PERGAMON

European Journal of Cancer 39 (2003) 1973–1975

European Journal of Cancer

www.ejconline.com

Editorial Comment

Editorial comment on Geographical differences in cancer incidence in the Belgian Province of Limburg by Bruntinx and colleagues

L. Jarup*, N. Best

The Small Area Health Statistics Unit (SAHSU), Department of Epidemiology and Public Health, Imperial College London, St Mary's Campus, 16 South Wharf Road, London W2 1PF, UK

Received 25 February 2003; accepted 28 March 2003

The relationships between risk factors and cancer occurrence at a geographical level are not easily understood, interpretation being greatly dependent on scale. Maps of cancer incidence on a global scale have been produced by the International Agency for Research on Cancer (IARC), showing clear spatial patterns between different parts of the world [1]. The overall pattern shows that total cancer incidence is higher in affluent parts of the world, most markedly for cancer types closely related to socio-economic status. For example, it is evident that lung cancer is a disease mainly affecting the developed world (so far), whereas liver cancer incidence is high in countries where hepatitis B and C infections are endemic.

At a continental area level, the United States (US) National Cancer Institute (NCI) produces interactive maps of cancer mortality at US county or State Economic Area levels (one or more socio-economically similar counties within a state) [2]. The maps are available on the NCI website (www.nci.nih.gov) and show (for example) lung cancer patterns, which are seemingly related to socio-economic status (a good proxy for smoking), whereas patterns are less clear for other cancers. NCI notes that "the study of geographic patterns of cancer may provide important clues to the causes of cancer and improvements in cancer control", but also that it "does not provide information about *why* death rates may be higher in certain localities than in others", although "it can generate leads for in-depth epidemiologic studies that may shed light on factors contributing to cancer risks." Risk factors, which may be heterogeneously distributed, include occupational and environmental exposure, smoking habits and socio-economic status (SES). These data would be difficult and expensive to collect, but some (e.g. SES) may be available through the census.

It should also be noted that geographical variations in disease rates might reflect differences in access to medical care, such as screening and diagnostic procedures. Thus, routinely collected cancer data may reflect differences between cancer registries, if several different registries contribute to the data used for analysis in a geographical area.

At a small area level (such as a municipality), spatial patterns of disease are often far more difficult to interpret. Using modern geographical information systems (GIS), data are usually aggregated up to a practicable level, depending on which geographical units have relevant data available. The level of data aggregation will impact the analysis and different results may be obtained depending on the definition of the areas. A well-known example of this problem relates to the cholera epidemic in London in the 19th century. John Snow, often referred to as the father of epidemiology as well as GIS, mapped individual cholera cases in Soho, and was able to relate the occurrence of cases to a certain water pump. It has later been demonstrated that if John Snow had used modern GIS tools, aggregating his point data to arbitrary geographical units, he would have found differing spatial patterns depending on the choice of boundaries [3]. Thus, the impact of spatial units used for any analysis of spatial disease patterns should be recognised. Administrative units (such as postcodes, municipalities, enumeration districts or wards) may not be optimal, but are often readily available; grid squares (or similar) are often better, but it may prove difficult (and expensive!) to get the necessary data aggregated into such units.

It is important to note that cancer rates may be elevated in some geographical areas purely due to chance. Furthermore, cancer maps are related to the place of residence at the time of diagnosis (incidence) or death (mortality). For cancers with a long latency time (most

* Corresponding author. Tel.: +44-20-7594-3337; fax: +44-20-7402-8837.

E-mail address: l.jarup@imperial.ac.uk (L. Jarup).

0959-8049/03/$ - see front matter
doi:10.1016/S0959-8049(03)00317-4

solid tumours), high migration rates will make any spatial patterns associated with potential exposures less obvious. For rare cancers, the number of cases in many of the geographical units will be small or even nil, leading to randomly fluctuating rates.

For these reasons, the choice of statistical methods used to produce small area scale disease maps is extremely important. The Standardised Incidence or Mortality Ratio (SIR, SMR) is one of the most widely used estimates of disease risk in epidemiological studies. The standard error of this estimate is inversely proportional to the square root of the expected number of cases in the study area. Since the latter are typically very low for rare diseases in small areas, but can vary by an order of magnitude or more over the study region due to differences in population density and age structure, this leads to risk estimates in different areas that have very different precisions. Maps displaying such point estimates are thus very hard to interpret. Statistical smoothing of risk estimates based on Bayesian inference methods has been advocated in the statistical literature for some time [4–6]; it is encouraging to see that in this issue Buntinx and his colleagues are endorsing their application in the medical/epidemiological literature [7]. Bayesian smoothing works by taking a form of weighted average of the SIR (SMR) in a particular small area and the overall risk for either the whole study region (leading to global smoothing) or for a local region surrounding the small area (local smoothing). The resulting smoothed risk estimates for each area are more precise and more robust against false-positive estimates than are SIRs (SMRs), and hence lead to more meaningful and interpretable disease maps, as can be seen from the comparison of SIR and smoothed cancer maps presented by Buntinx and colleagues [7]. None the less, some concerns have been voiced that Bayesian risk estimates may tend to over-smooth variations in disease risk, particularly if the data are very sparse [8,9], and it is important to be able to reassure the public that true clusters have not been smoothed away by using these methods. Further empirical work is needed to address this question, but it is clear that care must be taken not to over-interpret a *lack* of spatial variation in a disease map, in much the same way as epidemiologists and statisticians already caution against over-interpretation of apparent clusters.

A criticism that is sometimes levelled against mapping SIRs (SMRs) or smoothed versions of these, is that any two SIRs (SMRs) are not directly comparable since they are not based on the same standard population [10]. While this is theoretically correct, in practice, using (smoothed or unsmoothed) SIRs (SMRs) to compare the level of risk in two or more areas will only be misleading if the age and gender structure of the area populations are extremely disparate [11]—a situation that very rarely occurs in practice. The imprecision of alternative statistical estimates such as directly standardised rates when calculated on small area scale is a far more serious problem.

Papers describing relative risks (e.g. smoothed or unsmoothed SIRs) almost exclusively show point estimate maps, not taking into account the uncertainty in the risk estimates. The paper by Buntinx and colleagues is no exception [7], although 95% Confidence Interval (CI)s for the relative risk estimates for each municipality are given in a separate table. Nevertheless, maps are a far more powerful tool than tables for conveying information about geographical variations in risk and so techniques need to be developed to also map information about uncertainty in these risk estimates in a way that is easy for readers to understand, without the need for an in-depth statistical knowledge. A possible solution when using Bayesian smoothing methods is to map the posterior probability of any area's relative risk exceeding a prespecified threshold. Taking this threshold to be unity, these posterior probabilities may be interpreted as the strength of (statistical) evidence of excess risk in each area. High probabilities can be interpreted as providing clear evidence of an excess risk, while low probabilities have the reverse interpretation of providing clear evidence of a reduced risk [9].

As pointed out by Buntinx and colleagues [7], an important advantage of Bayesian smoothed risk estimates (or more specifically, of associated measures of uncertainty such as 95% CIs and threshold exceedance probabilities) is that they are automatically adjusted for the problem of multiple significance testing since they are based on the joint distribution of the smoothed risk estimates for all the areas in the map. This is in contrast with the usual P values associated with significance tests of the SIR (SMR), which are calculated independently for each area. The multiple testing problem still remains, however, if many different diseases (say, different cancer sites) are studied at the same time. If a sufficient number of different disease maps are produced for a particular study region, some diseases will show evidence of clusters purely by chance. This issue is not directly addressed by Buntinx and colleagues [7], although they make clear that any potential cancer clusters identified using their approach were carefully followed-up to check whether the result could be explained by data anomalies or known risk factors before being flagged as 'real' clusters. Recent statistical developments to extend Bayesian smoothing methods to *simultaneously* model joint spatial variations in risk of two or more diseases offer a more formal way to address this problem [12,13].

It is not always recognised that visualisation (e.g. choice of colour schemes) is very important for the perception of cancer maps. For example, red will symbolise danger for many people, whereas green may stand for 'no hazard', which will influence the intuitive inter-

pretation of the mapped risks. Since 'a good map of bad data looks better than a bad map of good data', it is imperative that the researcher makes sure that not only are the data and the analysis valid and accurate, but also presented in a sensible way.

Finally, it should be noted that spatial epidemiological studies using aggregated data are ecological, i.e. prone to ecological bias, associated with attempting to deduce individual-level effects from group-level data. Thus, geographical studies are valuable for hypothesis-generating, and although they cannot prove causality, they may provide guidance on how to design further detailed epidemiological studies, based on individual level data. Researchers need to recognise this and put their results into context when presenting geographical epidemiological studies.

References

1. Ferlay J, Bray F, Pisani P, Parkin MP. *GLOBOCAN 2000: Cancer Incidence, Mortality and Prevalence Worldwide, Version 1.0. IARC CancerBase No. 5*. Lyon, IARC Press, 2001.
2. Devesa, S.S., Grauman, D.G., Blot, W.J., Pennello, G., Hoover, R.N., Fraumeni, J.F., Jr., Atlas of Cancer Mortality in the United States, 1950–94 *NIH Publ No. (NIH) 99-4564*. NIH, 1999.
3. Monmonier M. *How to Lie with Maps*. Chicago and London, The University of Chicago Press, 1991.
4. Clayton DG, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987, **43**, 671–681.
5. Besag J, York J, Mollie A. Bayesian image restoration with applications in spatial statistics [with discussion]. *Annals of the Institute of Mathematical Statistics* 1991, **43**, 1–59.
6. Wakefield JC, Best NG, Waller LA. Bayesian approaches to disease mapping. In Elliott P, Wakefield JC, Best NG, Briggs DJ, eds. *Spatial Epidemiology: Methods and Applications*. Oxford, Oxford University Press, 2000 [chapter 7].
7. Buntinx F, Geys H, Lousbergh D, Broeders G, Cloes E, Dhollander D, De Beeck Lo, Vanden Brande, J, Van Waes A and Molenberghs G. Geographical differences in cancer incidence in the Belgian province of Limburg. *Eur J Cancer* (2003) **39**, 2058–2072.
8. Green PJ, Richardson S. Hidden Markov models and disease mapping. *Journal of the American Statistical Association* 2002, **97**, 1055–1070.
9. Jarup L, Best N, Toledano MB, Wakefield J, Elliott P. Geographical epidemiology of prostate cancer in Great Britain. *Int J Cancer* 2002, **97**, 695–699.
10. Julious SA, Nicholl J, George S. Why do we continue to use standardized mortality ratios for small area comparisons? *Journal of Public Health Medicine* 2001, **23**, 40–46.
11. Goldman DA, Brender JD. Are standardized mortality ratios valid for public health data analysis? *Statistics in Medicine* 2000, **19**, 1081–1088.
12. Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society A* 2001, **164**, 73–85.
13. Gelfand AE, Vounatsou P. *Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis. Technical Report*. Duke University, USA, Institute of Statistics and Decision Sciences, 2001.