DISEASE CLUSTERING STUDIES IN THE UK

- Richard J.Q. McNally (Newcastle University, UK)
- Richard.McNally@ncl.ac.uk

INTRODUCTION

- Spatial clustering
- Space-time clustering

- Cancer
- Diabetes
- Congenital anomalies
- Primary biliary cirrhosis

SPATIAL CLUSTERING (1)

- Cases allocated to a small area (census wards)
- Person-years by sex & age-group (e.g. <1, 1-4, 5-9, 10-14) estimated using decennial census population data
- Linear interpolation used to estimate between censuses

SPATIAL CLUSTERING (2)

- Expected numbers of cases for each small area computed by applying overall age- & sex-specific rates for relevant time period to the person-years at risk in each small area

SPATIAL CLUSTERING (3)

- Certain diseases show significant heterogeneity by larger area (region) & by area-specific socio-economic status

- To check we fit a regression model containing terms for region & for socio-economic status

SPATIAL CLUSTERING - METHODS

- Alexander & Boyle (1996) compared different methods, using simulated data for census wards

- Optimal method was one due to Potthoff & Whittinghill (1966) – most powerful

POTTHOFF-WHITTINGHILL (PW) METHOD (1)

- If no clustering, observed number of cases in a geographical area should follow a Poisson distribution

- Mean = expected number of cases in that area

PW METHOD (2)

- Then, variance of the observed number of cases = expected number of cases

- PW test used to test whether the ratio of the variance to the expected number is > 1

PW METHOD (3)

- If so, then data would be over-dispersed relative to the Poisson distribution

- Excess numbers of cases would arise in some areas more often than predicted by the Poison distribution

PW METHOD (4)

- Clusters may be seen
- May represent a general feature of the disease distribution

PW METHOD (5)

- Magnitude of any over-dispersion quantified as a factor by which the variance of the observations is increased
- $1+b$
- Define b as the extra-Poisson variation (EPV)

PW METHOD (6)

- EPV = b = 0.1
- If variance of the theoretical distribution of the number of cases in each ward 10% larger than predicted by the Poisson distribution
- See Muirhead and Butland (1996)

PW METHOD (7)

- Additional analyses can split EPV between & within subgroups
- If variance of the theoretical distribution of the number of cases in each ward 10% larger than predicted by the Poisson distribution
- See Muirhead and Butland (1996)

SPATIAL CLUSTERING - RESULTS

- Cancer – evidence of clustering based on place of diagnosis
- PBC – evidence of clustering based on place of diagnosis

SPACE-TIME CLUSTERING (1)

- Analyses focus on individual locations
- Centroids of postcodes obtained to an accuracy of 100m (Eastings and Northings)

SPACE-TIME CLUSTERING (2)

- Analyses based on Knox test and *K*-functions
- *K*-function analysis is a generalisation of the Knox test

KNOX TEST (1)

- Regards a pair of cases as being in “close proximity” IF

- they are diagnosed both at addresses that are close in space (< *s* km apart, e.g. *s* = 5)

KNOX TEST (2)

- & at times that are close (< *t* months apart, e.g. *t* = 12)

- *Critical values* chosen arbitrarily

KNOX TEST (3)

- Number of pairs of cases in *close proximity* calculated (*O*)
- Number of pairs of cases that are *close in space* calculated (*D*)
- Number of pairs of cases that are *close in time* calculated (*E*)

KNOX TEST (4)

- Assume that spatial & temporal proximity are independent
- Then, the expectation of *O*, is *D* x *T* / *N*
- *N* is total number of case pairs

KNOX TEST (5)

- If $O >> E$ there is evidence of space-time clustering
- Statistical tests used to determine whether excess is statistically significant

STRENGTH OF CLUSTERING (1)

- Magnitude of the excess (or deficit) estimated by $S = [(O - E) / E] \times 100$

- Variability of *S* depends on *E*

STRENGTH OF CLUSTERING (2)

- Related quantity whose variability is approximately independent of *E* is:

- $R = (O - E) / \sqrt{E}$

PROBLEMS WITH KNOX TEST

- Boundary problems may be important
- It can be difficult or impossible for a case near a geographical boundary or at the end of a time intervals to be close to many cases
- Arbitrariness of selected values (*s* and *t*)

K-FUNCTION ANALYSIS (1)

- Simplification of procedure due to Diggle *et al*
- Perform set of Knox-type calculations to obtain set of values of $R = (O - E) / \sqrt{E}$
- Approximate value of integral $\int R \, dsdt$

K-FUNCTION ANALYSIS (2)

- E.g. Obtain 225 values of R by varying critical values over a pre-specified set

- Vary critical values over a pre-specified set, e.g. for close times, $t = 0.1, 0.2, \ldots, 1.5$ years & for close points in space, $s = 0.5, 1, 1.5, \ldots, 7.5$ km

K-FUNCTION ANALYSIS (3)

- Distribution of *K*-function unknown - must be estimated by simulation
- For each simulation, randomly re-allocate dates of event (e.g. diagnosis, birth) to each of the cases in the analysis
- Calculate value of the *K*-function, *K(S)* from the simulated data

K-FUNCTION ANALYSIS (4)

- Repeat for a total of *n(SIM)* simulations
- Compare observed value of the *K*-function, *K(O)*, with the simulated values, *K(S)*
- *s* = 1,…, *n(SIM)*

K-FUNCTION ANALYSIS (5)

- *P*-values estimated by calculating proportion of the *n(SIM)* simulations for which *K(S)* > *K(O)*

- Hence assess statistical significance

K-FUNCTION ANALYSIS (6)

- LIMITATION
- *K*-function analysis yields no measure of the magnitude of the clustering effect
- Use $S = (O - E) / E$, derived from Knox test for specified critical values

NEAREST NEIGHBOUR (NN) APPROACH (1)

- 2 types of prior hypothesis:
- First relates to spatially fixed factors (although exposures may be temporally heterogeneous)
- Second relates to risk factors spread by person to person transmission

NN APPROACH (2)

- Fixed geographical distances can be used for testing:
- The first type of hypothesis
- The second type of hypothesis, but only when the population is relatively homogeneous

NN APPROACH (3)

- In practice this is often not true – especially when both urban & rural areas are included

- Any specified distance between 2 cases may have different meanings in urban & rural areas

NN APPROACH (4)

- For example, the sizes of school catchment areas will differ

- Use of fixed geographical distance thresholds may lead to underestimation of expected numbers in more densely populated areas

NN APPROACH (5)

- May inflate apparent clustering effects in more densely populated areas
- Leading to inflation of overall clustering effect

NN APPROACH (6)

- Use of fixed geographical distance thresholds may lead to underestimation of expected numbers in more densely populated areas

- NN threshold method takes localised variations in population density into account

NN APPROACH (7)

- For a specific ‘index’ case, the other cases that are in closest proximity are termed ‘nearest neighbours (NNs)’
- Rank the NNs according to distance from the index case: 1, 2, 3,…..
- Do for every case in dataset (by treating as an index case)

NN APPROACH (8)

- Make comparison with fixed analyses
- E.g. mean distance to 26th NN is approx 5 km
- Distances vary from 0.7 km – 245.3 km
- Choose 19th …..33rd NNs instead of fixed distances 0.5 – 7.5 km

SPACE-TIME CLUSTERING - RESULTS

- Cancer – space-time clustering (STC) based on birth & diagnosis (addresses & times)
- Diabetes – STC based on diagnosis
- Trisomy 21- STC based on delivery
- PBC – STC based on diagnosis

EXTENSIONS (1)

- Gender analyses
- Can analyse clustering pairs of ‘male: male’ or ‘female: female’ case pairs
- More meaningful to analyse ‘male: any’ or ‘female: any’ case pairs

EXTENSIONS (2)

- Urban / rural analyses
- Divide cases into those from ‘more densely populated areas’ & those from ‘less densely populated areas’
- Analyse ‘more densely populated: any’ & ‘less densely populated: any’ clustering pairs

EXTENSIONS (3)

- Cross-clustering between diseases (or specific sub-groups)
- Analyse clustering pairs ‘a: a’ & ‘b: b’
- Analyse cross-clustering pairs ‘a: b’

RESIDENTIAL HISTORY DATA

- Cancer – birth & diagnosis
- PBC – history of residential addresses
- Congenital anomalies – possibly (?)

THANK YOU FOR YOUR ATTENTION!