Check for updates

Received: 8 July 2021 | Accepted: 8 August 2022

DOI: 10.1111/biom.13739

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Delivering spatially comparable inference on the risks of multiple severities of respiratory disease from spatially misaligned disease count data

Duncan Lee | **Craig Anderson**

School of Mathematics and Statistics, University of Glasgow, Glasgow, Scotland

Correspondence
Duncan Lee, School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8SQ, Scotland.
Email: Duncan.Lee@glasgow.ac.uk

Abstract

Population-level disease risk varies between communities, and public health professionals are interested in mapping this spatial variation to monitor the locations of high-risk areas and the magnitudes of health inequalities. Almost all of these risk maps relate to a single severity of disease outcome, such as hospitalization, which thus ignores any cases of disease of a different severity, such as a mild case treated in a primary care setting. These spatially-varying risk maps are estimated from spatially aggregated disease count data, but the set of areal units to which these disease counts relate often varies by severity. Thus, the statistical challenge is to provide spatially comparable inference from multiple sets of spatially misaligned disease count data, and an additional complexity is that the spatial extents of the areal units for some severities are partially unknown. This paper thus proposes a novel spatial realignment approach for multivariate misaligned count data, and applies it to the first study delivering spatially comparable inference for multiple severities of the same disease. Inference is via a novel spatially smoothed data augmented MCMC algorithm, and the methods are motivated by a new study of respiratory disease risk in Scotland in 2017.

KEYWORDS

data augmentation, multiseverity disease risk modeling, multivariate conditional autoregressive models, spatial misalignment

1 | INTRODUCTION

Spatially aggregated disease count data are commonly used to monitor public health, including the identification of high-risk areas and the quantification of health inequalities. The World Health Organization defines *total inequality* as the overall variation in disease risk, and *social inequality* as the variation between different social groups. Existing studies have modeled these data for a single severity of disease outcome, with Lee (2018) focusing on mild cases treated in primary care, whereas Zhu et al. (2003) model more severe cases requiring hospitalization. However, this ubiquitous single-severity approach leads to an underestimation of the health burden of disease, because it ignores all cases of a different severity to that being modeled. Additionally, it does not allow us to examine how

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Biometrics* published by Wiley Periodicals LLC on behalf of International Biometric Society.

disease risk varies by severity, which is the novel insight provided by this paper.

Our motivating study focuses on respiratory disease in Scotland, and we have data on mild cases treated in primary care, moderately severe cases treated in hospital, and very severe cases resulting in death. Count data for the latter two are available for a set of known nonoverlapping areal units, whereas the primary care data have been spatially aggregated to doctors' surgeries whose catchment areas are partially unknown. Thus, the statistical challenge we address is the *change of support problem* (Gelfand et al., 2001), because we need to provide comparable inference for all disease severities on a common spatial scale.

Numerous methodologies have been proposed for overcoming spatial misalignment, including downscaling (Song et al., 2014), change of support (Nethery et al., 2019) and regression (Zhu et al., 2003) settings, and designing areal units that minimize spatial aggregation error (Bradley et al., 2017). For count data, Flowerdew and Green (1989) and Mugglin and Carlin (1998) focused on areal interpolation and Bradley et al. (2016) introduced change of support methodology, whereas Li et al. (2012) and Taylor et al. (2018) produce pseudospatially continuous inference from spatially aggregated data.

The methodological challenge addressed here extends this work to a multivariate setting, because we propose the first spatial realignment model for multivariate misaligned count data, some of which have partially unknown spatial supports. Our Bayesian hierarchical model delivers spatially comparable inference for all disease severities, and captures both spatial and between disease severity correlations via a latent Gaussian process modelled with a multivariate conditional autoregressive (MCAR, Gelfand and Vounatsou, 2003) prior distribution. Inference is based on a novel spatially smoothed data augmented (Tanner and Wong, 1987) Markov chain Monte Carlo (MCMC) algorithm, which extends the algorithm proposed by Taylor et al. (2018) that in our data context, suffers from computational convergence issues. The motivating study is outlined in Section 2, whereas our proposed methodology is presented in Section 3 and tested by simulation in Section 4. The study results are presented in Section 5, whereas Section 6 concludes the paper.

2 | SCOTLAND RESPIRATORY DISEASE STUDY

Our methodology is motivated by a new study of respiratory disease in mainland Scotland in 2017, which includes cases that are: (i) relatively minor and treated in primary care; (ii) moderately severe and require hospitalization; and (iii) severe and result in death. Our modeling aims are to: (a) identify the extent to which the highest risk areas differ by severity; and (b) quantify how the magnitude of health inequalities vary by severity?

2.1 | Disease data

Mild cases of disease are treated in primary care by doctors grouped within 869 general practice (GP) surgeries, and we obtained the total number of short-acting β_2 agonists prescribed by each GP surgery. These medications are used to relieve the symptoms of respiratory disease such as asthma and chronic obstructive pulmonary disease (COPD), and the medications included are listed and justified in Web Appendix A. Following Blangiardo et al. (2016), these data are used as a proxy measure of respiratory disease not requiring hospitalization, because complete incidence data are unavailable. Moderately severe cases of respiratory disease are represented by the numbers of admissions to hospital (ICD-10 codes J00-J99), and these data are spatially aggregated counts for the populations living in each of the 1252 intermediate zones (IZs) that comprise mainland Scotland. IZs are a Scottish Government developed small-area geography, and have an average population of around 4000. Finally, we have data relating to respiratory deaths (again ICD-10 codes J00-J99) for the same set of IZs, and a summary of these count data is provided in Web Appendix A.

We account for the differing population sizes and demographics across the spatial units using indirect standardization, which computes the expected numbers of disease events in each unit based on national age-sex-specific disease incidence rates. These expected counts are computed exactly for the hospitalization and death outcomes, because their national age-sex-specific rates are available from Public Health Scotland. In contrast, national rates of respiratory prescribing for different age-sex strata are not publicly available, so we estimate these unknown rates using data on the national age-sex-specific rates of asthma and COPD.

An exploratory measure of disease risk is the standardized incidence ratio (SIR), which is the observed numbers of disease cases/prescriptions divided by their expected numbers. Maps of the spatial distribution of the SIRs for each disease severity are displayed in Figure 1, where an SIR of 1.2 corresponds to a 20% increased risk compared to the national average. In the GP prescription map, each surgery is shown as a dot at the surgery location, because the residential locations and geographical extent of its patient population are partially unknown. The figures show that for all severities, the highest SIR values are in and around the city of Glasgow, which is well known to have some of the poorest health in western Europe.

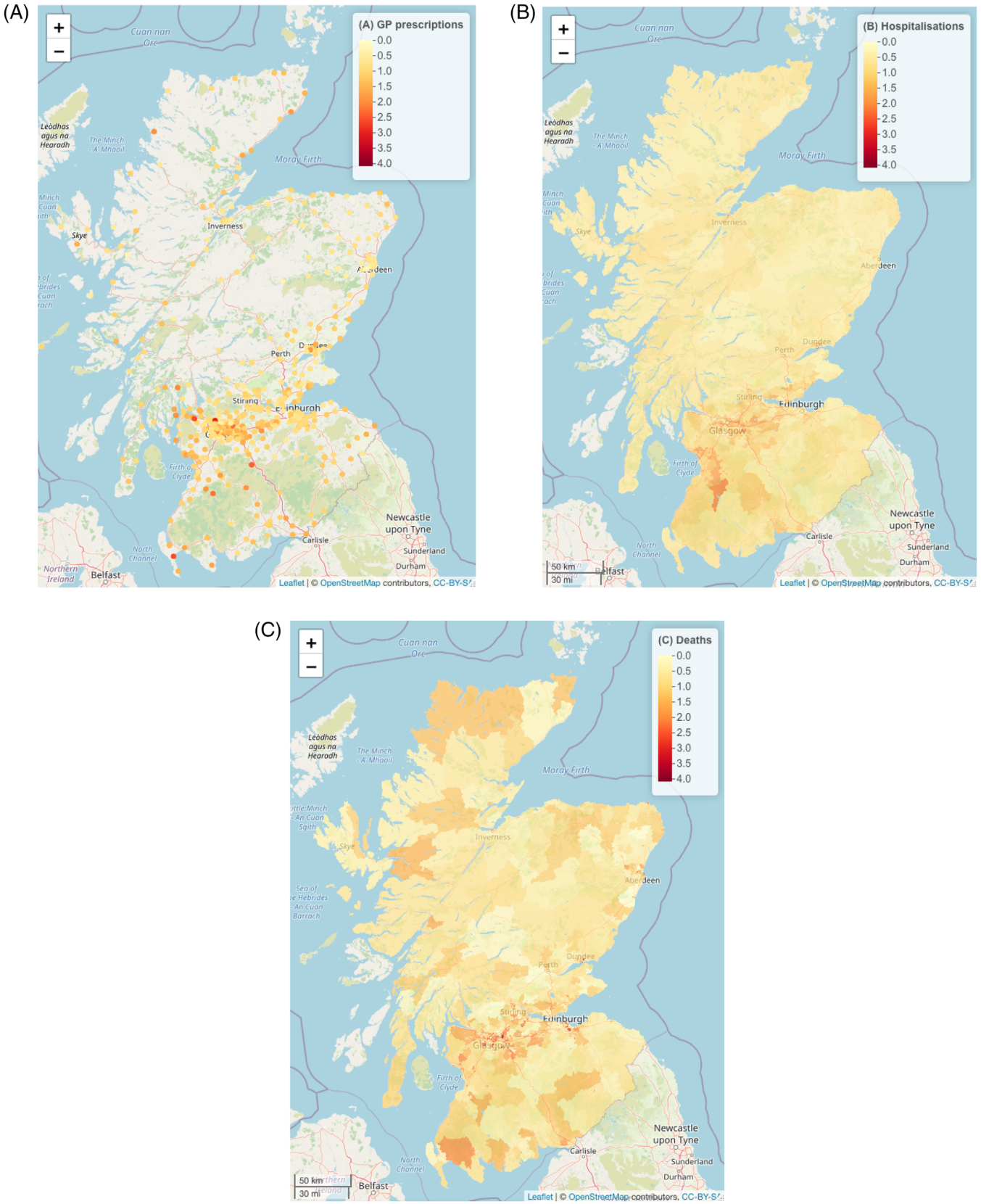

FIGURE 1 Maps displaying the SIR for: (A) GP prescriptions, (B) hospitalizations, and (C) deaths. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

2.2 | A common spatial inferential scale

Taylor et al. (2018) make inference on a regular grid when modeling data relating to known areal units, and base their spatial realignment on the area of intersection between each areal unit and grid square. However, the patient catchment area for each GP surgery is partially unknown, which makes it difficult to spatially allocate prescription counts based on areas of intersection. The only data available quantifying the spatial locations of the GP surgery practice populations are how they are distributed across IZs, which is the spatial scale at which we have hospitalization and death data. Specifically, we have data on how many people registered at each GP surgery live in each IZ, and a summary of these population intersections is provided in Web Appendix A.

Therefore, as we have disease data for two of the three severities at the IZ level and indirect data relating the GP surgery patient populations to the IZ scale, we use IZs as our common inferential scale. We note that while we know how many people registered with each GP surgery live in each IZ, we do not know which of these individuals obtained prescriptions for respiratory disease. Hence, we cannot directly compute the number of prescriptions at the IZ scale, which necessitates the spatial realignment methodology that we propose. Spatial misalignment such as this can occur routinely when jointly modeling spatially aggregated data from different sources, such as when health care is provided by different providers.

2.3 | Socioeconomic deprivation data

We quantify the social inequality in disease risk by estimating the effect of socioeconomic deprivation on each severity of disease. Socioeconomic deprivation is measured in Scotland by the Scottish Index of Multiple Deprivation (SIMD), which comprises indicators in the domains of geographical access to services, crime, education, employment, health, housing, and income. However, as our outcome variables are health related, we ignore indicators from the health domain. The remaining 16 indicators have pairwise correlations ranging between -0.81 and 0.98. Thus, we create one measure of socioeconomic deprivation for each domain of the SIMD using principal component analysis, and details are given in Web Appendix A.

3 | METHODOLOGY

3.1 | Data description and spatial scales

The observed and expected GP prescription counts for the study region $\mathcal{S}$ are, respectively, denoted by $\mathbf{Y}_1(\mathcal{S}) = \{Y_1(\mathcal{A}_1), \ldots, Y_1(\mathcal{A}_K)\}$ and $\mathbf{e}_1(\mathcal{S}) = \{e_1(\mathcal{A}_1), \ldots, e_1(\mathcal{A}_K)\}$, where $\mathcal{S} = \{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ denotes the spatial supports of the $K = 869$ GP surgery patient populations. In contrast, the hospitalization and death data are available at $M = 1252$ IZs denoted as $\mathcal{S} = \{\mathcal{H}_1, \ldots, \mathcal{H}_M\}$, where the observed and expected hospitalization counts are denoted by $\mathbf{Y}_2(\mathcal{S}) = \{Y_2(\mathcal{H}_1), \ldots, Y_2(\mathcal{H}_M)\}$ and $\mathbf{e}_2(\mathcal{S}) = \{e_2(\mathcal{H}_1), \ldots, e_2(\mathcal{H}_M)\}$, whereas the corresponding death counts are denoted by $\mathbf{Y}_3(\mathcal{S}) = \{Y_3(\mathcal{H}_1), \ldots, Y_3(\mathcal{H}_M)\}$ and $\mathbf{e}_3(\mathcal{S}) = \{e_3(\mathcal{H}_1), \ldots, e_3(\mathcal{H}_M)\}$. The SIMD indicators are available at the IZ scale, and the bth indicator is denoted by $x_b(\mathcal{H}_i)$. Our methodology provides spatially comparable inference on all three severities of disease at the IZ scale, and to achieve this, we use population intersection data $\{P(\mathcal{A}_k \cap \mathcal{H}_i)\}$, where $P(\mathcal{A}_k \cap \mathcal{H}_i)$ denotes the number of people who live in the ith IZ and are registered with the kth GP surgery. Thus, the population totals for each GP surgery and IZ are $P(\mathcal{A}_k) = \sum_{i=1}^{M} P(\mathcal{A}_k \cap \mathcal{H}_i)$ and $P(\mathcal{H}_i) = \sum_{k=1}^{K} P(\mathcal{A}_k \cap \mathcal{H}_i)$, respectively.

3.2 | Risk model

The first step is to estimate the expected numbers of GP prescriptions $\mathbf{e}_1^{\mathcal{H}}(\mathcal{S}) = \{e_1(\mathcal{H}_1), \ldots, e_1(\mathcal{H}_M)\}$ at the IZ level using population-weighted interpolation similar to Flowerdew and Green (1993), which is described in Web Appendix B. In contrast, the observed numbers of GP prescriptions at the IZ scale, $\{Y_1(\mathcal{H}_i)\}$, are estimated within the inferential algorithm described below. At the IZ level, we model the multivariate disease counts using a Poisson log-linear structure, whose mean depends on covariates and a latent Gaussian process. The latter is commonly used in the spatial modeling literature (Gelfand and Schliep, 2016), but alternatives include the multivariate log-gamma distribution (Bradley et al., 2018) and a Pólya-Gamma augmentation (Bansal et al., 2021). The data likelihood model is given by

$$\begin{aligned} Y_j(\mathcal{H}_i) &\sim \text{Poisson}\{e_j(\mathcal{H}_i)\theta_j(\mathcal{H}_i)\} \quad \text{for } i = 1, \ldots, M \text{ and } j = 1, \ldots, 3 \\ \ln\{\theta_j(\mathcal{H}_i)\} &= \mathbf{x}(\mathcal{H}_i)^\top \boldsymbol{\beta}_j + \phi_j(\mathcal{H}_i), \end{aligned} \tag{1}$$

where $\mathbf{x}(\mathcal{H}_i)$ denotes a vector of known covariates. The regression parameters $\boldsymbol{\beta}_j$ differ by disease severity, and each parameter is assigned a weakly informative-independent Gaussian prior with mean zero and variance 100,000 to allow the data to speak for themselves.

The random effects $\boldsymbol{\phi} = \{\boldsymbol{\phi}(\mathcal{H}_1), \ldots, \boldsymbol{\phi}(\mathcal{H}_M)\}_{3M\times 1}$, where $\boldsymbol{\phi}(\mathcal{H}_i) = \{\phi_1(\mathcal{H}_i), \ldots, \phi_3(\mathcal{H}_i)\}$, are modeled with an MCAR prior (Gelfand and Vounatsou, 2003), which captures both spatial and between outcome correlations. MCAR priors are based on a binary $M \times M$ spatial neighbor-

hood matrix $\mathbf{W} = (w_{ir})$, and here $w_{ir} = 1$ if IZs $(\mathcal{H}_i, \mathcal{H}_r)$ share a common border and $w_{ir} = 0$ otherwise (with $w_{ii} = 0 \ \forall \ i$). The MCAR prior we use has joint distribution $\boldsymbol{\phi} \sim \text{N}[\mathbf{0}, \{\mathbf{Q}(\mathbf{W}, \rho) \otimes \boldsymbol{\Upsilon}^{-1}\}^{-1}]$, where $\boldsymbol{\Upsilon}_{3\times3}$ is the between severity conditional covariance matrix and $\mathbf{Q}(\mathbf{W}, \rho)_{M\times M} = \rho\{\text{diag}(\mathbf{W1}) - \mathbf{W}\} + (1-\rho)\mathbf{I}$ is the spatial precision matrix corresponding to the conditional autoregressive prior proposed by Leroux et al. (2000). This prior is used rather than the BYM prior (Besag et al., 1991) because it has an explicit spatial dependence parameter ρ that one can interpret, as well as only having a single random effect for each areal unit (the BYM model has two random effects for each areal unit). The spatial dependence implied by the MCAR model can be seen from its full conditional form for $\boldsymbol{\phi}(\mathcal{H}_i)|\boldsymbol{\phi}(-\mathcal{H}_i)$ (where $\boldsymbol{\phi}(-\mathcal{H}_i) = \boldsymbol{\phi} \setminus \boldsymbol{\phi}(\mathcal{H}_i)$) given by

$$\boldsymbol{\phi}(\mathcal{H}_i)|\boldsymbol{\phi}(-\mathcal{H}_i), \mathbf{W}, \boldsymbol{\Upsilon}, \rho \sim \text{N}\left(\frac{\rho\sum_{r=1}^{M} w_{ir}\boldsymbol{\phi}(\mathcal{H}_r)}{\rho\sum_{r=1}^{M} w_{ir} + 1 - \rho}, \frac{1}{\rho\sum_{r=1}^{M} w_{ir} + 1 - \rho}\boldsymbol{\Upsilon}\right). \tag{2}$$

Here, ρ is a spatial dependence parameter, with $\rho = 1$ corresponding to strong spatial dependence (the conditional expectation of $\boldsymbol{\phi}(\mathcal{H}_i)$ is the mean of the random effects in neighboring IZs), while if $\rho = 0$, then $(\boldsymbol{\phi}(\mathcal{H}_i), \boldsymbol{\phi}(\mathcal{H}_r))$ are independent. We specify a noninformative uniform prior for ρ on the unit interval, that is, $\rho \sim \text{Uniform}(0, 1)$, and a weakly informative Inverse-Wishart$(4, \mathbf{I})$ prior for the cross-severity covariance matrix $\boldsymbol{\Upsilon}$.

3.3 | Inference

We fit the model using a spatially smoothed data-augmented MCMC algorithm, which jointly updates the IZ-level model parameters $\boldsymbol{\Omega} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_3, \boldsymbol{\phi}, \boldsymbol{\Upsilon}, \rho\}$ and the IZ-level GP prescription counts $\mathbf{Y}_1^{\mathcal{H}}(\mathcal{S}) = \{Y_1(\mathcal{H}_1), \ldots, Y_1(\mathcal{H}_M)\}$ conditional on the disease and covariate data $\mathcal{D}$ and the population data $\mathcal{P}$. Our algorithm extends the proposal of Taylor et al. (2018) that is summarized in Web Appendix B, and iterates the following steps.

(1) Sample from $f\{\boldsymbol{\Omega}|\mathbf{Y}_1^{\mathcal{H}}(\mathcal{S}), \mathcal{D}, \mathcal{P}\}$, the conditional distribution of the model parameters $\boldsymbol{\Omega}$ given the current values of the IZ-level prescription counts $\mathbf{Y}_1^{\mathcal{H}}(\mathcal{S})$ and the data $\{\mathcal{D}, \mathcal{P}\}$.
(2) Sample from $f\{\mathbf{Y}_1^{\mathcal{H}}(\mathcal{S})|\boldsymbol{\Omega}, \mathcal{D}, \mathcal{P}\}$, the conditional distribution of the IZ-level prescription counts $\mathbf{Y}_1^{\mathcal{H}}(\mathcal{S})$ given the model parameters $\boldsymbol{\Omega}$ and the data $\{\mathcal{D}, \mathcal{P}\}$.

Step (1) samples from the posterior distribution of models (1)–(2), which is achieved using Gibbs sampling (for $\boldsymbol{\Upsilon}$) and Metropolis–Hastings (for $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J, \boldsymbol{\phi}, \rho\}$) steps. Step (2) is the data augmentation step, and we undertake it at every Cth iteration of the MCMC algorithm because this lessens the dependence between successive values of $\boldsymbol{\Omega}$ used in this step. As described in Web Appendix B, we sample from $f\{\mathbf{Y}_1^{\mathcal{H}}(\mathcal{S})|\boldsymbol{\Omega}, \mathcal{D}, \mathcal{P}\}$ using K multinomial sampling steps, because each GP surgery's total number of prescriptions $Y_1(\mathcal{A}_k)$ has to be allocated to the M IZs in an integer fashion. Letting $Y_1(\mathcal{A}_k \cap \mathcal{H}_i)$ denote the number of prescriptions allocated to IZ $\mathcal{H}_i$ from the patients registered to GP surgery $\mathcal{A}_k$, step (2) involves sampling from

$$\begin{aligned}&\{Y_1(\mathcal{A}_k \cap \mathcal{H}_1), \ldots, Y_1(\mathcal{A}_k \cap \mathcal{H}_M)\} \sim \text{Multinomial}\\ &\{n = Y_1(\mathcal{A}_k)|\pi_{k1}, \ldots, \pi_{kM}\}.\end{aligned} \tag{3}$$

The IZ-level prescription counts are then computed as $Y_1(\mathcal{H}_i) = \sum_{k=1}^{K} Y_1(\mathcal{A}_k \cap \mathcal{H}_i)$. The multinomial weights in (3) are modeled on the unnormalized scale $\{\pi_{ki}^*\}$, where $\pi_{ki} = \pi_{ki}^* / \sum_{r=1}^{M} \pi_{kr}^*$ so that $\sum_{i=1}^{M} \pi_{ki} = 1$. A natural model for these unnormalized weights π_{ki}^* is

$$\pi_{ki}^* = P(\mathcal{A}_k \cap \mathcal{H}_i) \times \theta_1(\mathcal{H}_i), \tag{4}$$

which allocate GP prescriptions to IZs proportionally to the product of the population overlap via $P(\mathcal{A}_k \cap \mathcal{H}_i)$ and the IZ-level risk via $\theta_1(\mathcal{H}_i)$. The latter are the current values obtained from step (1) of the algorithm, and most $\pi_{ki}^* = 0$ as $P(\mathcal{A}_k \cap \mathcal{H}_i) = 0$ for most combinations of GP surgery and IZ. The choice of weights in (4) is not unique, because Li et al. (2012) and Taylor et al. (2018) proposed slightly different specifications. Our weights are most similar to those of Taylor et al. (2018) outlined in Web Appendix B, although ours are based on population intersection rather than land area intersection.

However, initial analyses showed that these weights can result in either a nonconverging MCMC algorithm or highly inaccurate parameter estimation. For example, in simulated data sets, the relative rates of GP prescriptions at the IZ level $\{\widehat{\theta}_1(\mathcal{H}_i)\}$ were routinely estimated as either very close to 0 or very large at around 20, whereas the true values lay in the interval (0.25, 4). The cause of this problem is having to simultaneously estimate the model parameters and the IZ-level disease counts, and could be phrased as a *rich get richer* problem. This is because if a large value of $Y_1(\mathcal{H}_i)$ is simulated by chance, then the next sample of $\boldsymbol{\phi}_1(\mathcal{H}_i)$ and hence $\theta_1(\mathcal{H}_i)$ is inflated. These inflated values then feed back to generate an even larger value of $Y_1(\mathcal{H}_i)$ in the next MCMC iteration via (4), and this cycle continues leading to inaccurate parameter estimation. Therefore, we propose the following spatially smoothed weights:

$$\pi_{ki}^* = P(\mathcal{A}_k \cap \mathcal{H}_i) \times \widetilde{\theta}_1(\mathcal{H}_i),$$

$$\widetilde{\theta}_1(\mathcal{H}_i) = \exp\{\mathbf{x}(\mathcal{H}_i)^\top \boldsymbol{\beta}_1 + \widetilde{\phi}_1(\mathcal{H}_i)\},$$

$$\widetilde{\phi}_1(\mathcal{H}_i) = \alpha\phi_1(\mathcal{H}_i) + (1-\alpha)\frac{\sum_{j=1}^{M} w_{ij}\phi_1(\mathcal{H}_j)}{\sum_{j=1}^{M} w_{ij}}. \quad (5)$$

This replaces $\theta_1(\mathcal{H}_i)$ in (4) with the spatially smoothed alternative $\widetilde{\theta}_1(\mathcal{H}_i)$, which leads to less extreme values of $\{\widetilde{\theta}_1(\mathcal{H}_i)\}$ in the multinomial weights and thus prevents the *rich get richer* phenomenon. This spatial smoothing is applied to the random effects $\phi_1(\mathcal{H}_i)$, because it is these that are assumed to be spatially smooth by (2). Here, $\alpha \in [0,1]$ is a spatial smoothing parameter, and using this as a simple spatial smoother was suggested by Banerjee et al. (2004). Clearly setting $\alpha = 1$ in (5) is equivalent to (4). This spatial smoothing step is only applied in (3) and not at any other part of the MCMC algorithm, and we assess its performance in the simulation study in the next section for a range of values of α.

4 | SIMULATION STUDY

We present a simulation study to assess: (i) how accurate is the inference from our spatial realignment methodology? and (ii) which value of α gives the best estimates? This study provides novel insight into spatial realignment for count data, because Bradley et al. (2016) only considered the ability of the models to estimate fitted values and not covariate effects, whereas Taylor et al. (2018) did not undertake any assessment of inferential accuracy.

4.1 | Data generation

Data are generated to match the Scotland study with prescription data for $K = 869$ GP surgeries $\{\mathcal{A}_k\}$ and hospitalizaton and death data for $M = 1252$ IZs $\{\mathcal{H}_i\}$, and our aim is to make inference at the IZ level. Initially, true IZ-level disease risks are generated from $\theta_j(\mathcal{H}_i) = \exp\{\texttt{Educ}(\mathcal{H}_i)\beta_{j_1} + \texttt{Acc}(\mathcal{H}_i)\beta_{j_2} + \phi_j(\mathcal{H}_i)\}$, where $(\texttt{Acc}(\mathcal{H}_i), \texttt{Educ}(\mathcal{H}_i))$, respectively, denote the access and education domains of the SIMD from the motivating study. The true values are $\beta_{j_2} = 0$ for all three severities j, whereas $\beta_{1_1} = 0.075, \beta_{2_1} = 0.05, \beta_{3_1} = 0.025$ to match the estimates from the real data. The random effects $\{\phi_j(\mathcal{H}_i)\}$ are generated from a zero mean multivariate normal distribution, whose covariance structure is the Kronecker product of a spatial covariance matrix and a between severity covariance matrix. The former is defined by a spatial exponential covariance structure, where the spatial range parameter is fixed so that the correlation between two IZs whose centroids is 1 km apart is 0.9. The between severity correlation matrix is chosen to match the correlations observed in the real data, namely, 0.5 between GP prescriptions and hospitalizations, 0.45 between GP prescriptions and deaths, and 0.9 between hospitalizations and deaths. Finally, the variance of the random effects v^2 is allowed to vary in our simulation design.

For the hospitalization and death outcomes, the expected counts $e_j(\mathcal{H}_i)$ come from the real data, whereas $Y_j(\mathcal{H}_i)$ are generated from (1). In contrast, as the GP prescription data need to be aggregated to the GP surgery level $\{\mathcal{A}_k\}$, we first generate observed and expected disease counts for the intersection populations $\{\mathcal{A}_k \cap \mathcal{H}_i\}$. In the majority of these $P(\mathcal{A}_k \cap \mathcal{H}_i) = 0$, and hence $Y_1(\mathcal{A}_k \cap \mathcal{H}_i) = 0$ and $e_1(\mathcal{A}_k \cap \mathcal{H}_i) = 0$. In contrast, if $P(\mathcal{A}_k \cap \mathcal{H}_i) > 0$, then we fix $e_1(\mathcal{A}_k \cap \mathcal{H}_i) = \gamma P(\mathcal{A}_k \cap \mathcal{H}_i)$, where γ controls disease prevalence and is chosen to match the real data. Finally, we generate $Y_1(\mathcal{A}_k \cap \mathcal{H}_i) \sim \text{Poisson}\{e_1(\mathcal{A}_k \cap \mathcal{H}_i)\theta_1(\mathcal{H}_i)\}$, which has a similar mean model to (1). Then the observed and expected disease counts at the GP surgery level are computed via $Y_1(\mathcal{A}_k) = \sum_{i=1}^{M} Y_1(\mathcal{A}_k \cap \mathcal{H}_i)$ and $e_1(\mathcal{A}_k) = \sum_{i=1}^{M} e_1(\mathcal{A}_k \cap \mathcal{H}_i)$. Thus, the simulated data being modeled are GP prescription data at the GP surgery level, and hospitalization, death, and covariate data at the IZ level.

4.2 | Models and scenarios

Models (1)–(2) are fitted to the data with $\alpha = 0.5, 0.7, 0.9, 0.95, 0.99, 1$, where $\alpha = 1$ corresponds to a nonsmoothed data augmentation algorithm. Additionally, we compare two other models, the first being applied directly to the IZ-level data (denoted as `IZ`), which allows us to quantify the loss in estimation accuracy from having spatial misalignment. The second is a simple spatial realignment approach based on population-weighted interpolation (denoted as `Naive`), which allows us to evidence the utility of our methods against a simpler alternative. Full details of both these models are given in Web Appendix C. All models are applied to 100 simulated data sets generated under each of six scenarios, which include all pairwise combinations of the following two factors.

- **Spatial risk variation.** It is controlled by comparing $v^2 = 0.05 \rightarrow \theta_j(\mathcal{H}_i) \in [0.45, 2]$ and $v^2 = 0.2 \rightarrow \theta_j(\mathcal{H}_i) \in [0.2, 4]$.
- **Accuracy of** $e_1(\mathcal{H}_i)$. In the motivating study, the GP prescription counts $\{e_1(\mathcal{H}_i)\}$ are estimated rather than computed exactly, due to a lack of data on national age-sex-specific GP prescribing rates. We examine the effect

that such estimation error has on model performance, by fitting the models using $\{e_1(\mathcal{H}_i)\}$ calculated with three levels of error. The first of these is no error, whereas the remaining two add zero-mean Gaussian random noise to the true values with standard deviations are $\omega = 3$ and $\omega = 6$. Both these correspond to small amounts of error because the mean value of $\{e_1(\mathcal{H}_i)\}$ is 1500.

Inference from each model is based on a single MCMC chain run for 250,000 iterations. The first 50,000 iterations are removed as the burn-in period, and the remaining 200,000 are thinned by 100 to reduce their autocorrelation, resulting in 2000 posterior samples for inference. The data augmentation step is run every 30 iterations, and pilot analyses suggested that these specifications were sufficient for the MCMC algorithm to converge.

4.3 | Results

The results for the GP prescription outcome are presented here because it is spatially misaligned with the set of IZs used for inference. The results for the hospitalization and death outcomes where there is no spatial misalignment are presented in Web Appendix C for completeness. The accuracy of the IZ-level disease risks $\{\theta_1(\mathcal{H}_i)\}$ and the two covariate effects $(\beta_{1_1}, \beta_{1_2})$ are summarized below by bias and relative mean absolute error (RMAE), the latter being presented as a percentage of the true value. The models produce generally unbiased estimates of disease risk across the scenarios, with biases being less than 0.6 in absolute value in almost all cases. The main exception to this is when $\alpha = 1$, which has four scenarios with larger positive biases (all those where $\omega > 0$).

The percentage RMAEs for $\{\theta_1(\mathcal{H}_i)\}$ are displayed in the top panel of Table 1, which shows that as expected having no spatial misalignment in the data (model `IZ`) leads to the best results (lowest RMAE) in all cases. In contrast, having spatially misaligned data causes around a threefold increase in RMAE when there is relatively little spatial variation in disease risk ($\upsilon^2 = 0.05$), and around a fivefold increase when there is large spatial variation in disease risk ($\upsilon^2 = 0.2$). However, in both cases, the percentage RMAE values are still relatively small, being mostly between 6% and 16% of the size of the true risks.

The `Naive` interpolation model consistently performs worse than the model proposed here for a range of values of α, suggesting that the complexity of our data augmentation approach is warranted. The data augmentation approach used by Taylor et al. (2018) ($\alpha = 1$) performs comparably with our spatially smoothed variant when there is no estimation error in $\{e_1(\mathcal{H}_i)\}$ ($\omega = 0$), but has much larger RMAEs when the expected counts are unknown exactly ($\omega > 0$) as in the motivating study. Typically, large values of α less than 1 produce the smallest RMAEs, with $\alpha = 0.95, 0.99$ being best when $\omega = 0$ and $\alpha = 0.9, 0.95$ generally being best when $\omega > 0$. The slight exception to this is in the most extreme case where the risk surface is highly spatially variable ($\upsilon^2 = 0, 2$) and the error in $\{e_1(\mathcal{H}_i)\}$ is largest ($\omega = 6$), where smaller values of α provide better results. However, in most cases, there are relatively little differences in RMAEs when $\alpha \in [0.7, 0.95]$, suggesting that the results are relatively robust.

The results for the education variable are included to illustrate whether the models can accurately estimate covariate effects when they are present. Overall, the results show a very similar pattern to those for the risk estimates, and their RMAE values are displayed in the middle panel of Table 1. Unsurprisingly, the best results occur if there is no spatial misalignment (model `IZ`), whereas the simple interpolation model (`Naive`) exhibits hugely attenuated estimates close to zero in all cases, which causes the large RMAEs of around 70% of the true value. The standard data augmentation algorithm ($\alpha = 1$) produces comparable results to our smoothed version if $\omega = 0$, but exhibits large positive biases and hence large RMAE values when there are errors in the expected counts. The results for our smoothed data augmentation algorithm ($\alpha < 1$) are not hugely different when $\alpha \in [0.7, 0.95]$, although $\alpha = 0.95$ typically performs the best across the range of scenarios considered.

The results for the access variable in the bottom panel of Table 1 allow us to investigate whether the models can accurately estimate when a covariate has no relationship with disease risk. As the true value of $\beta_{1_2} = 0$, the results presented are raw mean absolute errors. The `Naive` interpolation model produces the smallest MAEs, which is artificial because its covariate effect estimates are greatly attenuated toward zero (see above). Our spatially smoothed model performs similarly for $\alpha \in [0.7, 0.95]$, and its MAE values are not that much larger than those from the model with no spatial misalignment (`IZ`), suggesting that spatial misalignment has a relatively small effect in this setting. Additionally, in common with the previous results if $\alpha = 1$ and $\omega > 0$, then the MAE values are greatly inflated compared to our spatially smoothed model. Finally, Web Appendix D presents smaller scale additional simulation studies that investigate the impact of changing additional aspects of this study.

5 | RESULTS FROM THE RESPIRATORY DISEASE STUDY

5.1 | Model fitting

We fit the spatial realignment model to the data with three different covariate combinations: (i) no covariates, (ii)

TABLE 1 Percentage relative (to the true value) mean absolute errors (RMAE) of the regression parameters and the risk estimates relating to the GP prescriptions outcome for each scenario and model. Note that as the true value of the regression parameter β_{1_2} relating to the access covariate is zero, these results are raw mean absolute errors

	Model							
Scenario	**IZ**	**Naive**	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 0.9$	$\alpha = 0.95$	$\alpha = 0.99$	$\alpha = 1$
Risk - $\{\theta_1(\mathcal{H}_i)\}$								
$\upsilon^2 = 0.05, \omega = 0$	2.17	8.39	6.76	6.53	6.19	6.07	6.04	6.28
$\upsilon^2 = 0.05, \omega = 3$	2.17	8.44	6.81	6.59	6.28	6.21	6.62	38.10
$\upsilon^2 = 0.05, \omega = 6$	2.17	8.41	6.80	6.61	6.43	6.60	13.54	111.73
$\upsilon^2 = 0.2, \omega = 0$	2.25	16.13	13.63	13.05	12.05	11.58	10.92	11.27
$\upsilon^2 = 0.2, \omega = 3$	2.25	16.15	13.63	13.05	12.11	11.75	13.86	91.71
$\upsilon^2 = 0.2, \omega = 6$	2.25	16.12	13.71	13.77	15.64	18.37	30.91	127.78
Education - β_{1_1}								
$\upsilon^2 = 0.05, \omega = 0$	6.28	69.70	10.50	10.24	9.78	9.23	9.03	9.14
$\upsilon^2 = 0.05, \omega = 3$	5.98	69.65	9.19	8.90	8.38	8.18	8.72	24.19
$\upsilon^2 = 0.05, \omega = 6$	6.42	70.13	10.83	10.51	10.07	9.79	12.33	79.00
$\upsilon^2 = 0.2, \omega = 0$	11.45	71.35	20.13	19.27	17.75	16.84	15.02	14.69
$\upsilon^2 = 0.2, \omega = 3$	12.67	70.11	21.00	20.27	18.99	17.93	17.39	72.51
$\upsilon^2 = 0.2, \omega = 6$	13.09	71.36	21.45	21.09	19.10	19.21	21.43	104.71
Access - β_{1_2}								
$\upsilon^2 = 0.05, \omega = 0$	0.59	0.34	0.86	0.84	0.80	0.79	0.77	0.77
$\upsilon^2 = 0.05, \omega = 3$	0.56	0.30	0.74	0.72	0.71	0.71	0.73	4.08
$\upsilon^2 = 0.05, \omega = 6$	0.67	0.38	0.93	0.91	0.88	0.87	1.10	14.34
$\upsilon^2 = 0.2, \omega = 0$	1.19	0.75	1.96	1.87	1.76	1.70	1.68	1.62
$\upsilon^2 = 0.2, \omega = 3$	1.37	0.79	2.03	1.99	1.88	1.85	1.86	13.40
$\upsilon^2 = 0.2, \omega = 6$	1.26	0.70	1.97	1.86	1.74	1.71	2.16	17.94

covariates chosen by a model building strategy (denoted as `full`), and (iii) each SIMD covariate included in a separate model. The first two of these are included as a sensitivity analysis to see what impact it has on the resulting spatial risk maps (motivating question (a)), whereas the single covariate models allow us to quantify the consistency of the social inequalities in disease risk (motivating question (b)) across the SIMD domains. In building the `full` model, the education, employment, and income domains of the SIMD are collinear, with Pearson's correlation coefficients above 0.9. Therefore, we only included the education domain from these three, because it minimizes the Akaike information criterion (AIC) when incorporated in a simple Poisson log-linear model in conjunction with the access, crime, and housing variables.

We fitted the model with $\alpha = 0.8, 0.9, 0.95, 0.99, 1$ for each of these covariates combinations, and inference in each case is based on 15,000 MCMC samples obtained from three parallel Markov chains. Each chain was run for 1,100,000 iterations with a burn-in period of 100,000, and the remaining samples were thinned by 200 to reduce their autocorrelation. Convergence of the Markov chains for selected parameters was visually assessed using traceplots and numerically assessed using the Gelman–Rubin diagnostic (Gelman et al., 2013), and when $\alpha = 0.8, 0.9, 0.95$, the results show no evidence against convergence. However, when $\alpha = 0.99, 1$, the Markov chains show evidence of nonconvergence, which is visually evident from the traceplots and numerically supported by Gelman–Rubin diagnostics well above 1.1. This lack of convergence is due to the *rich get richer* phenomenon described earlier, and we illustrate this by presenting the estimated risk map for the GP prescriptions outcome when $\alpha = 1$ in Web Appendix E. Thus, for the rest of this section, we only present the results for models that showed no evidence against convergence, namely, when $\alpha = 0.8, 0.9, 0.95$.

5.2 | Overall model fit and estimated dependence structures

The overall fit of each model to the data is summarized in panel 1 of Table 2 by the widely applicable information criterion (WAIC), with the estimated number of independent parameters $p.w$ in brackets. The table shows that the `full` model provides the best fit to the data for every value

TABLE 2 Summary of the models fitted to the data, including their overall fit as measured by the WAIC and the effective number of independent parameters $p.w$ (in brackets), as well as posterior means and 95% credible intervals for the correlation parameters

Quantity	Covariate model	Spatial smoothing parameter $\alpha = 0.8$	$\alpha = 0.9$	$\alpha = 0.95$
	None	29,470 (1580)	29,190 (1467)	29,046 (1418)
	Full	29,352 (1468)	29,104 (1322)	28,943 (1261)
	Access	29,466 (1578)	29,187 (1464)	29,051 (1418)
WAIC	Crime	29,443 (1554)	29,163 (1438)	29,025 (1387)
($p.w$)	Education	29,363 (1476)	29,095 (1357)	28,976 (1264)
	Employment	29,383 (1513)	29,111 (1402)	28,967 (1348)
	Housing	29,436 (1516)	29,159 (1396)	29,052 (1323)
	Income	29,376 (1498)	29,101 (1383)	28,992 (1303)
	None	0.93 (0.88, 0.97)	0.89 (0.83, 0.94)	0.83 (0.75, 0.90)
	Full	0.89 (0.82, 0.94)	0.81 (0.72, 0.89)	0.70 (0.60, 0.80)
Residual	Access	0.93 (0.88, 0.97)	0.88 (0.82, 0.94)	0.82 (0.74, 0.89)
spatial	Crime	0.95 (0.91, 0.98)	0.91 (0.85, 0.95)	0.84 (0.76, 0.91)
correlation	Education	0.93 (0.88, 0.97)	0.87 (0.80, 0.93)	0.77 (0.68, 0.85)
(ρ)	Employment	0.93 (0.89, 0.97)	0.87 (0.80, 0.93)	0.77 (0.67, 0.86)
	Housing	0.97 (0.94, 0.99)	0.95 (0.91, 0.98)	0.91 (0.85, 0.96)
	Income	0.94 (0.90, 0.97)	0.89 (0.83, 0.95)	0.79 (0.70, 0.88)
	None	0.49 (0.44, 0.54)	0.48 (0.43, 0.53)	0.45 (0.40, 0.50)
	Full	0.15 (0.07, 0.22)	0.13 (0.06, 0.20)	0.11 (0.04, 0.18)
Residual	Access	0.49 (0.43, 0.53)	0.47 (0.42, 0.52)	0.45 (0.40, 0.50)
severity	Crime	0.36 (0.30, 0.42)	0.36 (0.29, 0.42)	0.34 (0.28, 0.40)
correlation	Education	0.11 (0.04, 0.18)	0.10 (0.03, 0.17)	0.08 (0.01, 0.16)
($\mathbf{\Sigma}_{12} = \frac{\Upsilon_{12}}{\sqrt{\Upsilon_{11}\Upsilon_{22}}}$)	Employment	0.12 (0.05, 0.19)	0.10 (0.03, 0.18)	0.08 (0.01, 0.16)
	Housing	0.44 (0.38, 0.49)	0.43 (0.37, 0.48)	0.41 (0.34, 0.46)
	Income	0.10 (0.03, 0.17)	0.09 (0.01, 0.16)	0.07 (−0.01, 0.14)
	None	0.45 (0.37, 0.53)	0.44 (0.36, 0.53)	0.43 (0.34, 0.51)
	Full	0.01 (−0.16, 0.17)	0.00 (−0.22, 0.22)	0.00 (−0.25, 0.25)
Residual	Access	0.44 (0.36, 0.52)	0.44 (0.35, 0.52)	0.42 (0.33, 0.51)
severity	Crime	0.29 (0.19, 0.39)	0.30 (0.19, 0.40)	0.29 (0.17, 0.41)
correlation	Education	−0.10 (−0.25, 0.05)	−0.13 (−0.32, 0.04)	−0.19 (−0.41, 0.05)
($\mathbf{\Sigma}_{13} = \frac{\Upsilon_{13}}{\sqrt{\Upsilon_{11}\Upsilon_{33}}}$)	Employment	0.00 (−0.12, 0.13)	−0.01 (−0.15, 0.13)	−0.02 (−0.17, 0.12)
	Housing	0.41 (0.30, 0.51)	0.42 (0.30, 0.54)	0.46 (0.32, 0.59)
	Income	−0.05 (−0.19, 0.08)	−0.07 (0.21, 0.68)	−0.09 (−0.27, 0.09)
	None	0.93 (0.89, 0.96)	0.93 (0.89, 0.96)	0.93 (0.88, 0.97)
	Full	0.75 (0.62, 0.88)	0.90 (0.69, 0.99)	0.98 (0.92, 1.00)
Residual	Access	0.92 (0.88, 0.95)	0.92 (0.88, 0.96)	0.92 (0.87, 0.96)
severity	Crime	0.89 (0.83, 0.94)	0.89 (0.83, 0.94)	0.90 (0.82, 0.99)
correlation	Education	0.76 (0.64, 0.86)	0.80 (0.66, 0.93)	0.95 (0.87, 0.99)
($\mathbf{\Sigma}_{23} = \frac{\Upsilon_{23}}{\sqrt{\Upsilon_{22}\Upsilon_{33}}}$)	Employment	0.81 (0.71, 0.89)	0.81 (0.71, 0.89)	0.82 (0.70, 0.96)
	Housing	0.90 (0.83, 0.95)	0.92 (0.85, 0.97)	0.98 (0.89, 1.00)
	Income	0.79 (0.69, 0.88)	0.80 (0.69, 0.90)	0.90 (0.71, 0.99)

of α, although the difference in WAIC between it and the model with only the education covariate is very small. Additionally, for each covariate model, $\alpha = 0.95$ provides the best overall fit to the data of the different α values considered.

Panel 2 in Table 2 displays posterior means and 95% credible intervals for the spatial dependence parameter ρ, which shows that the data contain substantial spatial dependence (ρ close to 1) even after the covariate effects have been accounted for. The posterior mean does reduce slightly as α increases, which is due to the reduced amount of spatial smoothing induced by the data augmentation step (5). The between severity correlations are summarized in panels 3–5 of the table, and display the between severity correlations for the random effects derived from $\boldsymbol{\Upsilon}$. These correlations are highest for the model with no covariates, which is because in the covariate adjusted models, the covariates account for some of the correlations in the data. For the no covariate model, the GP prescription data are moderately correlated with both hospitalizations and deaths, with correlations between 0.4 and 0.5. In contrast, the correlation between hospitalizations and deaths is higher being around 0.93. This stronger correlation is not surprising, because a death from respiratory disease is often preceded by a stay in hospital for that individual. In contrast, some of the mild cases of respiratory disease (e.g., asthma) will be successfully managed without the patient requiring hospitalization, which explains the lower correlations in this case.

5.3 | Spatial risk surfaces

The estimated risk surfaces from the full covariate model show little change when varying $\alpha \in \{0.8, 0.9, 0.95\}$, and the biggest differences naturally occur for the GP prescription outcome because these disease counts are estimated rather than known at the IZ scale. The mean absolute differences (over the IZs) between the estimated risk surfaces obtained from $\alpha \in \{0.8, 0.9, 0.95\}$ range between 0.038 and 0.085, which compares to disease risk estimates that range between 0.1 and 4.3 across the IZs. The differences between the estimated risks from the `full` model and the model with no covariates are also relatively small, with mean average differences of 0.126 (GP prescriptions), 0.038 (hospitalizations), and 0.091 (deaths) when $\alpha = 0.95$. Thus, in what follows we present the results from the `full` model with $\alpha = 0.95$, because it best fits the data as measured by the WAIC and the simulation study showed it performed consistently well across the scenarios considered.

The correlation between the posterior mean risk surfaces for hospitalizations and deaths is 0.947, whereas the corresponding correlations between GP prescription and the other two outcomes are lower at 0.635 with hospitalizations and 0.610 with deaths. Figure 2 displays maps of the estimated risk surfaces for the two largest cities Edinburgh and Glasgow for all three severities, and we focus on these cities because displaying Scotland-wide maps as in Figure 1 makes it hard to see the small-scale risk variation. The most striking feature of these maps is that Glasgow exhibits much higher risks than Edinburgh on average for all severities, which is due to the well-known *Glasgow effect* that Glasgow exhibits some of the poorest health in the western world.

The risk patterns are largely consistent across the three severities, and in Edinburgh, the high-risk areas are mainly in Craigmillar in the east and Broomhouse in the west. In Glasgow, the main high-risk areas include Drumchapel in the west, Castlemilk in the south, and a large part of the east end of the city. The other striking feature of the Glasgow maps is that visually, there appears to be more high-risk areas as the severity of disease increases, with the high-risk areas for GP prescriptions expanding when increasing in severity to hospitalizations and then deaths. This is most apparent in the north-east of the city in locations such as Riddrie, which has a risk 22% below the Scottish average ($\widehat{\theta}_1(\mathcal{H}_i) = 0.78$) for GP prescriptions, compared to risks that are 59% above ($\widehat{\theta}_2(\mathcal{H}_i) = 1.59$) and 61% above ($\widehat{\theta}_3(\mathcal{H}_i) = 1.61$) the Scottish average for hospitalizations and deaths, respectively.

5.4 | Health inequalities

The inequalities in disease risk are summarized in Table 3 for all severities, where Panel (A) presents total inequality, while Panel (B) displays social inequality. Total inequality is summarized by the standard deviations and interquartile ranges in the posterior mean risk surfaces, which come from the `full` model with $\alpha = 0.95$. The table shows that the level of total inequality reduces as the severity of disease increases, which is a consistent finding across both standard deviation and interquartile range metrics. For example, the interquartile range in risk is 0.602 for GP prescriptions, 0.579 for hospitalizations, and 0.469 for deaths, a reduction of 22% in the inequality of deaths compared to the inequality in GP prescriptions.

The level of social inequality is summarized in Panel (B) of Table 3 by the relative risk between each SIMD covariate and disease risk. All results relate to $\alpha = 0.95$ and come from single covariate models, and to ensure comparability across the different SIMD domains, all relative risks relate to a one-standard-deviation increase in the covariate in question. For completeness, the estimated covariate effects from the full model are presented in Web Appendix E. With the exception of the access domain that specifically

FIGURE 2 Maps displaying the posterior mean risk estimates for Edinburgh (left) and Glasgow (right) for GP prescription rates (top), hospitalizations (middle) and deaths (right). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

TABLE 3 Summary of the health inequalities in respiratory disease risk for all three severities. Panel (A) presents total inequality, whereas Panel (B) displays social inequality

(A) Total inequality			
	Severity		
Inequality metric	**GP prescription**	**Hospitalization**	**Death**
Standard deviation	0.503	0.403	0.342
Interquartile Range	0.602	0.579	0.469
(B) Social inequality			
	Severity		
SIMD domain	**GP prescription**	**Hospitalization**	**Death**
Access	0.944 (0.912, 0.977)	0.915 (0.896, 0.934)	0.920 (0.889, 0.951)
Crime	1.297 (1.245, 1.350)	1.205 (1.186, 1.226)	1.235 (1.200, 1.271)
Education	1.465 (1.425, 1.513)	1.321 (1.303, 1.339)	1.324 (1.294, 1.356)
Employment	1.486 (1.439, 1.536)	1.288 (1.270, 1.306)	1.279 (1.246, 1.312)
Housing	1.259 (1.204, 1.314)	1.280 (1.257, 1.305)	1.331 (1.295, 1.369)
Income	1.502 (1.451, 1.553)	1.295 (1.278, 1.313)	1.293 (1.261, 1.325)

measures geographical access to services rather than general socioeconomic deprivation, all estimated relative risks and corresponding 95% credible intervals are greater than one. This shows consistent convincing evidence of social inequality in disease risk for all severities, as increases in the levels of socioeconomic deprivation are associated with significant increases in disease risk. The sizes of these effects are large and range between a 20.5% increased risk of hospitalization for the crime indicator, to a 50.2% increase in GP prescription rates when the income deprivation indicator increases by one standard deviation. The levels of social inequality seem to be highest for the GP prescription outcome, because it has the largest inequality for the crime, education, employment, and income domains of the SIMD.

6 | DISCUSSION

This paper develops a new approach for modeling multivariate spatially misaligned disease count data with known and partially unknown spatial supports, which delivers inference on a common spatial scale using a spatially smoothed data-augmented MCMC algorithm. The simulation study shows that the spatial smoothing in the data augmentation step provides reliable inference across a range of scenarios, and consistently outperforms the data-augmented MCMC algorithm proposed by Taylor et al. (2018) in the data context considered here. The study suggests that $\alpha \in [0.9, 0.95]$ provides the most accurate inference, but that the results do not vary greatly for values down to $\alpha = 0.7$. This preference for a small amount of spatial smoothing is because as α gets closer to zero, the current estimate of risk used in the data augmentation step is overly smooth, leading to oversmoothing and poorer estimation in the IZ-level risk estimates. In contrast, as α approaches 1, this oversmoothing reduces, leading to better estimation. However, if α gets very close to 1, the MCMC algorithm is affected by the *rich get richer* phenomenon, leading to inaccurate parameter estimation. In the simulation study, the presence of spatial misalignment reduces inferential accuracy as expected, with the RMAE for disease risk increasing by between three- and fivefold depending on the scenario, whereas the corresponding result for covariate effects is only a 1.5-fold increase. However, for disease risk, the absolute size of the RMAE in the presence of spatial misalignment is only between 6% and 16% of the true risk size, suggesting that reliable inference can still be obtained from spatially misaligned data.

The novel insight provided by this paper is how disease risk varies by the severity, which extends the existing literature that ubiquitously focuses on a single severity of outcome. The first key finding from our Scotland respiratory disease study is that the spatial risk trends are largely consistent across the different severities, with mainly the same subregions exhibiting elevated risks for each severity. This is likely to be because areas that exhibit elevated risks will have increased exposures to factors such as smoking and air pollution, which thus affect all severities of outcome. Additionally, elevated risk areas are likely to self-perpetuate across severities, because having a larger proportion of people suffering from mild disease is likely to lead to this cohort having higher rates of more severe disease.

The second key finding from our Scotland study is that both total and social health inequalities reduce as the severity of disease outcome increases, with mild disease

treated in primary care being much more unevenly distributed across different sections of society than moderate cases requiring hospitalization or severe cases resulting in death. This reduced inequality for deaths is likely to be because everybody dies eventually, and as respiratory disease is one of the most common causes of death, it affects people from all communities. In contrast, respiratory diseases such as asthma that can be treated in primary care are more likely to affect younger people (see table 8.2 in Scottish Government, 2018), and the incidence rates of such cases tend to be driven by risk factors such as those highlighted above whose magnitude varies spatially. However, this examination of social health inequalities was based on the estimated effects of covariates, which were obtained from a data set containing residual spatial autocorrelation. This autocorrelation was modeled by spatially correlated random effects as is standard practice in the literature, but Hodges and Reich (2010) have shown that confounding can happen in this context because the random effects can themselves be correlated with the covariates. A number of approaches such as Hughes and Haran (2013) have been proposed to deal with this potential issue when there is no spatial misalignment in the data, but the extension of this work to our spatial misalignment setting has not been explored and is an important avenue for future work.

Finally, this paper represents the start of a long-term focus on spatially misaligned count data modeling, which has received much less attention in the literature to date compared to spatial misalignment in a continuous data setting. Our use of IZs here as the common inferential scale was a pragmatic choice based on data availability, but may suffer from the well-known modifiable areal unit problem (MAUP, Openshaw and Taylor, 1979). Li et al. (2012) and Taylor et al. (2018) overcome this limitation by providing pseudospatially continuous inference on a regular grid, but they mainly focus on the simpler setting where the areal units have known spatial extents. In our study, the GP surgery catchment areas have partially unknown spatial supports, precluding the use of area of intersection in the spatial misalignment mechanism as used by the above authors. Therefore, future work will begin by investigating how to deliver pseudospatially continuous inference on a regular grid in the context of areal units with partially unknown spatial supports.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of two reviewers and an associate editor, whose suggestions have improved the motivation for and content of this paper.

DATA AVAILABILITY STATEMENT

The General Practice prescription and Scottish Index of Multiple Deprivation data are available at https://www.opendata.nhs.scot/dataset/prescriptions-in-the-community and https://www.gov.scot/publications/scottish-index-multiple-deprivation-2016/ , while the hospitalization and death data can be requested from Public Health Scotland (https://publichealthscotland.scot/).

ORCID

Duncan Lee https://orcid.org/0000-0002-6175-6800

REFERENCES

Banerjee, S., Carlin, B. & Gelfand, A. (2004) *Hierarchical modelling and analysis for spatial data*. Boca Raton: Chapman & Hall.

Bansal, P., Krueger, R. & Graham, D. (2021) Fast Bayesian estimation of spatial count data models. *Computational Statistics & Data Analysis*, 157, 107152.

Besag, J., York, J. & Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, 43, 1–59.

Blangiardo, M., Finazzi, F. & Cameletti, M. (2016) Two-stage Bayesian model to evaluate the effect of air pollution on chronic respiratory diseases using drug prescriptions. *Spatial and Spatio-temporal Epidemiology*, 18, 1–12.

Bradley, J., Holan, S. & Wikle, C. (2018) Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, 13, 253–310.

Bradley, J., Wikle, C. & Holan, S. (2016) Bayesian spatial change of support for count-valued survey data with application to the American community survey. *Journal of the American Statistical Association*, 514, 472–487.

Bradley, J., Wikle, C. & Holan, S. (2017) Regionalization of multiscale spatal processes by using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society Series B*, 79, 815–832.

Flowerdew, R. & Green, M. (1989) Accuracy of spatial databases. In: *Statistical methods for inference between incompatible zonal systems*. Taylor and Francis, London, pp. 239–247.

Flowerdew, R. & Green, M. (1993) *Developments in areal interpolation methods and GIS*. Berlin, Heidelberg: Springer, pp. 73–84.

Gelfand, A. & Schliep, E. (2016) Spatial statistics and Gaussian processes: a beautiful marriage. *Spatial Statistics*, 18, 86–104.

Gelfand, A. & Vounatsou, P. (2003) Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4, 11–15.

Gelfand, A., Zhu, L. & Carlin, B. (2001) On the change of support problem for spatio-temporal data. *Biostatistics*, 2, 31–45.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013) *Bayesian data analysis*, 3rd edition. Boca Raton: Chapman and Hall/CRC.

Hodges, J. & Reich, B. (2010) Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64, 325–334.

Hughes, J. & Haran, M. (2013) Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75, 139–159.

Lee, D. (2018) A locally adaptive process-convolution model for estimating the health impact of air pollution. *The Annals of Applied Statistics*, 12, 2540–2558.

Leroux, B., Lei, X. & Breslow, N. (2000). Statistical models in epidemiology, the environment and clinical trials. In: Halloran, M. & Berry, D. (Eds.), *Estimation of disease rates in small areas: a new mixed model for spatial dependence*. New York: Springer-Verlag, pp. 135–178.

Li, Y., Brown, P., Gesink, D. & Rue, H. (2012) Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, 21, 479–507.

Li, Y., Brown, P., Rue, H., al-Maini, M. & Fortin, P. (2012) Spatial modelling of lupus incidence over 40 years with changes in census areas. *Journal of the Royal Statistical Society: Series C*, 61, 99–115.

Mugglin, A. & Carlin, B. (1998) Hierarchical modeling in geographic information systems: population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics*, 3, 111–130.

Nethery, R., Sandler, D., Zhao, S., Engel, L. & Kwok, R. (2019) A joint spatial factor analysis model to accommodate data from misaligned areal units with application to Louisiana social vulnerability. *Biostatistics*, 20, 468–484.

Openshaw, S. & Taylor, P. (1979) *A million or so correlation coefficients: three experiments on the modifiable areal unit problem*. Pion, London, pp. 127–144.

Scottish Government (2018) The Scottish Health Survey. https://www.gov.scot/publications/scottish-health-survey-2018-volume-1-main-report/.

Song, Y., Li, Y., Bates, B. & Wikle, C. (2014) A Bayesian hierarchical downscaling model for south-west Western Australia rainfall. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 715–736.

Tanner, M. & Wong, W. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–540.

Taylor, B., Andrade-Pacheco, R., & Sturrock, H. (2018) Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society Series A*, 181, 1125–1150.

Zhu, L., Carlin, B. & Gelfand, A. (2003) Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmentrics*, 14, 537–557.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures referenced in Sections 2– 5 are available with this paper at the Biometrics website on Wiley Online Library. Software in the form of R functions to fit the model together with simulated data is also available at the Biometrics website on Wiley Online Library.

How to cite this article: Lee, D. & Anderson, C. (2022) Delivering spatially comparable inference on the risks of multiple severities of respiratory disease from spatially misaligned disease count data. *Biometrics*, 1–14. https://doi.org/10.1111/biom.13739