Geo-epidemiology and environmental co-variate mapping of primary biliary cholangitis and primary sclerosing cholangitis

Jessica Katharine Dyson, Alasdair Blain, Mark David Foster Shirley, Mark Hudson, Steven Rushton, David Emrys Jeffreys Jones

Table of contents

Supplementary case-finding and search strategies

The case-finding methodologies used were based on the principles proposed by Metcalf et al for the conduct of rigorous epidemiological studies.[1] The study area includes the counties of County Durham, Cumbria, Northumberland, Tees Valley and Tyne and Wear. Postcode areas included in the study were CA, DH, DL, NE, SR and TS. Newcastle upon Tyne Hospitals NHS Foundation Trust is the regional tertiary hepatology referral unit in the study area so some patients across the region receive clinical care for their liver disease in Newcastle as well as their local hospital. These patients were only included in the database once.

To review the potential impact of patient mobility, postcode at diagnosis and postcode at time of study entry were collected where available.

Medical coding

The medical coding department in each NHS Trust was contacted to provide a list of patients with appropriate WHO ICD-10 codes. This identified any patient who had had an inpatient admission for whom these ICD-10 codes appeared on their discharge summary. Details of ICD-10 codes are available at http://www.who.int/classifications/icd/icdonlineversions/en/. There are specific codes for AIH (K75.4) and PBC (K74.3). However, no specific code exists for PSC. K83.0 is the code for cholangitis but is non-specific and significantly limits the use of this methodology for this diagnosis as it generates a huge number of patients with an alternative diagnosis. This is consistent with previous work by Molodecky et al who examined the validity of administrative data for diagnosing PSC in a population-based study. They found that the optimal algorithm included 1 PSC code and 1 IBD code but still only yielded a sensitivity of 56% and positive predictive value (PPV) of 59% and concluded that true PSC cases could not be identified using administrative data.[2] As a result, medical coding data were not used in the identification of PSC patients.

Histopathology

Each histopathology department was contacted and asked to provide a list of patients who had undergone liver biopsy where the formal report included relevant SNOMED codes or if possible, the text could be searched for the word-strings ‘primary biliary cirrhosis, autoimmune hepatitis or primary sclerosing cholangitis’. The appropriate SNOMED codes were determined after discussion with expert histopathologists at Freeman Hospital (***Supplementary table 1***).

Immunology

A targeted search was performed by the immunology laboratories in all hospitals for all positive relevant antibodies: AMA, M2 antibody, SMA, LKM, SLA/LP and LC1. ANA and pANCA were initially included in the search but due to the lack of specificity of these tests for AILD further use of this method was not employed. Repeated tests on the same patient were excluded. AMA results with a titre <1:80 were excluded due to the false positive results seen at this level.

Radiology

A radiology search was performed for all MRCP imaging at each Trust and reports were reviewed manually. Where possible, reports were searched for the ‘word string’, primary sclerosing cholangitis. This also generated reports stating that there was no evidence of PSC but these were manually removed.

Outpatient clinic letters

At four Trusts (Gateshead, North Tees, South Tees, and West Cumberland), clinic letters were kept in Microsoft Word format on the hospital server and it was possible to search these directories for each diagnosis using word strings: PBC, AIH, PSC, primary biliary cirrhosis, autoimmune hepatitis, primary sclerosing cholangitis). In other units, it was possible to access notes electronically to review the clinical scenario and determine whether the patients should be included in the dataset. All clinic letters generated from the dedicated autoimmune liver disease clinic in Newcastle between 2014 and 2016 were reviewed.

Endoscopy reports

At Gateshead, where it was not possible to search using radiology, a search was done of all endoscopic reports with ‘PSC’ in the indication field. This did not yield any previously unidentified patients and this technique was not employed in other centres.

Confirmation of diagnosis

Once patients with a possible diagnosis of AILD were identified, the clinical case notes and electronic records were reviewed by the lead researcher to confirm or refute the diagnosis to make a final decision as to inclusion in the database. If patients were cared for in more than one hospital for their AILD, duplicates were removed from the database. In cases where it was not possible to determine all the variables at diagnosis (e.g. IgG status) a clinical decision was made as to whether to include them. Patients with unknown postcode or who had been diagnosed in postcode areas outside the defined geographical area (n=299) were excluded from further analysis.

Based on the time period in each Trust where data was available using the various relevant case-finding methods, the appropriate time periods for epidemiological modelling of prevalent patients were selected. For PBC this was 2011-2016 and for AIH 2012-2016. This was to ensure that the patient identification was as rigorous and comprehensive as possible for the time period studied. Given that PSC is an even rarer disease (prevalence only 6 per 100,000 population) the complete dataset was included for modelling.

Supplementary modelling methods

Point-based analyses

[3]Analyses were undertaken using the SPLANCS package in 'R' with a spatial range of 20,000 metres in 200 metre steps and with a time range dictated by the first and last years of diagnosis in 1-year steps. 200 simulations were used.

Postcodes were used to allow for collation of different ecological data for use as covariates in the models. Postcode information is available from http://census.edina.ac.uk/pcluts.html. There are currently approximately 1.8 million postcode units in use; each covers up to approximately 100 households. The National Grid Reference positions for the centroids of each postcode were used as a reference point to define where cases occurred. GeoConvert (http://geoconvert.mimas.ac.uk/index.html) and Batch Geocoding (http://www.doogal.co.uk/BatchGeocoding.php) were used to convert postcodes and National Grid references to Easting and Northing.

If there were 30 adults resident at a postcode then the National Grid Reference of the centroid for that postcode was replicated 30 times in the control data set. This meant that there was a National Grid Reference for the residential address of every adult in the study region. The model was adjusted for population size to avoid just finding clusters where the population density was highest.

Spatio-temporal clustering occurs when an excess of cases occurs within a limited geographical area over a limited period of time. K-function analyses examine spatial dependence over a range of spatial scales and can be calculated over any range of time and space between the upper and lower limits of the geography and time domains in a dataset.[4] The envelope size for the K-function was set to avoid edge effects and adjusted to fit the size of the study area (defined as a polygon enclosing all data points).

Area-based analyses

Bayesian area-based analyses (using Conditional Autoregressive (CAR) Models) compares observed with expected cases on the basis of the population size present in an area and the burden of disease in that area. The proportion of the total adult population (from 2010 Census data) that resided within each postcode district was used to estimate the expected cases within that district. Only the adult population was included given that the study population was adults only. The models including the spatial covariates used Bayesian Markov chain Monte Carlo (MC) simulation methods within WinBUGS 1.14. A 10,000 iteration 'burn-in' was followed by a 50,000 iteration sample. For the Deviance Information Criterion (DIC), significance was determined by the distribution of the posterior samples. If 97.5% of the sample was above/below 0 then it was considered as significantly high/low risk.

Average values for stream sediment pH and the heavy metals for each postcode area were calculated. NA values were changed to 0 for appropriate covariates (urbanness, traffic, landfill sites, coal mines, lead mines, sandstone quarries, limestone quarries, Townsend score) and changed to the value from the neighbouring postcode district where there was no sampling rather than a real '0' value (cadmium, arsenic, lead, manganese, iron, stream sediment pH). In the Index of Multiple Deprivation, no values are independent of each other.

In all cases, the Monte Carlo error for each area was <5% of the standard deviation, indicating sufficient iterations of the model had been run after convergence. The R 'maptools' and 'spdep' libraries were used. Areal boundaries were defined according to postcode districts. Data were then merged to create a data frame with postcodes and the number of cases in each of the 150 areas, substituting NA for 0 where there were no cases. Although each postcode district contains a similar number of people, there is variation in the sizes of postcode districts. The area of each polygon was extracted from the shapefile in R.

To depict the spatial distribution of relative risk across the study area, postcode district maps were obtained from UK Borders and the map for the study area was read in as a spatial polygon (***Fig. S4***). Area-based analyses for mines, landfill sites and quarries were performed after controlling for the area size of each postcode district. Kriging maps are a more sophisticated way of showing distribution of environmental factors in an area because they use much smaller geographical units than the risk maps using postcode districts.

Structural equation modelling

In order to use actual counts of diseases in the SEM, the response variable for modelling used was ‘observed’ minus ‘expected’ cases. The ‘expected’ count was a population-based estimate of the expected number of cases given the population in each area of interest (postcode district). The model was adjusted for population size by assuming that the expected level of disease was dependent on population size as well as any of the covariates.

This methodology was used to develop a hypothetical model, evaluating processes that have direct effects on an outcome (i.e. disease incidence) and indirect effects that can be mediated by other processes (i.e. latent variables that might not be directly measurable). In the SEM, the paths between variables were defined in equation form with the response variable (cases) being related to two or more predictor variables (e.g. coal mines, landfill sites, urbanness) with response variables in one equation forming the predictors in others. SEM tests whether variables were inter-related by analyzing their variances and covariances with goodness-of-fit criteria for each model being used to compare and identify the simplest model and best explanation for the available data.

The conceptual model was built in ‘R’ and challenged using the lavaan package. For the SEM, anything with a Root Mean Square Error of Association (RMSEA) over 0.1 represented a poor fit and models with an RMSEA <0.05 considered to have a good fit.. The comparative fit index (CFI) was used to compare how good the model was at predicting cases compared to what we know about the system with a CFI of 1 indicating a perfect match. Variables with large values were divided to make each of a comparable size for the model i.e. lead/100, traffic/1000, iron/10000, manganese/1000 and arsenic/1000.

Spatial covariates

Data sources were identified for each of the spatial covariates that were available at the level of the geographical unit required for the modelling. The Rural-Urban Classification was based on the 2011 census. Traffic count datasets for 2000-2015 were downloaded from the Department for Transport website (http://www.dft.gov.uk/traffic-counts/) for the North-East and North-West of England. The annual volume of traffic and the annual average daily flow (AADF), presented as vehicles per day, which gives the number of vehicles driving on a stretch of road on an average day of the year. The average number of motor vehicles per postcode area was used as the covariate. The use of a traffic dataset spanning 15 years accounts for variation in traffic over that time period. Data about 1477 landfill sites in the National Grid areas NT, NU, NX, NY and NZ were obtained from the Environment Agency

Geostore and information about 2371 coal-mines, 173 lead mines, 4590 sandstone quarries and 2390 limestone quarries across the NENC region was collated using the British Geological Society (BGS) BRITPITS Database from North-East and North-West England (licence number 2016/076BP ED).The Geochemical Baseline Survey of the Environment (G-BASE) project undertakes annual geochemical sampling within the UK and describes levels of spatial distribution of potential toxins in the landscape. Stream sediment pH data were available for 6,247 sites and data for 5 heavy metals (cadmium, arsenic, lead, manganese and iron) from 11,741 sampling sites within the relevant region. The Townsend Material Deprivation Score is an absolute measure of deprivation derived from 4 Census tables: percentage of households without access to a car or van, percentage of households with overcrowding, percentage of households not owner-occupied and percentage of economically active residents who are unemployed. The Townsend Material Deprivation Score allows areas to be ‘ranked’ as a means of expressing relative deprivation. A greater score implies a greater degree of deprivation. The Index of Multiple Deprivation (IMD) is an overall relative, ranked measure of deprivation constructed by combining 7 domains (***Supplementary table 5***) with rank 1 being the most deprived and rank 32,844 being the least.

Supplementary figures

Fig. S1: Patient distribution across the spatial polygon of study area with 150 postcode districts. A: PBC patients (red dots, n=2150), B: PSC patients (green dots, n=472), C: AIH patients (blue dots, n=963). These data are not corrected for population density.

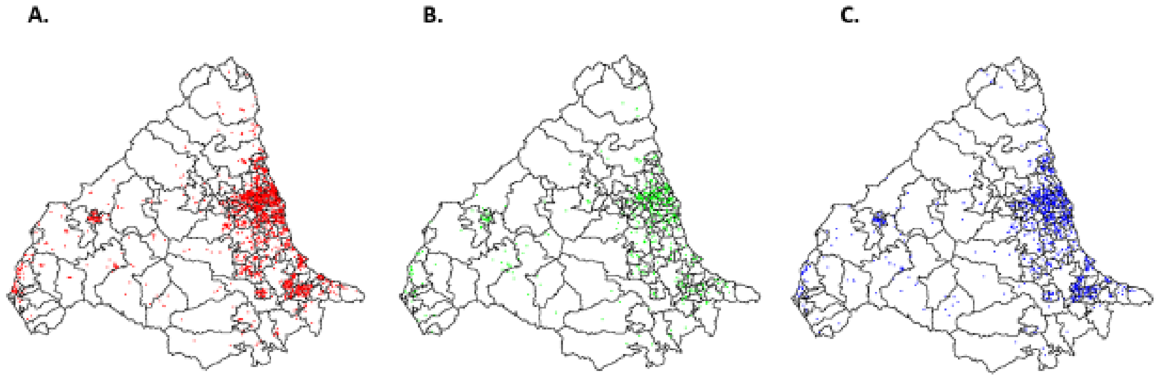


Fig. S2: Monte Carlo test of significance of space-time clustering for cases in the study area where both their postcode and year of diagnosis were known. A: 1780 PBC patients; B: 451 PSC patients; C: 889 AIH patients. A total of 200 simulations undertaken in each case. Filled bar (the data statistic) represents the position of observed data in the set. There was no clustering of cases of PBC, PSC or AIH when a temporal dimension was introduced with the test statistics lying within the body of the histograms.

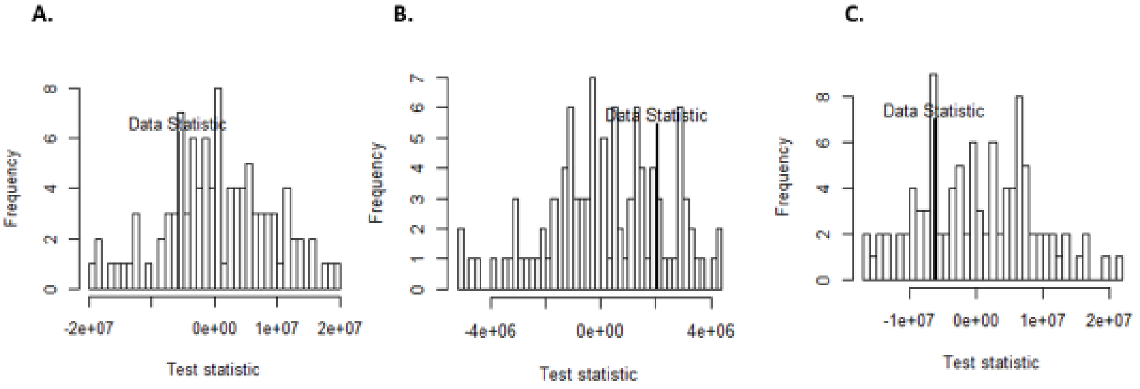


Fig. S3: Location of coal mines in the study area overlaid on kriging maps of A: urbanness; and B: Townsend score.

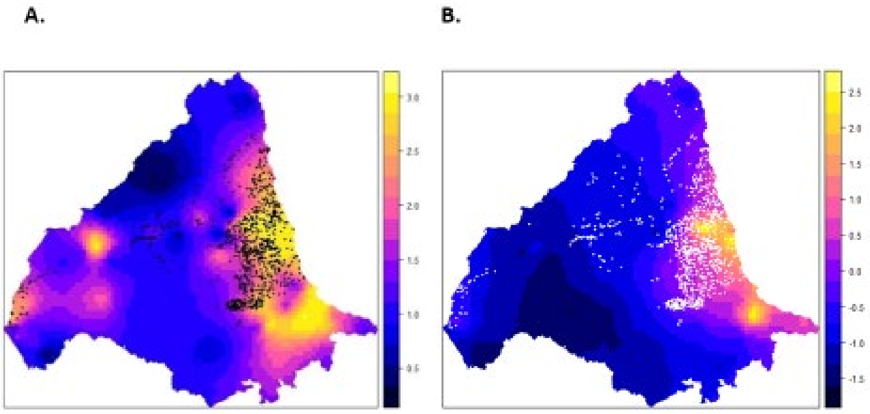
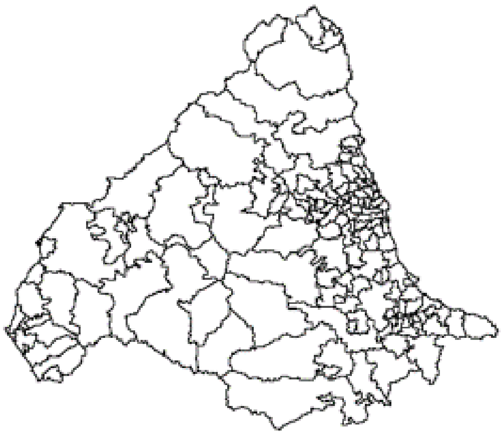


Fig. S4: Spatial polygon of study area with 150 postcode districts across the North-East of England and North Cumbria



References1. Metcalf JV, James OFW. The geoepidemiology of primary biliary cirrhosis. *Seminars in liver disease* 1996; **17**: 13-22.
2. Molodecky NA, Myers RP, Barkema HW, Quan H, Kaplan GG. Validity of administrative data for the diagnosis of primary sclerosing cholangitis: a population-based study. *Liver International* 2011; **31**(5): 712-20.
3. Bailey TC GA. Interactive spatial data analysis: Longman Group Limited; 1995.
4. Diggle PJ, Chetwynd AG, Haggkvist R, Morris SE. Second-order analysis of space-time clustering. *Stat Methods Med Res* 1995; **4**(2): 124-36.