Table A2 displays the covariates of four pairs of neighborhoods, chosen to maximize the discrepancy in the proportion of Black inhabitants (B1 and B2), in the proportion of women
in the labor force (L1 and L2), in the proportion of women who graduated from college (C1 and C2) and in the level of violent crime (V1 and V2). It is interesting to notice that the pairs of neighborhoods do not differ only in the level of the chosen predictor: for example B1 and B2 have very different proportions of women who are college graduates, L1 and L2 differ also in the proportion of women below the poverty line, and V1 and V2 have very different levels of housing violations. The last four rows of the table compare the predicted risk for each pair of neighborhoods, for stillbirth and preterm birth outcomes, together with the fraction of the two odds-ratio.

C.2 Spatial distribution of random effects

In Figure A1, we plot the posterior means of the neighborhood random effects $\tilde{\alpha}_i$’s for the 363 census tracts in our models. Figure 4 of the main manuscript is a bit easier to interpret than Figure A1, since Figure 4 depicts our models’ predicted neighborhood risk probabilities of adverse pregnancy outcomes. However, we include Figure A1 here for the sake of completeness.

![Maps for neighborhood random effects](image)

Figure A1: Maps for neighborhood random effects, computed from the non-SMOTE models, using the whole dataset. **Left panel**: Results for stillbirth. **Right panel**: Results for preterm birth.

C.3 MCMC diagnostics

Figure A2 shows the trace plots and autocorrelation plots of the MCMC samples for several of the regression coefficients in our Bayesian mixed effect logistic regression models. Figure A2 shows that the MCMC algorithm described in Section B converged within 5500 iterations for the data that we analyzed in Section 5 of the main manuscript. Moreover, by thinning the chain every 10 iterations, our final MCMC samples were not overly correlated.
Figure A2: Trace plots of the two MCMC chains and autocorrelation plots of the thinned samples, for several coefficients in our models: in the first two rows we report results for preterm birth, fit with the non-SMOTE model with independent random effects, while the last two rows are for stillbirth with non-SMOTE CAR model.
References

[1] Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88(422), 669–679.

[2] Balocchi, C. and S. T. Jensen (2019). Spatial modeling of trends in crime over time in Philadelphia. *The Annals of Applied Statistics* 13(4), 2235–2259.

[3] Banerjee, S., B. Carlin, and A. Gelfand (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

[4] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 36(2), 192–225.

[5] Bivand, R. and N. Lewin-Koh (2019). *maptools: Tools for Handling Spatial Objects*. R package version 0.9-9.

[6] Bivand, R. and D. W. S. Wong (2018). Comparing implementations of global and local indicators of spatial association. *TEST* 27(3), 716–748.

[7] Canelón, S. P., H. H. Burris, L. D. Levine, and M. R. Boland (2021). Development and evaluation of MADDIE: Method to acquire delivery date information from electronic health records. *International Journal of Medical Informatics* 145, 104339.

[8] Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16(1), 321–357.

[9] Dahl, D. B., D. Scott, C. Roosen, A. Magnusson, and J. Swinton (2019). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-4.

[10] Dekker, G. A., S. Y. Lee, R. A. North, L. M. McCowan, N. A. B. Simpson, and C. T. Roberts (2012). Risk factors for preterm birth in an international prospective cohort of nulliparous women. *PLOS ONE* 7(7), 1–9.

[11] Ferré, C., W. Callaghan, C. Olson, A. Sharma, and W. Barfield (2016). Effects of maternal age and age-specific preterm birth rates on overall preterm birth rates — United States, 2007 and 2014. *MMWR Morbidity and Mortality Weekly Report* 65(43), 1181–1184.

[12] Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing* 24(6), 997–1016.

[13] Genz, A., F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn (2020). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-0.

[14] Goldenberg, R. L., J. F. Culhane, J. D. Iams, and R. Romero (2008). Epidemiology and causes of preterm birth. *The Lancet* 371(9606), 75–84.

[15] Gray, R., S. Bonellie, J. Chalmers, I. Greer, S. Jarvis, and C. Williams (2008). Social inequalities in preterm birth in Scotland 1980-2003: findings from an area-based measure of deprivation. *BJOG* 115(1), 82–90.
[16] Kahle, D. and H. Wickham (2013). ggmap: Spatial visualization with ggplot2. *The R Journal 5*(1), 144–161.

[17] Kramer, M. R., H. L. Cooper, C. D. Drews-Botsch, L. A. Waller, and C. R. Hogue (2010). Metropolitan isolation segregation and black–white disparities in very preterm birth: A test of mediating pathways and variance explained. *Social Science and Medicine 71*(12), 2108 – 2116.

[18] Kramer, M. R. and C. R. Hogue (2008). Place matters: Variation in the black/white very preterm birth rate across u.s. metropolitan areas, 2002–2004. *Public Health Reports 123*(5), 576–585. PMID: 18828412.

[19] Kramer, M. S. (2003). The Epidemiology of Adverse Pregnancy Outcomes: An Overview. *The Journal of Nutrition 133*(5), 1592S–1596S.

[20] Lee, D. (2011). A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology 2*(2), 79–89.

[21] Leroux, B. G., X. Lei, and N. Breslow (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, New York, NY, pp. 179–191. Springer New York.

[22] Lhila, A. (2009). Does government provision of healthcare explain the relationship between income inequality and low birthweight? *Social Science and Medicine 69*(8), 1236 – 1245.

[23] Li, H., C. A. Calder, and N. Cressie (2007). Beyond moran’s i: Testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis 39*(4), 357–375.

[24] Macdorman, M. F. and E. C. W. Gregory (2015). Fetal and perinatal mortality: United states, 2013. *National Vital Statistics Reports 64*(8), 1–24.

[25] Makowski, D., M. S. Ben-Shachar, S. Chen, and D. Lüdecke (2019). Indices of effect existence and significance in the bayesian framework. *Frontiers in Psychology 10*, 2767.

[26] Mandujano, A., T. P. Waters, and S. A. Myers (2013). The risk of fetal death: current concepts of best gestational age for delivery. *American Journal of Obstetrics and Gynecology 208*(3), 207.e1 – 207.e8.

[27] Martin, J., B. Hamilton, M. Osterman, A. Driscoll, P. Drake, and National Center for Health Statistics (U.S.) (2018). Births: Final data for 2017. *National Vital Statistics Reports 67*(8).

[28] Mattison, D. R., K. Damus, E. Fiore, J. Petrini, and C. Alter (2001). Preterm delivery: a public health perspective. *Paediatric and Perinatal Epidemiology 15*(s2), 7–16.

[29] Mayne, S. L., L. R. Pool, W. A. Grobman, and K. N. Kershaw (2018). Associations of neighbourhood crime with adverse pregnancy outcomes among women in chicago: analysis of electronic health records from 2009 to 2013. *Journal of Epidemiology and Community Health 72*(3), 230–236.
[30] Meeker, J. R., S. P. Canelón, R. Bai, L. Levine, and M. R. Boland (2021). Individual-level and neighborhood-level risk factors for severe maternal morbidity. *Obstetrics & Gynecology*, to appear.

[31] Meghea, C. I. and W. Corser (2016). Electronic medical record use and maternal and child care and health. *Maternal and Child Health Journal 20*, 819–826.

[32] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*(1/2), 17–23.

[33] Muraskas, J. and K. Parsi (2008). Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *AMA Journal of Ethics 10*(10), 655–658.

[34] Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (2004). *Applied Linear Statistical Models (4th ed.)*. McGraw-Hill Irwin.

[35] Pebesma, E. J. and R. S. Bivand (2005, November). Classes and methods for spatial data in R. *R News 5*(2), 9–13.

[36] Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association 108*(504), 1339–1349.

[37] R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

[38] Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics 12*, 77.

[39] Shapiro-Mendoza, C. K., W. D. Barfield, Z. Henderson, A. James, J. L. Howse, J. Iskander, and P. G. Thorpe (2016). Cdc grand rounds: Public health strategies to prevent preterm birthcdc grand rounds: Public health strategies to prevent preterm birth. *MMWR Morbidity and Mortality Weekly Report 65*(32), 826–830.

[40] South, A. P., D. E. Jones, E. S. Hall, S. Huo, J. Meinzen-Derr, L. Liu, and J. M. Greenberg (2012). Spatial analysis of preterm birth demonstrates opportunities for targeted intervention. *Maternal and Child Health Journal 16*, 470–478.

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

[42] Starling, J. E., J. S. Murray, C. M. Carvalho, R. K. Bukowski, and J. G. Scott (2020). Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *The Annals of Applied Statistics 14*(1), 28–50.

[43] Torgo, L. (2013). *DMWR: Functions and data for “Data Mining with R”*. R package version 0.4.1.
[44] Warren, J., M. Fuentes, A. Herring, and P. Langlois (2012). Spatial-temporal modeling of the association between air pollution exposure and preterm birth: Identifying critical windows of exposure. *Biometrics* 68(4), 1157–1167.

[45] Watanabe, S. and M. Opper (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11(12), 3571–3594.

[46] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[47] Wickham, H., R. François, L. Henry, and K. Müller (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.5.

[48] Xu, J., S. L. Murphy, K. D. Kochanek, and B. A. Bastian (2016). Deaths: Final data for 2013. *National Vital Statistics Reports* 10(10), 1–119.

[49] Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer* 3(1), 32–35.

[50] Zahrieh, D., J. J. Oleson, and P. A. Romitti (2019). Bayesian point process modeling to quantify geographic regions of excess stillbirth risk. *Geographical Analysis* 51(3), 381–400.

[51] Zhao, Y., Z. S.-Y. Wong, and K. L. Tsui (2018). A framework of rebalancing imbalanced healthcare data for rare events’ classification: A case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering 2018*, Article ID 6275435.