Detecting Small-Area Similarities in the Epidemiology of Childhood Acute Lymphoblastic Leukemia and Diabetes Mellitus, Type 1: A Bayesian Approach

Richard G. Feltbower¹, Samuel O. M. Manda², Mark S. Gilthorpe², Mel F. Greaves³, Roger C. Parslow¹, Sally E. Kinsey⁴, H. Jonathan Bodansky⁵, and Patricia A. McKinney¹

¹ Pediatric Epidemiology Group, Centre for Epidemiology and Biostatistics, University of Leeds, Leeds, United Kingdom.
² Biostatistics Unit, Centre for Epidemiology and Biostatistics, University of Leeds, Leeds, United Kingdom.
³ Leukemia Research Fund Centre, Institute of Cancer Research, London, United Kingdom.
⁴ Pediatric Oncology and Hematology Department, St. James’s University Hospital, Leeds, United Kingdom.
⁵ Diabetes Centre, The General Infirmary, Leeds, United Kingdom.

Received for publication May 21, 2004; accepted for publication January 12, 2005.

Childhood acute lymphoblastic leukemia and diabetes mellitus, type 1, have common epidemiologic and etiologic features, including correlated international incidence and associations with infections. The authors examined whether the diseases’ similar large-scale distributions are reflected in small geographic areas while also examining the influence of sociodemographic characteristics. Details of 299 children (0–14 years) with acute lymphoblastic leukemia and 1,551 children with diabetes diagnosed between 1986 and 1998 were extracted from two registers in Yorkshire, United Kingdom. Standardized incidence ratios across 532 electoral wards were compared using Poisson regression, confirming significant associations between population mixing and the geographic heterogeneity of both conditions. Bayesian methods analysis of spatial correlation between diseases by modeling a bivariate outcome based on their standardized incidence ratios was applied; spatial and heterogeneity components were included within a hierarchical random effects model. A positive correlation between diseases of 0.33 (95% credible interval: ±0.20, 0.74) was observed, and this was reduced after control for population mixing (r = 0.18), population density (r = 0.14), and deprivation (r = 0.06). The Bayesian approach showed a modest but nonsignificant joint spatial correlation between diseases, only partially suggesting that the risk of both was associated within some electoral wards. With Bayesian methodology, population mixing remained significantly associated with both diseases. The links between diabetes and acute lymphoblastic leukemia observed for large regions are weaker for small areas. More powerful replications are needed for confirmation of these findings.

Bayesian analysis; child; diabetes mellitus, type 1; hierarchical model; infection; leukemia; spatial model

Acute lymphoblastic leukemia is a distinct morphologic form of leukemia, albeit with a diversity of molecular subsets, and in children, common (precursor B-cell) acute lymphoblastic leukemia is the predominant subtype of leukemia, comprising the largest subgroup of malignant disease in this age group (1). Childhood diabetes, type 1, or insulin-dependent diabetes is the consequence of an immune-mediated destruction of the insulin-producing beta cells of the pancreas (2) and is a major contributor to the burden of chronic disease in the childhood population of more developed countries (3).

Although these conditions appear to be biologically unconnected, there are common threads in both their epidemiology and etiology based on evidence that environmental exposures are likely to have a strong influence on disease occurrence. The concordance rate for twins aged 2–6 years with acute lymphoblastic leukemia is reported as 5 percent (4) and, for monozygotic twins with diabetes, type 1, concordance estimates range from 13 percent to 50 percent (5, 6), the latter higher rates possibly reflecting a stronger attributable genetic susceptibility that is more clearly defined for diabetes, type 1 (7), than for acute lymphoblastic leukemia (8). Slowly rising incidence rates of both acute lymphoblastic leukemia (9, 10) and diabetes (11, 12) have been observed over recent decades in many Western countries. The consistency of these increases across different
populations and the high quality of the registers make it unlikely that improved diagnostic accuracy can account for these changes. Additionally, for diabetes, type 1, increasing rates have been seen in childhood populations migrating from areas of low to high incidence (13), rises which have persisted into later generations (14).

A common and recurring theme in the etiology of acute lymphoblastic leukemia and diabetes, type 1, has been the possible involvement of infections (4, 15). One explanation involves a paradox of development in Westernized societies: that common infections in infancy may protect from later disease by appropriate modulation of the naïve immune system (16). Contrariwise, in the absence of such early exposures, later infection, for example with social mixing of children, may precipitate abnormal immune reactions and disease. Inherited genetic factors may also influence susceptibility. This scenario has been referred to as the “hygiene hypothesis” in the context of both diabetes, type 1, and allergy/asthma (17) and as the “delayed infection hypothesis” for childhood acute lymphoblastic leukemia (18).

Although the underlying immunologic pathologies of childhood allergies and diabetes, type 1, are dissimilar (T-helper 2 vs. T-helper 1 T-cell overactivity), their shared environmental associations are reflected in international correlations between their respective incidence rates (19). The presence of two population-based registers of childhood diabetes (20) and cancers (21) from a defined geographic area of the United Kingdom has provided the unique opportunity of simultaneously investigating distributions of these conditions in small geographic areas.

Population mixing has received attention in geographic and infectious disease epidemiology as a potential proxy measure of exposure to infections. This measures the degree of population migration and the extent to which incoming migrants originate from different areas. Localized excesses of childhood leukemia have been associated with “rural population mixing” occurring in discrete areas and under certain specific circumstances, such as extreme population influxes (22–24). This study uses a reproducible measure of “population mixing” applied to the incidence of a specific subtype of childhood leukemia and diabetes, in a representative childhood population. The effect of population mixing or mobility has been examined separately for both diabetes (25) and leukemia (22–24, 26–29), and some early work looked at the epidemiology of both conditions (30).

Both acute lymphoblastic leukemia and diabetes have similar large-scale distributions (19), and our study examines whether this is reflected in small geographic areas. We aimed to explore 1) whether the occurrences of both diseases were spatially associated with each other in small areas and 2) whether area-based sociodemographic risk factors attenuate the degree of spatial correlation in the incidence between these conditions. Given the sparsely distributed case data, we adopted a Bayesian approach to examine smoothed standardized incidence ratios and modeled their joint spatial association using a bivariate random effects model. The Bayesian model risk estimates were contrasted with those derived from the classical “frequentist” approach.

**MATERIALS AND METHODS**

**Data collection**

We extracted data on children aged less than 15 years and diagnosed with acute lymphoblastic leukemia or diabetes, type 1, between 1986 and 1998 from two population-based disease registers covering the former Yorkshire Regional Health Authority in the north of the United Kingdom (20, 21). The registers have full ethical approval and carry out active ascertainment checks from multiple sources, with diagnoses confirmed from hospital records (20, 31). Children were diagnosed as having diabetes, type 1, according to World Health Organization guidelines (32), while acute lymphoblastic leukemia was confirmed histopathologically for 96 percent of the children. The registers cover a geographic area of 12,000 km² and a childhood population of 700,000. We limited the case series to a period centered at the time of the 1991 national census in the United Kingdom to ensure the inclusion of relevant sociodemographic denominator data.

Patients’ addresses and postcodes at the time of diagnosis were validated using QuickAddress software (Experian, Ltd., London, United Kingdom) (www.qas.com/uk/products/consumer/cleaning.asp) and linked to one of 532 electoral wards in existence in Yorkshire at the time of the 1991 census. These small geographic areas have a median childhood population count of 700 (range: 0–5,900). Population estimates from the 1991 census were used to calculate age- and sex-standardized incidence rates (the 1991 census; Crown Copyright; purchased by the Economic and Social Research Council, Swindon, United Kingdom).

**Analytical models**

We initially examined the distribution of cases for both diseases across wards by calculating and mapping spatially (locally) smoothed estimates of the standardized incidence ratios, using the approach of Besag et al. (33). Generally, one might assume that areas in close proximity to one another may have similar standardized incidence ratios, leading to spatially (locally) structured variation in relative risks. Standardized incidence ratio estimates by ward are presented as graphs (figures 1 and 2) according to the following ranges based on a previous geographic mapping analysis carried out in the region (34): <85, 85–94, 95–104, 105–114, and ≥115.

We then implemented a multivariate spatial model, proposed by Leyland et al. (35), within a Bayesian framework that can differentiate between the relative contribution of the spatial part and the degree to which the disease rates exhibit extra-Poisson variation. This approach helps to improve the efficiency of parameter estimates from one disease by incorporating information on a second disease and crucially allows estimation and inference to be made about the correlation between the risks of the two diseases.

**Statistical analysis**

For each disease, a Poisson regression model was fitted to the observed numbers of cases in each ward using the log of the number of expected cases as the offset derived from...
age- and sex-specific incidence rates for Yorkshire between 1986 and 1998. This was implemented within a classical framework. We compared separately the risk for both diseases from three sociodemographic factors previously linked to disease onset. These included the following: 1) population mixing, measured using the Shannon Index fully described elsewhere (25, 26), reflecting the diversity of origins of incomers into each ward and calculated for the childhood (0–14 years) population; 2) person-based childhood population density (25), which is a population-weighted average of population density (persons per hectare), an appropriate index for investigating infectious etiology as it reflects the density level at which a typical person lives (this is preferable to using an urban/rural indicator for a study testing an infectious hypothesis); and 3) deprivation, measured using the Townsend score (36), standardized to all wards in Yorkshire with the following variables contributing to the index: unemployment, household overcrowding, car ownership, and housing tenure.

For comparison with other studies (25–29) and ease of interpretation, incidence rate ratios and 95 percent confidence (table 2)/credible (tables 3 and 4) intervals are presented and can be interpreted as relative risks. These are presented according to categories based on the rankings of the values, separately for each factor, across all wards. They were defined as follows:

- Population mixing: <10th percentile, 10th–90th percentiles (referent group), and >90th percentile. This grouping enables effects to be detected at the extremes of the range, as there is little variation in the value of the Shannon Index for the majority of wards.
- Population density: three equal groups of wards (lowest density taken as the referent group).
- Deprivation: five equal groups of wards (least deprived taken as the referent group).

All three variables were included separately in the Poisson model; no other confounding factors were added to this initial model. The effect on standardized incidence ratios of including all three covariates was then assessed. We were concerned about the sparseness of data across wards for acute lymphoblastic leukemia and therefore examined the
goodness of fit for the Poisson models using a prediction model. The observed counts were compared graphically by the degree of symmetry with those predicted from a simulated data set based on the model incorporating all three terms. We also allowed for any potential overdispersion by incorporating extra-Poisson variation into the classical analysis using negative binomial regression. All classical statistical analyses were performed using Stata statistical software (Stata Corporation, College Station, Texas).

Finally, we modeled the two disease counts jointly, examining the effects from each covariate using two Bayesian forms of spatial smoothing. This was done using spatially unstructured and structured random effects (also known as heterogeneity and spatial random effects, respectively). The heterogeneity random effects introduce extra-Poisson variation into the model that is due to omitted covariates; the spatial random effects control for unmeasured spatial covariates. Leyland et al. (35) explained how the proportion of variation attributable to the heterogeneity and spatial random effects is calculated, the latter being scaled by the modal count of the number of neighbors for any ward (35). The proportion of variation due to spatial effects is provided, along with interval estimates, in the tables next to the random effects estimates, with the modal count taken to be 5.45: this was calculated empirically at each iteration of the Markov chain Monte Carlo simulation (Appendix 2).

Suppose $O_{i1}$ and $O_{i2}$ represent the observed disease counts for acute lymphoblastic leukemia and diabetes mellitus, type 1, in the $i$th electoral ward, with expected counts $E_{i1}$ and $E_{i2}$, respectively. The observed count outcomes are modeled as Poisson $O_{ih} \sim \text{Poi}(E_{ih}R_{ih})$; $i = 1, \ldots, 532$; and $h = 1, 2$ (37), where $R_{ih}$ is the relative risk of disease $h$ in ward $i$. The maximum likelihood estimate of the relative risk of disease $h$ in area $i$ is the usual standardized morbidity/incidence ratio, $\hat{R}_{ih} = O_{ih}/E_{ih}$. However, it is now common to introduce area-level covariates while accounting for both unstructured heterogeneity and spatial dependent effects in relative risks when modeling small-area count data. With the model of Besag et al. (33), the logarithm of the disease-specific relative risk is modeled as:

$$
\log(R_{i1}) = \alpha_{01} + \beta_{11}x_i + u_{i1} + v_{i1}
$$

$$
\log(R_{i2}) = \alpha_{02} + \beta_{21}x_i + u_{i2} + v_{i2}
$$

(1)
where

- $\beta_0h$ is an intercept of the log relative risk for disease $h$ in ward $i$,
- $x_i$ is a covariate vector with the corresponding parameter $\beta_h$,
- $u_{1i}$ and $u_{2i}$ are the independent unstructured random effects (representing global smoothing), and
- $v_{1i}$ and $v_{2i}$ are the spatially structured random effects (representing local smoothing).

The above models are sometimes called “convolution” models. An exchangeable normal prior for the unstructured random effects is easily specified as $u_{1h} \sim N(0, \sigma^2_{u1})$. However, for the spatially correlated random effects, we adopted the approach by Leyland et al. (35) in assuming that the spatial random effect $v_{ih}$ arises from a combination of independent random effects errors $e_{ih}$. Using the adjacency matrix in the parameterization of the Wishart distribution for the scale matrix, as a form of sensitivity analysis (1), and

$$v_{ih} = \sum_{j \in \Theta_i} e_{ij} / n_i,$$

where $\Theta_i$ is the set of wards sharing a common boundary to ward $i$, and $n_i$ is the number of neighbors in the set. The $e_{ij}$ are regarded as the effects of area $j$ on its neighbors, and the summation in equation 2 gives the spatial effect $v_{ih}$ for disease $h$ in area $i$. For simplicity, we take the $e_{ij}$ as normal random variables, $e_{ij} \sim N(0, \kappa^2_{ij})$. This formulation gives the total variation in the logarithm of relative risk for an area as the sum of variances in the heterogeneity and spatial effects and is dependent on the number of neighbors of the area, that is:

$$\text{Variance}(\Sigma_{hi}) = \sigma^2_{u1} + \kappa^2_{ij} / n_i.$$  

In our main model of the paper, we took the four random effect terms $u_{1i}$, $u_{2i}$, $e_{1i}$, and $e_{2i}$ to have arisen from a multivariate normal distribution with zero mean vector and covariance matrix $\Sigma$ (35, 38), though we chose to adopt a hierarchical Bayesian approach (39) rather than use an iterative generalized least-squares estimation in a multilevel context. In our Bayesian analysis of the model, all fixed effect parameters were given noninformative proper normal (0, 1,000) prior distributions. However, for the covariance matrix $\Sigma$ for the four random effects, we used both informative and noninformative specifications for the scale matrix in the parameterization of the Wishart distribution for $\Sigma^{-1}$, the precision matrix, as a form of sensitivity analysis (Appendix 1).

Posterior estimation of all the model parameters was done using the Gibbs sampling algorithm (40) implemented in the software package WinBUGS (41). The model so far defined provided only direct simulated values for the variance/covariance between the heterogeneity effects but not between the spatial effects ($v_{1i}$ and $v_{2i}$) or the total risk variation ($u_{1i} + v_{1i}$) and ($u_{2i} + v_{2i}$). These were computed empirically at each iteration of the Gibbs sampler. For each model considered, three parallel Gibbs sampler chains from independent starting positions were run for 50,000 iterations. All fixed effects and covariance parameters were monitored for convergence. Trace plots of sample values of each of these parameters showed that they were converging to the same distribution. Using Gelman-Rubin reduction factors (42), we formally assessed convergence of the three chains that was estimated to be near 1.0 by 15,000 iterations. For posterior inference, we used a combined sample of the remaining 35,000 iterations. Finally, the effect on the degree of spatial correlation between both diseases was examined before and after allowing for each sociodemographic factor previously linked to the spatial distribution of disease incidence. A copy of the WinBUGS code is provided in Appendix 2.

**RESULTS**

Among children aged 0–14 years, we identified 299 who were diagnosed with acute lymphoblastic leukemia and 1,551 who were diagnosed with diabetes, type 1. Table 1 describes the distribution of patients by gender and 5-year age groups, showing a slight excess of males for both disease groups and different age incidence peaks (0–4 years for acute lymphoblastic leukemia and 10–14 years for diabetes). Lower rates of acute lymphoblastic leukemia and diabetes were seen in the more urban county of West Yorkshire than in other parts of the region, while higher rates of both conditions were observed in the more rural county of North Yorkshire. The southeastern part of the region below the Humber estuary showed large fluctuations in standardized incidence ratios for both disease groups (figures 1 and 2). The mean numbers of cases distributed across all 532 wards in Yorkshire were 0.6 (range: 0–5) and 2.9 (range: 0–17) for acute lymphoblastic leukemia and diabetes overall; 0.3 (range: 0–4) and 0.7 (range: 0–7) for acute lymphoblastic leukemia and 10–14 years for diabetes were seen in the more urban county of West Yorkshire than in other parts of the region, while higher rates of both conditions were observed in the more rural county of North Yorkshire. The southeastern part of the region below the Humber estuary showed large fluctuations in standardized incidence ratios for both disease groups (figures 1 and 2). The mean numbers of cases distributed across all 532 wards in Yorkshire were 0.6 (range: 0–5) and 2.9 (range: 0–17) for acute lymphoblastic leukemia and diabetes overall; 0.3 (range: 0–4) and 0.7 (range: 0–7) for acute lymphoblastic leukemia and diabetes overall; 0.3 (range: 0–4) and 0.7 (range: 0–7) for cases aged 0–4 years; and 0.5 (range: 0–3) and 2.2 (range: 0–15) for cases aged 5–14 years.

**Classical approach**

Table 2 shows the unadjusted and adjusted incidence rate ratios for each covariate and disease separately. Higher rates
of diabetes and acute lymphoblastic leukemia were present in areas of low population mixing, and this effect remained after adjusting for the other two covariates. In areas with very high population mixing, significantly lower rates of acute lymphoblastic leukemia were observed, although no similar association in incidence was seen for diabetes.

An inverse association was present for population density for each condition, with lower rates associated with higher levels of density. However, once the effects from population mixing and deprivation were taken into consideration, the association with population density disappeared for diabetes and was reversed for acute lymphoblastic leukemia. There was some evidence of a negative association between deprivation and diabetes, with lower rates observed in more deprived areas. There was no systematic relation between deprivation and acute lymphoblastic leukemia.

To investigate whether age was an influencing factor, we analyzed the data by subdividing the diseases into groups aged 0–4 and 5–14 years. The results (not presented), although based on small numbers, showed findings similar to those of the original analysis: Population mixing was associated with acute lymphoblastic leukemia for children aged 0–4 years.

### TABLE 2. Incidence rate ratios and 95% confidence intervals derived using a classical approach for each disease modeled separately, Yorkshire, United Kingdom, 1986–1998

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Unadjusted estimates</th>
<th>Adjusted estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Incidence</td>
<td>95% confidence</td>
</tr>
<tr>
<td></td>
<td>rate ratio</td>
<td>interval</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Diabetes mellitus, type 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population density</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Medium</td>
<td>0.92</td>
<td>0.77, 1.09</td>
</tr>
<tr>
<td>High</td>
<td>0.75</td>
<td>0.63, 0.88</td>
</tr>
<tr>
<td>Population mixing (Shannon Index)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10th percentile</td>
<td>1.50</td>
<td>1.11, 2.02</td>
</tr>
<tr>
<td>10th–90th percentile</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>&gt;90th percentile</td>
<td>1.01</td>
<td>0.88, 1.16</td>
</tr>
<tr>
<td>Townsend deprivation score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (least deprived)</td>
<td>1.00</td>
<td>$\rho_{\text{trend}} = 0.01$</td>
</tr>
<tr>
<td>2</td>
<td>1.05</td>
<td>0.86, 1.29</td>
</tr>
<tr>
<td>3</td>
<td>1.03</td>
<td>0.84, 1.26</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>0.71, 1.04</td>
</tr>
<tr>
<td>5 (most deprived)</td>
<td>0.76</td>
<td>0.63, 0.91</td>
</tr>
<tr>
<td><strong>Acute lymphoblastic leukemia</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population density</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Medium</td>
<td>0.92</td>
<td>0.62, 1.38</td>
</tr>
<tr>
<td>High</td>
<td>0.83</td>
<td>0.56, 1.23</td>
</tr>
<tr>
<td>Population mixing (Shannon Index)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10th percentile</td>
<td>1.36</td>
<td>0.67, 2.74</td>
</tr>
<tr>
<td>10th–90th percentile</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>&gt;90th percentile</td>
<td>0.74</td>
<td>0.54, 0.99</td>
</tr>
<tr>
<td>Townsend deprivation score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (least deprived)</td>
<td>1.00</td>
<td>$\rho_{\text{trend}} = 0.14$</td>
</tr>
<tr>
<td>2</td>
<td>1.28</td>
<td>0.78, 2.11</td>
</tr>
<tr>
<td>3</td>
<td>1.24</td>
<td>0.77, 2.00</td>
</tr>
<tr>
<td>4</td>
<td>1.11</td>
<td>0.70, 1.77</td>
</tr>
<tr>
<td>5 (most deprived)</td>
<td>0.92</td>
<td>0.59, 1.44</td>
</tr>
</tbody>
</table>

* Mutually adjusted for the other two covariates.
both 0–4 and 5–14 years, as was deprivation for diabetic patients aged 0–4 years, although not significantly so for those aged 5–14 years.

Although all three variables included in the model were positively correlated, we saw no evidence of multicollinearity through inflated standard errors of the model estimates. Variance inflation factors, calculated manually in Stata software using linear regression for each covariate, were all below 2.5. Population mixing also exhibited the least degree of correlation of any of the covariates; for example, areas with high levels of mixing had an equal number of areas in the medium and highest population density categories. A graphical comparison between the observed counts and predicted counts from a simulated model showed good symmetry for each disease.

**Bayesian approach**

We modeled the effects of both disease counts together as a bivariate outcome, comparing the variation attributable to spatial and nonspatial (heterogeneity) random effects before and after adjusting for sociodemographic covariates. Assuming dependent random effects between diseases with no adjustment for covariates, we found that on average 50 percent of the variation occurred through the spatial component for diabetes and acute lymphoblastic leukemia, with the remainder occurring through effects of heterogeneity. We found a modest degree of positive spatial correlation between diseases of 0.33 (95 percent credible interval: −0.20, 0.74).

Compared with the classical univariate model (table 2), the parameter estimates remained largely the same after allowing for dependent random effects and the contribution of each sociodemographic covariate on its own (table 3). After accounting separately for population mixing, population density, and deprivation, we found that the spatial correlation between diseases fell from 0.33 to 0.18 (95 percent credible interval: −0.62, 0.82), 0.14 (95 percent credible interval: −0.50, 0.78), and 0.06 (95 percent credible interval: −0.59, 0.69), respectively. This corresponded to an approximate reduction of 45 percent, 58 percent, and 82 percent, respectively. Adding the spatial component of variation into a model already containing the heterogeneity part significantly improved model fit using the Deviance Information Criterion (Appendix 1).

After adjustment for all three covariates simultaneously (table 4), the spatial correlation fell to 0.12 (95 percent credible interval: −0.63, 0.73). The parameter estimates were similar to the adjusted incidence rate ratios presented in table 2 from the classical approach. There was little change in the proportion of variation attributable to spatial and nonspatial effects among models 1–4: 45–60 percent was due to the spatial component for both diseases. We performed a sensitivity analysis based on different prior specifications of the scale matrix (Appendix 1). The fixed-effect estimates and random-effect variances remained largely the same, although the unadjusted spatial correlation fell to around 0.10.

A small positive spatial correlation of 0.10 (95 percent credible interval: −0.55, 0.78) and 0.13 (95 percent credible interval: −0.32, 0.55) was also observed between the unstructured random-effect components ($u_1$ and $u_2$) and the overall residuals ($v_1 + v_1$ and $u_2 + v_2$), again suggesting that a modest correlation existed between the “net” risks of each disease not explained by the three covariates in the model.

Finally, we examined the effect of estimating a single coefficient for each disease, rather than two. We found that they were similar to those presented in table 4 for population mixing, deprivation, and population density (results not shown) and were typically a weighted average of the estimates from the two separate diseases.

**DISCUSSION**

The environmental causes of acute lymphoblastic leukemia and diabetes, type 1, have received considerable attention but continue to elude unambiguous identification. Despite the reported similarities in their epidemiology, this matter had only been addressed formally in a study comparing the international correlation of their respective incidence rates, which found a strong positive association of 0.53 (95 percent credible interval: 0.36, 0.72) (19). We performed an analogous study investigating whether the risk for both diseases was similar across small geographic areas in the north of the United Kingdom, showing a positive and statistically nonsignificant joint spatial correlation of 0.33 (95 percent credible interval: −0.20, 0.74). This finding was also illustrated by mapping the spatial distribution of each disease: Generally rates were lower in the more populated county of West Yorkshire than in other areas and higher in the less populated county of North Yorkshire.

The spatial correlation was reduced toward zero once we allowed for the effects of deprivation, whereas a much smaller effect was observed on the size of the correlation after adjustment for population mixing and population density. The finding that population mixing explained a smaller proportion of the joint spatial association between diseases than deprivation or population density did is a new observation. This is not necessarily inconsistent with previously reported studies (22–29). One possible explanation for this is that population mixing may have had a strong but erratic effect across wards, yielding an association with disease occurrence in the classical approach, whereas deprivation/population density had a modest but consistent effect across small areas, which disappeared when considering the joint spatial correlation.

There was considerable heterogeneity in incidence across wards that accounted for half of the observed variation in disease outcome for each condition, and it may be unsurprising that the data only partially suggested that the risk of acute lymphoblastic leukemia and diabetes was associated within wards. We cannot exclude the possibility that the reduction in the degree of spatial correlation at the small geographic level compared with the international level was caused by the modifiable areal unit problem (43), whereby the correlation typically becomes stronger the greater the level of geographic aggregation.
### TABLE 3. Fixed and random effects estimates (median and 95% credible interval) from a Bayesian bivariate model with dependent errors, Yorkshire, United Kingdom, 1986–1998

<table>
<thead>
<tr>
<th></th>
<th>Model 1: population density</th>
<th>Model 2: population mixing</th>
<th>Model 3: deprivation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diabetes</td>
<td>Leukemia</td>
<td>Diabetes</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>95% credible interval</td>
<td>Median</td>
</tr>
<tr>
<td>Random effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heterogeneity</td>
<td>0.02</td>
<td>0.00, 0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Spatial</td>
<td>0.01</td>
<td>0.00, 0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>Proportion of total variation (%)</td>
<td>45</td>
<td>13, 79</td>
<td>59</td>
</tr>
<tr>
<td>Spatial correlation</td>
<td>0.14</td>
<td>−0.50, 0.78</td>
<td>0.18</td>
</tr>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population density</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>1.00</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Medium</td>
<td>0.93</td>
<td>0.77, 1.09</td>
<td>0.99</td>
</tr>
<tr>
<td>High</td>
<td>0.76</td>
<td>0.64, 0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>Population mixing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;10th percentile</td>
<td>1.47</td>
<td>1.05, 1.94</td>
<td>1.34</td>
</tr>
<tr>
<td>10th–90th percentile</td>
<td>1.00</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>&gt;90th percentile</td>
<td>1.03</td>
<td>0.90, 1.20</td>
<td>0.74</td>
</tr>
<tr>
<td>Townsend deprivation score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (least deprived)</td>
<td>1.00</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>2</td>
<td>1.06</td>
<td>0.85, 1.32</td>
<td>1.37</td>
</tr>
<tr>
<td>3</td>
<td>1.04</td>
<td>0.85, 1.26</td>
<td>1.30</td>
</tr>
<tr>
<td>4</td>
<td>0.86</td>
<td>0.71, 1.04</td>
<td>1.17</td>
</tr>
<tr>
<td>5 (most deprived)</td>
<td>0.76</td>
<td>0.64, 0.91</td>
<td>0.98</td>
</tr>
</tbody>
</table>
Comparison with previous findings

Ecologic studies have addressed the role of infections with the onset of diabetes and acute lymphoblastic leuke-mia by the use of proxy measures of exposure to “com-munity infections” such as differences in population density (44, 45) and population mixing (22–29). If the hygiene hypothesis was operating, then areas with low lev-els of population mixing would have higher rates of acute lymphoblastic leukemia and diabetes, and this has been demonstrated to be the case even after adjustment for pop-ulation density and deprivation (23, 25, 28). This pattern was reflected in our results for both conditions, where we observed similar effect estimates using either a classical or a Bayesian approach. There was no clear association with either population density or deprivation once we adjusted for each of the other two factors. Previous inverse associ-ations between diabetes/acute lymphoblastic leukemia and population density may therefore have been accounted for if population mixing and deprivation were already consid-ered in the modeling process.

Unusually high levels of population mixing occurring in specific settings have been shown to be associated with an increased risk of acute lymphoblastic leukemia (22–24). The increase may be the result of susceptible individuals who have an abnormal response to a common infection that orig-inates from incoming migrants (22–24, 27). However, our findings showing a reduced risk of acute lymphoblastic leukemia linked to high levels of mixing are in contrast to but not directly comparable with studies of small rural populations in the United Kingdom (22, 24) and extreme population growth in urbanized Hong Kong (23). Our findings for acute lym-phoblastic leukemia are similar to those of a recent inde-pendent study in the United Kingdom (29) that used the Shannon Index applied at the same geographic scale.

Strengths and weaknesses of a Bayesian ecologic approach

All ecologic analysis should be interpreted with some degree of caution because estimated fixed effects from exposure to infections at the community level may not di rectly resemble those at the individual level. More importantly, in the modeling of disease outcomes for rare conditions, such as childhood cancer and childhood diabetes, Poisson regression is prone to overdispersion where the table:
sample variance is higher than the mean. This can often occur when counts of disease cluster within certain geographic areas, and it may be difficult to ignore if we are interested in a common etiologic factor. The dearth of disease counts across wards, particularly for acute lymphoblastic leukemia in our data set, is difficult to overcome when limited to a predefined geographic scale in which to incorporate sociodemographic data, such as population mixing. Further, in view of the sparse nature of the data across wards, it was impossible to examine the spatial correlation between diseases in a classical manner without smoothing incidence rates across wards using Bayesian methods. It was for all these reasons that we chose to adopt a Bayesian smoothing approach in modeling the spatial correlation between diseases, and this is the method we feel is most appropriate for these types of data.

We were also able to account for the degree of variation in incidence accounted for by heterogeneity and spatial effects. The robust nature of the data was signified by the fact that the parameter estimates for all three fixed effects using a classical approach were almost identical to the Bayesian results, and the ability to specify prior knowledge made no difference in the results. In order to exclude any unknown ecologic confounding, we found no association between population mixing and childhood nephrotic syndrome, a disease without an infectious etiology, using data from the same period and geographic area (results not shown).

Nonetheless, our methodology can be easily applied to other studies investigating the simultaneous occurrence of disease incidence or mortality. In studies investigating rare conditions, such as childhood cancer and diabetes, it is imperative that the identification of case information is derived from sources with high levels of ascertainment; this was a key feature of our study with access to two high-quality, population-based disease registries in Yorkshire (20, 21).

**Summary**

Although childhood acute lymphoblastic leukemia and diabetes display strongly correlated incidence rates at the national level, this association is weakened when investigating the distribution across small geographic areas. Deprivation and, to a lesser degree, population density appear to explain more of the spatial correlation between the diseases than population mixing does. The parallels in the descriptive epidemiology of diabetes, type 1, and childhood acute lymphoblastic leukemia and the role of deprivation/population density/population mixing suggest that a joint investigation of common causal pathways and underlying genetic susceptibility in individuals would be informative. First, however, our findings need to be replicated in other populations, and we are planning to conduct more extensive geographic analyses for other countries.

**ACKNOWLEDGMENTS**

The costs of running the cancer and diabetes registers were supported by the Candlelighters Trust and the Leeds Teaching Hospitals National Health Service Trust. This diabetes registry work was also undertaken by the University of Leeds, which received funding from the Department of Health.

The authors thank the Office for National Statistics for the provision of population data and special migration statistics. They also acknowledge the earlier work of Dr. Anthony Staines that laid the groundwork for the study and the fruitful discussions with Dr. Ian Lewis. The authors are grateful to Margaret Buchan, Carolyn Stephenson, and Sheila Jones for data collection and for the cooperation of all the pediatricians, pediatric oncologists, physicians, diabetes specialist nurses, and general practitioners in Yorkshire.

The views expressed in this article are those of the authors and not necessarily those of the Department of Health.

**REFERENCES**


APPENDIX 1

A standard approach to choosing a prior distribution for the covariance matrix $\Sigma$ is to specify a Wishart($\nu$, $\Omega$) distribution on its inverse $\Sigma^{-1}$, which is the precision matrix (akin to using a gamma distribution on the inverse of a variance in univariate situations). The parameters of the Wishart($\nu$, $\Omega$) distribution are the degrees of freedom $\nu$, which control the shape of the distribution and the scale parameter $\Omega$ of the same dimension as $\Sigma$. The WinBUGS parameterization of the Wishart($\nu$, $\Omega$) distribution gives the expected value of the precision matrix $\Sigma^{-1}$ as $\mathbb{E}(\Sigma^{-1}) = \nu\Omega^{-1}$; thus, $\Omega/\nu$ is a prior guess at the covariance matrix $\Sigma$. As with other distributions dependent on degrees of freedom, the distribution has longer tails for lower values of $\nu$. In order to avoid unnecessary convergence problems, we set $\nu = 4$, resulting in a minimal informative prior on $\Sigma$.

We used population mixing to elicit plausible values for the overall variances of the random effects for each disease. From previous studies (28, 29), the univariate relative risk (RR) for acute lymphoblastic leukemia ranged from 0.65 to 1.50. If we assume that this translates to a $1.50/0.65 = 2.3$-fold variation in risk between two areas at the extreme ends of the 95 percent reference range,
then 95 percent of areas would have their log risk in a range with width equal to $2 \times 1.96 \sqrt{\text{variance}}$. Mathematically, this is as follows:

$$\log \left( \frac{RR(0.975)}{RR(0.025)} \right) = \log(2.3) = 2 \times 1.96 \sqrt{\text{variance}}$$

This solves to give a variance of 0.0456 for the overall area effects with respect to acute lymphoblastic leukemia. Similarly, for diabetes, the univariate risk ranged from 1.0 to 1.5 (25), giving a 1.5-fold variation in risk between two areas at opposite ends of the reference range; this equates to a prior overall variance estimate of 0.0106. We also assumed that these prior variances split equally between the unstructured and spatial random effects. With a modal number of 5.45 for the number of neighbors for each area and using equation 3, we obtained 0.0228, 0.0053, 0.1243, and 0.0289 as plausible prior values for $\sigma_1^2$, $\sigma_2^2$, $\kappa_1^2$, and $\kappa_2^2$, respectively. These prior values were then scaled up by a factor of $v = 4$ and entered diagonally on the scale matrix $\Omega$ in the Wishart($v$, $\Omega$) distribution for $\Sigma^{-1}$. The off-diagonal entries were set to zero, assuming that there was no correlation among the four random effects, at least between the two unstructured random effects. We used other prior values of the variances, first, by assuming that these were large (25, 100, 25, 100) and, second, by setting them all equal to 1 on the diagonal of the matrix $\Omega$.

For model fit, we used the Deviance Information Criterion developed by Spiegelhalter et al. (46). The Deviance Information Criterion is defined as the sum of the posterior mean of the deviance (which measures model fit) and the effective number of parameters in the model (which measures model complexity). The Deviance Information Criterion is particularly useful in situations involving complex hierarchical models in which it is difficult to compute the actual number of parameters being used. It is a generalization of the Akaike Information Criterion and works on a similar principle.

**APPENDIX 2**

The WinBUGS code for our main model (bivariate disease outcome counts, with all fixed covariate effects and a multivariate normal prior for the four random effects to capture interdependence between the outcomes) is provided below for reference.

Model

```r
{ 
  for (i in 1 : N_Wards) { # likelihoods for the observed 
    ALL and T1D disease counts 
    O[i,1] <- obsALL[i]; O[i,2] <- obsT1D[i]; 
    E[i,1] <- expALL[i]; E[i,2] <- expT1D[i]; 
    for (j in 1:2) { R[j,i] <- 100*mu[i,j]/E[i,j] 
    O[i,j] ~ dpois(mu[i,j]) 
    log(mu[i,j]) <-log(E[i,j]) + A[j] + b[j,1]*pmix1[i]+b[j,2]*pmix2[i]+ 
    b[j,3]*town1[i]+b[j,4]*town2[i]+b[j,5]*town3[i]+ 
    b[j,6]*town4[i]+ b[j,7]*pdens1[i]+b[j,8]*pdens2[i]+ 
    u[i,j]+ v[j]
  } 
  # Error priors 
  u[i,1:4] ~ dnorm(nought[1:4],P[1:4, 1:4]) 
  v[i,1] ~ <- mean(e.ALL[C[i] + 1 : C[i + 1]] 
  v[i,2] ~ <- mean(e.T1D[C[i] + 1 : C[i + 1]]) 
  # Sum of heterogeneity and spatial error effects, i.e., 
  total unexplained variation 
  sum_effects[i,1] <-u[i,1]+v[i,1] 
  sum_effects[i,2]<-u[i,2]+v[i,2] 
  # Quantities needed for spatial variances/covariances 
  for (i in 1:N) 
    diffv1_v2[i] <- (v[i,1]-mean(v[1:2]))*(v[i,2]-mean(v[1:2])) 
  # Quantities needed for total unexplained variances/ 
  # covariances 
  difftheta1_2[i] <- (sum_effects[i,1]-mean(sum_effects 
  [1:2]))*(sum_effects[i,2]-mean(sum_effects[2:4])) 
  } 
  # Empirical variance-covariance estimates/Correlation 
  estimates 
  for (j in 1:2) { var.spatial[j] <- sd(v[j])*sd(v[j]) 
    # Total variation 
    var.total[j] <- sd(u[j])*sd(u[j]) + sd(v[j])*sd(v[j]) 
    var.sum_effects[j] <- sd(sum_effects[j])*sd 
      (sum_effects[j]) 
  } 
  corr.spatial1_2 <- sum(diffv1_v2[i])/(N-1)*sd(v[1])* 
    sd(v[2]) 
  corr.sum_effects1_2 <- (sum(difftheta1_2[i]))/(N-1)*sd 
    (sum_effects[1])*sd(sum_effects[2]) 
  } 
  # Proportion of spatial variance 
  prop_var_spat1 <- sd(v[1])*sd(v[1]) / 
    (sd(u[1])*sd(u[1]) + sd(v[1])*sd(v[1])) 
  prop_var_spat2 <- sd(v[2])*sd(v[2]) / 
    (sd(u[2])*sd(u[2]) + sd(v[2])*sd(v[2])) 
  # Prior for the intercepts, fixed effects 
  for (j in 1:2) { A[j] ~ dnorm(0,0.001) 
  for (k in 1:8) { b[j,k] ~ dnorm(0,0.001) } 
  } 
  for (j in 1:2) { var_mean[j] <- dim(v[j])/exp(b[j,k]) } 
  # Prior for mean and precision matrix in the 4-variate 
  normal distribution 
  for (i in 1:4) { nought[j] <- 0 
  } 
  # Computing the dispersion/correlation matrices for the 
  4-variate normal distribution 
  for (i in 1:4) { for (j in 1:4) { R[i,j] ~ V[i,j]/sqrt(V[i,i]* 
    V[j,j]) } 
  # Specification of the scale matrix in the Wishart prior 
  Q[1,1] ~ 0.0228*4; Q[2,2]<-0.0053*4; Q[3,3] ~ 
    0.1243*4; Q[4,4]<-0.0289*4; 
  Q[1,2] <- 0.0; Q[1,3]<-0.0; Q[1,4] <- 0.0; 
  Q[2,1] <- 0.0; Q[2,3]<-0.0; Q[2,4] <- 0.0; 
  Q[3,1] <- 0.0; Q[3,2]<-0.0; Q[3,4] <- 0.0; 
  Q[4,1] <- 0.0; Q[4,2]<-0.0; Q[4,3] <- 0.0; 
```
Abbreviations in Appendix 2: ALL, acute lymphoblastic leukemia; T1D, type 1 diabetes.

Data node map contains a set of adjacent wards for each ward, and C contains a cumulative total count of the number of neighbors for each ward; for example, if wards 1 and 2 have five and seven neighbors, then \( C = c(0, 5, 12, \ldots, \text{All Neighbors}) \).