Health & Place 16 (2010) 684–693

Contents lists available at ScienceDirect

Health & Place

journal homepage: www.elsevier.com/locate/healthplace

Small area estimation of sparse disease counts using shared component models-application to birth defect registry data in New South Wales, Australia

Arul Earnest [a,*], John R. Beard [a,c], Geoff Morgan [a,b], Douglas Lincoln [a], Richard Summerhayes [c], Deborah Donoghue [a], Therese Dunn [a,b], David Muscatello [d], Kerrie Mengersen [e]

[a] Northern Rivers University Department of Rural Health, The University of Sydney, 55 Uralba Street, Lismore, New South Wales 2480, Australia
[b] North Coast Area Health Service, New South Wales, Australia
[c] Graduate Research College, Southern Cross University, New South Wales, Australia
[d] Centre for Epidemiology and Research, New South Wales Department of Health, New South Wales, Australia
[e] Faculty of Science, Queensland University of Technology, Queensland, Australia

ARTICLE INFO

Article history:
Received 1 October 2009
Received in revised form
10 February 2010
Accepted 22 February 2010

Keywords:
CAR
Sparse
Spatial
Defects

ABSTRACT

In the field of disease mapping, little has been done to address the issue of analysing sparse health datasets. We hypothesised that by modelling two outcomes simultaneously, one would be able to better estimate the outcome with a sparse count. We tested this hypothesis utilising Bayesian models, studying both birth defects and caesarean sections using data from two large, linked birth registries in New South Wales from 1990 to 2004. We compared four spatial models across seven birth defects: spina bifida, ventricular septal defect, OS atrial septal defect, patent ductus arteriosus, cleft lip and or palate, trisomy 21 and hypospadias. For three of the birth defects, the shared component model with a zero-inflated Poisson (ZIP) extension performed better than other simpler models, having a lower deviance information criteria (DIC). With spina bifida, the ratio of relative risk associated with the shared component was 2.82 (95% CI: 1.46–5.67). We found that shared component models are potentially beneficial, but only if there is a reasonably strong spatial correlation in effect for the study and referent outcomes.

1. Introduction

In recent years, there has been considerable interest in the development and application of spatial models to analyse areal-level data. Most of these applications have been in the field of disease mapping. Bayesian methods, in particular, have been used to calculate smoothed relative risks of a particular disease at some areal level. While much attention has been given to developing temporal extensions to spatial models and methods for simultaneously analysing multiple outcomes, very little has been done to address the issue of analysing sparse datasets, where there could be an abundance of zero counts or large number of areas with extremely low expected counts of the disease.

The conditional autoregressive (CAR) model provides estimates of disease risk that borrow strength from neighbouring areas. One advantage of this approach is to provide more precise estimates of disease risk in areas with a small population, thus helping to overcome the problem of larger uncertainty associated with a smaller disease count. However, the performance of the model is questionable when neighbouring areas themselves are sparsely populated, as there is little information to borrow. This is particularly true of studies involving rare disease outcomes.

In epidemiological research, several outcomes can share the same risk factor (e.g. lung cancer, chronic obstructive pulmonary disease (COPD) with smoking) (Population Health Division, 2001). We hypothesised that Bayesian modelling which examined outcomes with shared risk factors simultaneously, may be a useful means of overcoming the problems of small area estimation of sparse counts. We also hypothesised that modelling two outcomes simultaneously should improve the estimation of the outcome with the sparse count, provided that both share a common spatially varying covariate, which need not be measured. We then tested this hypothesis in models incorporating birth defects and caesarean section rates. Birth defects are relatively rare events, but community perceptions of clusters of defects cause great concern. Therefore, they are a useful health outcome for evaluating methods for estimating spatially varying risks.

* Corresponding author. Tel.: +61 2 6620 7570; fax: +61 2 6620 7270.
E-mail address: arul_earnest@hotmail.com (A. Earnest).

1353-8292/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.healthplace.2010.02.006

Relatively few studies have examined the spatial variation in risks of specific birth defects, with neural tube defects being most commonly found to be spatially correlated (Meyer and Siega-Riz, 2002; Tuncbilek et al., 1999; Ericson et al., 1988; Borman and Cryer, 1993; Frey and Hauser, 2003; Rankin et al., 2005), followed by oral clefts (Saxen, 1975; Poletta et al., 2007). In anophthalmia and microphthalmia (Dolk et al., 1998) prevalence was found to be higher in rural versus urban areas, as it was in diaphragmatic hernia and gastroschisis (Rankin et al., 2005). Proximity of maternal residence to landfill sites was associated with certain birth defects such as neural tube defects, hypospadias and epispadias and abdominal wall defects (Elliott et al., 2001). Other studies have examined spatial distribution of birth defects in general (Rushton and Lolonis, 1996; Kuehl and Loffredo, 2006; Rushton et al., 1996). Caesarean sections are a common procedure and spatial variation in the distribution of caesarean section rates has also been well established (Taffel, 1994; Magadi et al., 2001; Baicker et al., 2006; Clarke and Taffel, 1996). One study found a four-fold variation between low and high use areas (Baicker et al., 2006).

Maternal age is commonly studied as a risk factor for both birth defects and caesarean counts, and is a readily available demographic variable in most birth defect registries. The effect of maternal age on the occurrence of birth defects is not uniform, with both very young maternal age and old maternal age associated with a different range of birth defects. For defects like Down Syndrome (trisomy 21), several studies (Gaulden, 1992; Hsieh et al., 1995; Reefhuis et al., 1999) have found a positive association between advanced maternal age and the risk of having babies with Down Syndrome. An analysis of two large birth registries combined together, for instance, showed that that mothers aged 40 years and above were 4.96 times (95% CI: 3.44–7.16) more likely than those aged below 40 to give birth to a baby with Down Syndrome, and this relationship was statistically significant ($p < 0.001$) (Reefhuis et al., 1999). Further, maternal age has been associated with caesarean section in a number of studies (Padmadas et al., 2000; Taffel, 1994; Seshadri and Mukherjee, 2005; Sims et al., 2000; Maslow and Sweeny, 2000; Witter et al., 1995; Parrish et al., 1994; Peipert and Bracken, 1993; Gordon et al., 1991). We use maternal age here as an example. We would like to emphasise that our subsequent model is not restricted to just maternal age, as it accommodates a number of variables, which can be latent, but varying spatially.

We evaluated the shared component Bayesian modelling approach by examining both birth defect and caesarean section counts simultaneously. We hypothesised that by modelling two related outcomes simultaneously, one should be able to better estimate the outcome with a sparse count, provided both share a common spatially varying covariate, which need not be measured.

2. Methods

The shared component model was developed by Knorr-Held and Best (2001) and was applied for the investigation of oral and oesophageal cancer mortality data for males in Germany. The model was initially used to separate the underlying risk surface for each outcome into a shared and outcome-specific component. The shared component was to be interpreted as a surrogate for unobserved covariates that display spatial structure and are common to both outcomes. The two outcome-specific risk components and the shared component are assumed to be independent, each with a spatial prior. The authors found two large clusters with a large shared component value, and they postulated that this was consistent with the distribution of risk factors in the neighbourhood. They also found distinct spatial patterns for each individual outcome.

We obtained data on birth anomalies in the state of New South Wales (NSW), Australia for the period 1990–2003 (inclusive) from the NSW Midwives Data Collection (MDC) and Birth Defects Register (BDR) databases. These mandatory registers of all births in the state are completed by clinicians at the time of the birth. In 1998, a 2% sample of the Midwives Data Collection records was validated against hospital records. The excellent quality of this database is reflected in high correlations and low missing data for almost all covariates (Centre for Epidemiology and Research, 2000). Further details on the two registries can be found elsewhere (Centre for Epidemiology and Research, 2007). Data from both registers was geocoded to the mother's usual address using software developed by NSW Health and the Australian National University using a previously described process (Summerhayes et al., 2006).

We calculated standardised expected counts for each of seven birth defects: spina bifida (SB), ventricular septal defect (VSD), ostium secundum atrial septal defect (ASD), patent ductus arteriosus (PDA), cleft lip and or palate (CLP), trisomy 21 (T21) and hypospadias (HPS). For example, the expected count of hypospadias in each statistical local area (SLA) was defined as $E_i = (\text{Births}_i/\text{Totalbirths}) \times \text{Totalhypospadias}$, where Births_i refers to total births in the ith SLA, and Totalbirths and Totalhypospadias refer to the overall number of births and hypospadias defects in the NSW study region for that particular time-period. These birth defects were studied for several reasons. They were the more common defects reported in Australia, and they covered a spectrum of body systems. They were also correlated with caesarean rates at varying strengths, which is important to allow us to quantify the performance of the shared component model, described in detail below.

We also calculated caesarean section counts for use in the shared component model. Analyses were undertaken at the statistical local area (SLA) level, for which there were 198 SLAs available. The median size of the SLAs in our study was 2069 km^2 (interquartile range 181–4103 km^2), while the median annual number of births in each SLA was 193 (interquartile range 66–670). SLA-specific relative risk estimates were calculated as the ratio of the observed and expected counts for each area. Caesarean section counts were obtained from the Midwives Data Collection registry, and we included both emergency and elective caesarean counts from 1990 to 2003.

Ethical approval was obtained for the use of the NSW Midwives Data Collection and the NSW Births Defects Registry data from the NSW Population and Health Services Research Ethics Committee, and for the study itself from the University of Sydney Ethics Committee.

We compared the following four models in our analysis: a simple CAR model, a CAR model with a zero-inflated Poisson (ZIP) extension to model the excess zeros, a shared component model and finally a shared component with a ZIP extension.

The simple CAR model consisted of both a spatially structured prior and spatially unstructured prior, which is sometimes known as the convolution prior or the Besag, York and Mollie (BYM) model (Besag et al., 1991). The BYM model allows for the smoothing of relative risk estimate in each region towards the mean risk in the neighbouring areas. This provides for a more precise or reliable estimate of both mean and variance compared to using the crude rate. Risks are also smoothed towards the global mean to account for overdispersion.

The basic shared component model was formulated as follows:

$$O_{1i} \sim Poi(E_{1i}\,\theta_i^{\delta}\,\phi_{1i})$$

$$O_{2i} \sim Poi(E_{2i}\,\theta_i^{1/\delta}\,\phi_{2i})$$