

# Geographical Variation of Non-Communicable Diseases and Environmental Risk Factors: Application of Bayesian Modeling and GIS

*Elena Moltchanova<sup>1</sup>, Mika Rytönen<sup>1</sup>, Anne Kousa<sup>2</sup>, Olli Taskinen<sup>1</sup>,  
Jaakko Tuomilehto<sup>1</sup> & Marjatta Karvonen<sup>1</sup> for the SPAT Study Group  
and the Finnish Childhood Diabetes Registry Group*

<sup>1</sup>National Public Health Institute  
Department of Epidemiology and Health Promotion  
Diabetes and Genetic Epidemiology Unit  
Helsinki, Finland  
Phone: +358 9 4744-8608 Fax: +358 9 4744-8338  
Email: elena.moltchanova@ktl.fi

<sup>2</sup>Geological Survey of Finland  
Kuopio, Finland  
Phone: +358 205 50 3644 Fax: +358 205 50 13  
Email: anne.kousa@gsf.fi

**Presented at GeoHealth 2002  
Victoria University of Wellington  
December 3-5<sup>th</sup> 2002**

## ABSTRACT

The incidence of Coronary Heart Disease (CHD) and childhood type 1 diabetes (DM1) is among the highest in the world in Finland. Also the occurrence of DM1 and CHD are known to have a large geographical variation in Finland. The recent advances in GIS, availability of geo-referenced data and developments in spatial methodology are valuable tools for exploratory analysis, testing and generating hypothesis and searching for unusual clustering. The aim is to present recent developments in methodology of geographical epidemiology research and to demonstrate its use in small-area study on DM1 and AMI in Finland.

### *Data*

**CHD:** Data on non-fatal definite acute myocardial infarctions (AMI) and mortality data (CHD mortality) included 19,000 cases from the cross-section years 1983, 1988 and 1993. The mid-year number of men aged 35-74 years in each cross-section year was used as a denominator.

**DM1:** Data on the incidence of childhood Type 1 diabetes for the period 1987-1996 included 3649 cases. The mid-year number of children aged 0-14 years in each cross-section years 1987, 1989, 1991, 1993 and 1995 was used as a denominator.

**Water macro- and trace elements:** The geochemical data on nitrate (NO<sub>3</sub>), zinc (Zn), water hardness (dH) and some other elements of the ground water were obtained from the hydrogeochemical database of the Geological Survey of Finland and interpolated onto a regular grid.

**Urban rural-rural status:** Also, the area of Finland was classified into rural or urban based on the values of over 30 indicators such as e.g. population density, economic activity etc.

### *Methods*

A full Bayesian approach was used to explore the relationships between the distribution (Poisson intensities) of disease occurrence and environmental risk factors. The Markov chain Monte Carlo (MCMC) computation techniques and Bayesian analysis allowed us to use more elaborate models and to take into account plausible prior knowledge of the phenomenon.

### *Results*

There was a clear spatial pattern in the incidence of DM1 with a constant high-risk area in the central Finland and also stable low risk areas in the southernmost Finland. Children living in rural heartland areas had a higher risk for type 1 diabetes than children living urban-adjacent rural areas. There was also a tendency to a higher risk of diabetes when compared to risk in rural heartland area to urban and remote rural areas.

The spatial pattern of AMI incidence showed that a high-risk area is still in eastern Finland, although it withdrew to the northeast while the overall incidence decreased. For the pooled data the only statistically significant influence was found in the water hardness, 1 additional unit of which decreases the average incidence by 1%.

### *Conclusions*

Environmental risk factors may have a meaningful role in the aetiology of chronic diseases such as CHD and DM1. The Bayesian hierarchical model offers a great opportunity to study environmental factors using them as covariate variables in the model. Although such a study cannot disclose the mechanism of disease occurrence they indicate at a population level associations between disease and potential environmental exposures.

**Keywords and phrases:** Bayesian spatial modeling, environmental covariates, DM1, AMI

## **1.0 INTRODUCTION**

The incidence of Coronary Heart Disease (CHD) and childhood type 1 diabetes (DM1) is among the highest in the world in Finland and is known to have large spatial variation. The recent advances in GIS, availability of geo-referenced data and developments in spatial methodology are valuable tools for exploratory analysis, testing and generating hypothesis and searching for unusual clustering. The purpose of this paper is to present recent developments in methodology of geographical epidemiology research and to demonstrate its use in small-area study on childhood type 1 diabetes and acute myocardial infarction in Finland. The geographical distribution of the incidence of DM1 and AMI were using a full hierarchical Bayesian approach. In addition, the Bayesian model was extended to include information on potential environmental and socio-economic risk factors.

## **2.0 DATA**

### **2.1 DIABETES DATA**

Data on the incidence of childhood type 1 diabetes nation-wide were obtained from the Prospective Childhood Diabetes Registry for the period 1987-1996 (3649 cases). Since 1987 all the newly diagnosed cases have been registered in the Prospective Childhood Diabetes Registry with high level of case ascertainment.

## **2.2 AMI DATA**

The data on men aged 35-74 year with first attack of AMI (18946 cases of both fatal and non-fatal events) were obtained from the nation-wide Death Register and the Hospital Discharge Register for the years 1983, 1988 and 1993. The personal identification number was used to perform a computerized records linkage of the National Death Register data and the Hospital Discharge Register data for deaths due to coronary heart disease (CHD) (ICD 9 codes 410-414) and hospitalization due to AMI (ICD codes 410-411). The record was linked to trace back the possible earlier events of AMI in each cases obtained from the National Death Register. For the environmental factor analysis the data from the three cross-sectional years were pooled together.

## **2.3 POPULATION-AT-RISK**

AMI: The mid-year number of men aged 35-74 years in each cross-section year was used as a denominator. The data for the years 1983 and 1993 were obtained from the Population Registry whereas the data for the year 1988 was estimated from them under an assumption of negligible migration.

DM1: In a mapping study the mid-year number of children aged 0-14 years in each cross-section years 1987, 1989, 1991, 1993 and 1995 was used as a denominator. In a study of the incidence variations by area rurality the annual population data 1987-1996 were used as a denominator. Both data sets were obtained from the National Population Registry, which is continuously updated by Statistics Finland.

## **2.4 ENVIRONMENTAL DATA**

The data on total water hardness (°dH), Ca, Mg, Fe, F, NO<sub>3</sub> (mg/l) and Cu, Zn and Al (µg/l) were available from the hydrogeochemical database of GTK (Geological Survey of Finland). The original data contained from 3621 up to 12401 samples taken mainly from dug wells and wells drilled in the bedrock.

## **2.5 URBAN/RURAL DIVISION DATA**

The 445 municipalities in the study area was classified into four categories: 1) urban areas, 2) urban-adjacent rural areas, 3) rural heartland areas and 4) remote/isolated areas. A total of 30 variables were used in the classification process and the following principal factors were considered to be important: natural conditions, internal differentiation, town-country interaction, rural development and local factors.

## **2.6 GEO-REFERENCING THE DATA**

Every resident in Finland has a unique personal identity number. Using this number 95% of all children with diabetes and men with AMI were located at the time of diagnosis according to the map coordinates.

The geo-referenced data sets (point data structure), both population at risk and case data, were converted into grid structure and aggregated into 10 x 10 km grid cells to ensure the protection of privacy of individuals.

The geographical scale of the study of incidence variation over the urban/rural continuum was fixed at municipality level because information of the rurality status of the area was available only on that level. Using unique municipality/county code, the urban/rural classes of the municipalities were linked to the incidence data.

The geochemical point data were interpolated into a regular grid by using ALKEMIA software Smooth interpolation method option.

### 3.0 BAYESIAN SPATIAL MODEL

Bayesian spatial conditional autoregressive model (CAR) with covariates, which is currently in wide use in the field of the disease mapping, was applied here. Here we propose one modification, which is pertinent to the sparsely populated areas.

Since Finland is a sparsely and unevenly inhabited country some of the cells based on fine grids are empty and have to be omitted from the analysis. For example, in the case of 10 km x 10 km grid over Finland (excluding Lapland) such cells comprise 5% of all the grid. However, once we take environmental factors into account, thus assuming that the disease risk is influenced by both demographic factors (i.e. people who actually live within the grid cell) and the environmental factors, which exist in each cell whether or not it is inhabited, the omission of unpopulated cell results in the loss of information. The following modification is thus proposed:

Let  $Y_{ik}$  denote the number of cases in the cell  $i$  and age group  $k$ . Further let  $N_{ik}$  denote the respective population at risk. The proposed probability distribution is then following:

$$P(Y_{ik} = y | N_{ik}, \mu_{ik}) = \begin{cases} 1 & \text{if } y = N_{ik} = 0 \\ \frac{e^{-\mu_{ik}} \mu_{ik}^y}{y!} & \text{if } 0 \leq y \leq N_{ik} \text{ and } N_{ik} > 0 \\ 0 & \text{elsewhere} \end{cases}$$

i.e. the Poisson distribution is assigned to the inhabited cells and the uninhabited cells naturally have no cases of the disease with the unit probability. Also we assign common regression structure to the  $\mu_{ik}$ :

$$\begin{aligned} \text{log}(\mu_{ik}^{(k)}) &= \alpha + \gamma^k + \beta^k + \xi^k + \lambda_i^k & \text{if } \lambda_i^k &= 0 \\ \text{log}(\mu_{ik}^{(k)}) &= \alpha + \gamma^k + \beta^k + \xi^k + \lambda_i^k + \text{log}(\lambda_i^k) & \text{if } \lambda_i^k &> 0 \end{aligned}$$

where

$\alpha$	is the background risk level
$\lambda_i$	is the local unexplained spatial variation of the risk level
$\beta$	is the effect of age on the risk level
$k$	is the age group, $k=0, \dots, K$
$\xi$	is a vector of environmental covariate effects
$Z_i$	is a vector of environmental covariates
$\gamma$	is a vector of demographic covariate effects
$P_i$	is a vector of demographic covariates

In the analyses of DM1 the age axis was divided into three age groups 0-4, 5-9, and 10-14 years. The cases were not differentiated by gender since there was no evidence of a potential incidence difference between genders in any areas. In the analyses of AMI the age axis was divided into eight age groups: 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, and 70-74 years.

The regression coefficients  $\beta, \gamma$  and  $\xi$  were given non-informative Normal priors  $N(0, 0.00001)$ , the background level  $\alpha$  was given an improper flat prior

$$p(\alpha) \propto 1$$

and the  $\lambda$  were given a CAR structure:

$$\lambda_i \sim N(\bar{\lambda}_{-i}, \sigma_i)$$

where

$\lambda_{.i}$	are spatial variation parameters in the neighborhood of $i$
$m_i$	is the number of neighbors for cell $i$
$\tau$	is the overall level of spatial precision (inverse spatial variation)

The neighbors were defined to be all those cells adjacent to the cell  $i$  through side or corner. Thus each cell could have at most eight (8) neighbors.

The parameter  $\tau$  was given vague Gamma prior  $\Gamma(0.001,0.001)$ . The model for diabetes was fitted using WinBUGS, the model for AMI -- a SAS program. Diagnostic tests have indicated that convergence was reached.

## 4.0 RESULTS

### 4.1 GEOGRAPHICAL OCCURRENCE OF DM1 AND AMI

The maps of spatial pattern of mean observed and estimated intensity of DM1 are presented in Figures 1-2. Most of the estimated high-risk areas of DM1 were in the wide belt crossing the central part of Finland. Although there was similarities in the spatial risk pattern of type I diabetes during the two 5-year periods, the pattern did not remain stable. However, the changes of the geographical pattern of risk in Northern Finland may be more due to the inability of the model to smooth the variation of the estimates since the population at risk was small in the northern part of the country.

The maps of spatial pattern of mean estimated intensity of AMI cases showed a temporal decline of the occurrence of AMI (Figures 3-5b). The spatial pattern of AMI incidence showed an increasing risk from the south-west to the north-east. From cross-section to cross-section the high-risk area withdrew deeper to the north-east along with the decreasing overall incidence.

The posterior estimated intensity of a disease in a grid cell is only a point estimate. To confirm the credibility of the displayed geographical pattern of the occurrence of both diseases the maps of Bayesian posterior probabilities of the estimated intensities being higher than the estimated country average were produced.

### 4.2 BAYESIAN ECOLOGICAL MODELING

#### Incidence of type 1 diabetes according to urban-rural status of the areas

Bayesian model-based approach was used to model variation in the incidence of DM1 among children between urban/rural municipalities. The average increase in the incidence was 2,4% per year during the period 1987-1996. Children living in rural heartland areas had a higher risk for DM1 than children living urban-adjacent rural areas. The risk of type 1 diabetes was 15 % higher in rural heartland than urban-adjacent rural areas. There was also a tendency to a higher risk of diabetes when compared to risk in rural heartland area to urban and remote rural areas. For example, the mean incidence was 8% ( $\theta_{24}=0.0731$ ) higher in rural heartland areas than urban areas in Finland during 1987-1996.

Parameter	mean	sd	2.50 %	median	97.50 %
$\theta_{12}$	0.0858	0.0725	-0.0572	0.0864	0.2262
$\theta_{13}$	-0.0637	0.0725	-0.2049	-0.0641	0.0800
$\theta_{14}$	0.0127	0.0753	-0.1313	0.0118	0.1614
$\theta_{23}^*$	0.1495	0.0513	0.0472	0.1501	0.2484
$\theta_{24}$	0.0731	0.0573	-0.0422	0.0740	0.1819
$\theta_{34}$	-0.0764	0.0531	-0.1821	-0.0757	0.0262

**Table 2.** Estimated effects of area rurality on the incidence of DM1 among 0-14 year olds in Finland.  $\theta_{ij}$  is the difference between the area types  $i$  and  $j$ , where 1= remote area , 2 = rural heartland, 3 = urban-adjacent rural area and 4 = urban area.

### Bayesian modeling of AMI incidence and geochemical covariates

The geochemical elements such as total water hardness, Ca, Mg, Fe, F, Cu, Al, Zn and  $\text{NO}_3$  concentration in the ground water were taken into analysis as the covariates. The estimates of the effect of geochemical covariates in the ground water on the variation of the incidence of AMI adjusted for age-group are shown in table 3. One additional unit ( $^{\circ}\text{dH}$ ) of water hardness on average decreases the risk of AMI by 1 %. The levels of the other elements in the ground water included in this study did not have any additional effect on the incidence and spatial variation of AMI.

Element	Posterior mean	95 % CI
Tot. water hardness ( $^{\circ}\text{dH}$ )	-0.0097160	(-0.0213600, -0.0003195)
Zn ( $\mu\text{g/l}$ )	-0.0006656	(-0.0061290, 0.0048140)
Al ( $\mu\text{g/l}$ )	-0.0002723	(-0.0007370, 0.0001862)
Cu ( $\mu\text{g/l}$ )	0.0400800	(-0.0652800, 0.1477000)
F (mg/l)	-0.0317200	(-0.1453000, 0.0898500)
Fe (mg/l)	0.1015000	(-0.1298000, 0.3176000)
$\text{NO}_3$ (mg/l)	0.0006068	(-0.0003548, 0.0015870)

**Table 3.** Estimated effect of geochemical covariates on the incidence of first AMI among 35-74 year old men in Finland.

## 5.0 CONCLUSIONS

Environmental risk factors may have a meaningful role in the aetiology of chronic diseases such as CHD and DM1. The Bayesian hierarchical model offers a great opportunity to study environmental factors using them as covariate variables in the model. Although such a study cannot disclose the mechanism of disease occurrence they indicate at a population level associations between disease and potential environmental exposures. It is also known that genetic factors play a large role in the risk of both diseases considered above. Their inclusion in the analyses may provide further useful insights into aetiology of DM1 and AMI.

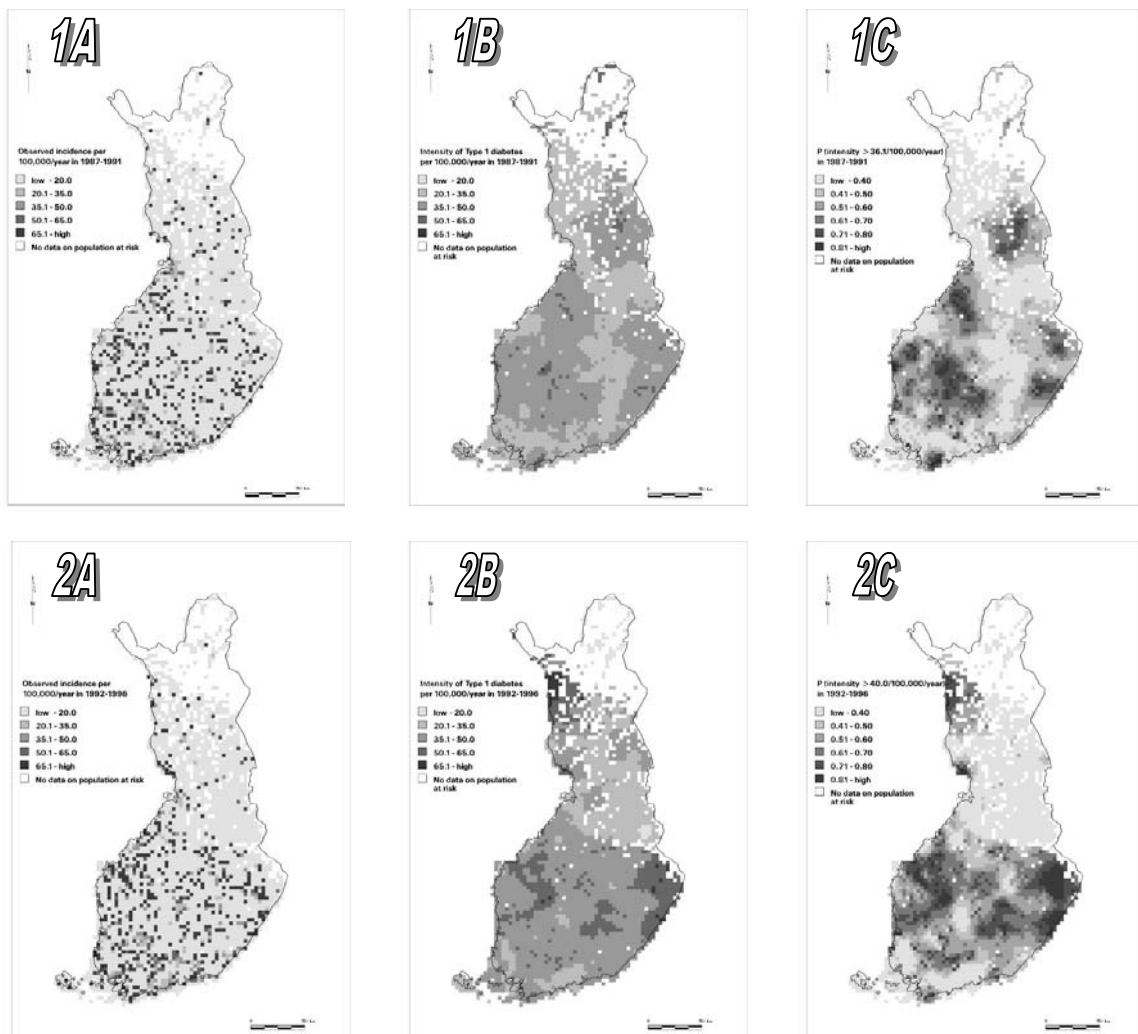
A major methodological limitation of the current study is the use of aggregated data and therefore inferences based on the analysis cannot be directly transferred to the individual level. One of the problems in the geographical analysis of chronic diseases is the lag-time between a potential exposure

and the occurrence of the disease symptoms. People may have been exposed much earlier and might have lived in a different location than where the first signs of the disease occur; thus an exposure experienced earlier in the life may become associated with an inaccurate geographic location. Ecological studies are most useful for generating and testing hypothesis (i.e. qualitative identification of an association) rather than quantitative estimation of the strength of an exposure-response relationship.

**Figures 1-2.** Risk of DM1 among 0-14 year old children in Finland during 1987-1991 (fig1) and 1992-1996 (fig2):

- (A) Observed age-standardized incidence,
- (B) Posterior mean age-standardized incidence,
- (C) Posterior probability of the risk exceeding the overall country risk

The same scale was used for figures 1A, 2A; for figures 1B, 2B; for figures 1C, 2C

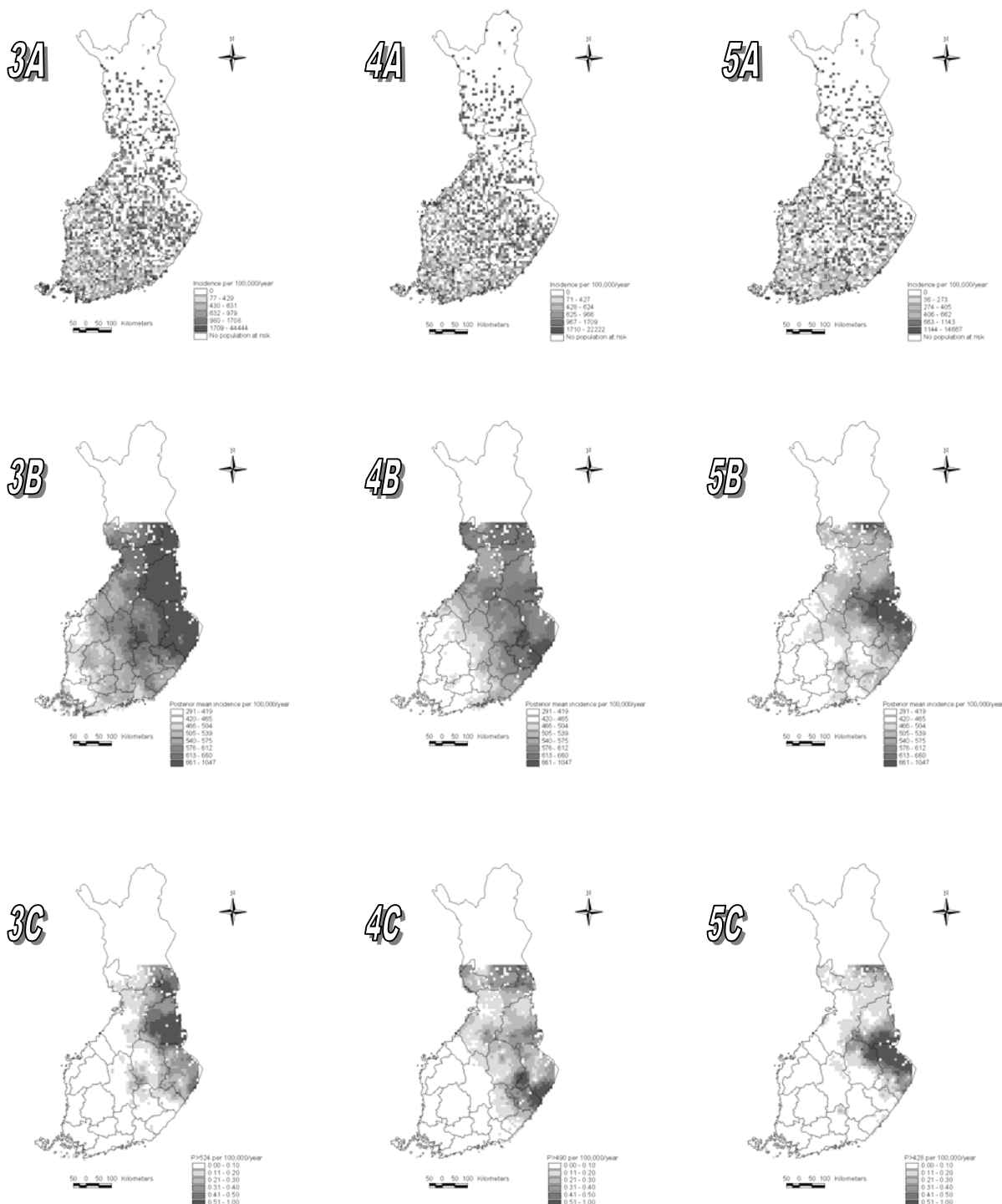


**Figures 3-5.** Risk of AMI among 35-74 year old men in Finland in 1983 (fig1), 1988 (fig2) and 1993 (fig3): (A) Observed age-standardized incidence,

(B) Posterior mean age-standardized incidence,

(C) Posterior probability of the risk exceeding the overall country risk

The same scale was used for figures 3A, 4A, 5A; for figures 3B, 4B, 5B; for figures 3C, 4C, 5C



## ACKNOWLEDGEMENTS

This work was funded by the Academy of Finland; project No. 41266, the NIH-grant DK-37957, The Alma and K.A. Snellman Foundation, Oulu, Finland and Wihuri Foundation, Finland. The data were provided by the Diabetes and Genetic Epidemiology Unit at the National Public Health Institute, Finland, and computing facilities were provided by the Rolf Nevanlinna Institute, University of Helsinki. GIS-facilities were partly offered by Department of Geography, University of Oulu, Finland. Elena Moltchanova was supported by the Doctoral Program in the COMAS Postgraduate School of the University of Jyväskylä.

### The SPAT-Project Research Group:

*Principal investigator:* Marjatta Karvonen

*Investigators:* Anne Kousa, Elena Moltchanova, Jukka Ranta, Mika Rytönen, Maarit Viik-Kajander

*Steering group:* Elja Arjas, Unto Häkkinen, Mikko Nenonen, Olli Nylander, Juha Pekkanen, Antti Penttinen, Jarmo Rusanen, Reijo Salminen, Timo Tarvainen, Jaakko Tuomilehto

### The Finnish Childhood IDDM Registry group:

Fagerlund A, Flittner M, Floman P, Gustafsson B, Haavisto H, Huupponen T, Hiltunen P, Hyttinen M, Jokisalo R, Kaski U, Komulainen J, Korpela P, Käär M-L, Lappalainen J, Lautala P, Lehtinen P, Lehtokoski-Lehtiniemi E, Lipsanen-Nyman M, Niemi K, Nuuja A, Ojajarvi P, Ollikainen J, Putto-Laurila A, Pöntynen S, Sankala J, Sillanpää T, Sipilä I, Tapanainen P, Tomminen A, Uotila T, Varimo P, Vuolukka P, Väre M.

## REFERENCES

Besag J, York J, Mollié A (1991) Bayesian Image Restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, pp. 1-59.

Gustavsson N, Lampio E, Tarvainen T (1997). Visualization of geochemical data on maps at the Geological Survey of Finland. *J Geochem Explor*, 59, pp. 197-207

Karvonen M, Moltchanova E, Viik-Kajander M, Moltchanov V, Rytönen M, Kousa A, Tuomilehto J for the SPAT study group (2002) Regional inequality in the risk of acute myocardial infarction (AMI) in Finland *HeartDrug*, 2, pp. 51-60.

Kousa A, MSc, Moltchanova E, Viik-Kajander M, Rytönen M, Karvonen M, for the Spat study group Geochemistry of the Ground Water and the Incidence of Acute Myocardial Infarction in Finland (submitted to *European Heart J*).

Lahermo P, Ilmasti M, Juntunen R, Taka M. The Geochemical Atlas of Finland, Part 1 (1990) The hydrochemical mapping of Finnish groundwater. Geological Survey of Finland, Espoo, 66.

Moltchanova E, Rytönen M, Kousa A, Taskinen O, Tuomilehto J, Karvonen M for the SPAT Study Group and the Finnish Childhood Diabetes Registry Group. Zinc and Nitrate in the Ground Water and the Incidence of Type 1 Diabetes in Finland (submitted to *Epidemiology*).

Ranta J, Penttinen A (2000) Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. *Statistics in Medicine*, 19, pp. 2345-2359.

Rytönen, M., Moltchanova, E., Ranta, J., Taskinen, O., J. Tuomilehto and M. Karvonen for the Spat study group and the Finnish Childhood Diabetes Registry Group (2002) The Incidence of Type 1 Diabetes Among Children in Finland: Rural-Urban Difference.(submitted to *Health and Place*)

Rytönen, M., Ranta, J., J. Tuomilehto and M. Karvonen for the Spat Group and the Finnish Childhood Diabetes Registry Group (2001) Bayesian analysis of geographical variation in the incidence of Type 1 diabetes in Finland. *Diabetologia* 44 (Suppl 3), pp. B37-B44.

Viik-Kajander M, Moltchanova E, Tuomilehto J, Salomaa V, Ketonen M, Palomäki P, Miettinen H, Torppa J, Karvonen M for the FINMONICA AMI register group "Geographical variation in the incidence of acute myocardial infarction (AMI) in Eastern Finland – a Bayesian prospective" (*Annals of Medicine*: in press).