A common error in the ecological regression of cancer incidence on the deprivation index

Gemma Renart,[1] *Marc Saez,*[1] *Carme Saurina,*[1] *Rafael Marcos-Gragera,*[1] *Ricardo Ocaña-Riola,*[2] *Carmen Martos,*[3] *Maria A. Barceló,*[1] *Federico Arribas,*[4] *and Tomás Alcalá*[4]

Suggested citation Renart G, Saez M, Saurina C, Marcos-Gragera R, Ocaña-Riola R, Martos C, et al. A common error in the ecological regression of cancer incidence on the deprivation index. Rev Panam Salud Publica. 2013;34(2):83–91.

ABSTRACT

Objective. *To determine if introducing age as another explanatory variable in an ecological regression model relating crude rates of cancer incidence and a deprivation index provides better results than the usual practice of using the standard incidence ratio (SIR) as the response variable, introducing the non-standardized index, and not including age in the model.*

Methods. *Relative risks associated with the deprivation index for some locations of cancer in Spain's Girona Health Region were estimated using two different models. Model 1 estimated relative risks with the indirect method, using the SIR as the response variable. Model 2 estimated relative risks using age as an explanatory variable and crude cancer rates as the response variable. Two scenarios and two sub-scenarios were simulated to test the properties of the estimators and the goodness of fit of the two models.*

Results. *The results obtained from Model 2's estimates were slightly better (less biased) than those from Model 1. The results of the simulation showed that in all cases (two scenarios and two sub-scenarios) Model 2 had a better fit than Model 1. The probability density for the parameter of interest provided evidence that Model 1 leads to biased estimates.*

Conclusions. *When attempting to explain the relative risk of incidence of cancer using ecological models that control geographic variability, introducing age as another explanatory variable and crude rates as a response variable provides less biased results.*

Key words Incidence; spatial analysis; risk; Spain.

In epidemiology, death or incidence ratios, standardized by factors known to confound the relationships of interest, are used to compare incidence and mortality in different geographic areas. The indirect method of standardization compares cases observed in a particular area with those one would expect to find within a certain reference population if the risks were the same for each age group. The standardizing factor is usually age distribution. The ratio of observed cases to expected cases, known as the standardized mortality ratio (SMR) or standard incidence ratio (SIR), is essentially a relative risk estimator for the area (i.e., an estimator of the risk of illness in an area in relation to the reference population). However, it has been demonstrated that problems arise with the use of age-adjusted rates in ecological regression models (1). Rosenbaum and Rubin (2) confirm that using standardized rates as a response variable in regression models leads to biased results because only the response values and not the predictor values are adjusted by the same confounding factor, usually distribution by age, resulting in what is known as the "mutual standardization problem." Anselin (3, 4) confirms that rates derived from both direct and indirect standard-

[1] Research Group on Statistics, Econometrics and Health (GRECS), Universitat de Girona, Girona, Catalonia, Spain. Send correspondence to Gemma Renart, gemma.renart@udg.edu
[2] Escuela Andaluza de Salud Pública (EASP), Granada, Spain.
[3] Centro Superior de Investigación en Salud Pública (CSISP), Valencia, Spain.
[4] Instituto Aragonés de Ciencias de la Salud (IACS), Zaragoza, Spain.

ization are calculated assuming a homogeneous relationship between risk and distribution by age in space (and time). The use of a non-standardized predictor variable implicitly assumes that the effect is constant across all strata of the confounding variable. Grisotto et al. (5), in line with Rosenbaum and Rubin (2), show that less biased estimators would be obtained by standardizing both the response and the predictor using the same variable, or by using the crude rates as response and including age in the regression models as one more explanatory variable.

The objective of this research was to show that introducing age as another explanatory variable in an ecological regression model relating crude rates of cancer incidence and a deprivation index provides better results than the usual practice of using the SIR as a response, introducing the non-standardized index, and not including age in the model.

MATERIALS AND METHODS

The current study was undertaken within the framework of the MEDEA project.[5] One of the project objectives was to estimate the relative risks associated with a deprivation index for some cancer locations in the Girona Health Region (GHR) in the province of Girona in northern Catalonia, an autonomous region in Spain, and to ascertain if the index could explain part of the spatial variability found in some of these locations (6, 7).

The analysis was performed on data provided by the Girona Cancer Registry (8, 9) for 1) incident cases of lung, tracheal, and bronchial cancer (codes C33–C34 in the International Classification of Diseases, 10th revision [ICD-10]); melanoma skin cancer (ICD-10 code C43); and non-Hodgkin's lymphoma (ICD-10 codes C82–C85 and C96) for men and women; 2) incident cases of larynx cancer (ICD-10 code C32) in men; and 3) breast cancer (ICD-10 code C50) in women.

All residents of the GHR (670 096 people, including 339 839 men and 330 257 women, according to the 2006 municipal population register) were included in the study population. The study took place from the years 1993 to 2006 (both inclusive), and the geographic area of analysis was the census tract.

The SIRs were calculated using the number of observed cases of the neoplasia of interest in the census tract i (with i = 1–500) during the period 1993–2006, and the number of expected cases of those diseases for the same tract. The reference population for the SIR was assumed to be the estimated population for each census tract in the GHR.

Although widely used, SIRs do have some limitations (7, 10, 11). These problems can be solved by smoothing. In this study, the Besag, York, and Mollié (BYM) model (12, 13) was used for Model 1, within a full Bayesian perspective. Two random effects were introduced into the model (spatial dependency and [nonspatial] unstructured variability) to gather all unexplained variability, and the parameters were assigned a probability distribution (prior distribution).

The method used to estimate the model parameters is explained in Annex 1.

To test the properties of the estimators and the goodness of fit of the two models, two scenarios and two sub-scenarios were simulated (Annex 2).

Model 1

Stratifying by sex, the BYM model (Model 1) was specified as a generalized linear mixed model (GLMM) with a Poisson response variable

$$O_i \sim Poisson(\mu_i E_i)$$

$$\log(\mu_i) = \alpha + \log(E_i) + \sum_{j=1}^{4} \beta_j Index_{ji} + S_i + \upsilon_i$$

where O_i is the number of cases observed in each census tract i; μ_i is the mean for the Poisson distribution ($E(O_i)$); E_i is the number of expected cases in each tract i; S_i is the random effect that captures the spatial variability; and υ_i is the random effect that captured the unstructured variability.

A deprivation index was introduced into the model as an explanatory variable to capture the specific socioeconomic contextual effects of geographic location on health. The index was constructed in accordance with the protocol established for the MEDEA project (14). $Index_{ji}$ in Model 1 denotes dummy variables relative to the quintile j for each census tract i of the deprivation index, and β_j is the associated parameter. The first quintile was used as the reference.

The random effects υ_i in Model 1 were independent and normally distributed, with zero mean and constant variance. For the random effect that captures the spatial variability, a conditional autoregressive model (CAR) S_i was used (15–17). The CAR model assumed dependency existed between neighboring areas, with "neighboring" defined as "adjacent" (18).

Model 2

For Model 2, crude rates were used as the response variable, incorporating the age structure of the population as an explanatory variable. The population was divided into five age groups (< 1 year, 1–14 years, 15–44 years, 45–64 years, and 65 years or more) corresponding to the five stages of health care (neonatal, child, young adult, adult, and old age). As cancer incidence is low for those under 14 years old, the first three age groups were combined into one cohort, for a total of three age groups (≤ 44 years, 45–64 years, and 65 years or older). This resulted in alternative specifications for Model 2

$$O_i \sim Poisson(\mu_i Pob_i)$$

$$\log(\mu_i) = \alpha + \log(Pob_i) + \sum_{j=1}^{4} \beta_j Index_{ji} + \gamma_1 P4564_i + \gamma_2 P65M_i + S_i + \upsilon_i$$

where Pob_i denoted the population of census tract i, $P4564_i$ denotes the percentage of the population between ages 45 and 64 years (both inclusive), and $P65M_i$ denotes the population 65 years or older. The first age group (≤ 44 years) was not included in the model to avoid problems of colinearity with the two other age groups.

RESULTS

The results obtained from the model estimators are shown in Table 1. Model 1 used the SIR as the response variable without standardizing the deprivation index. Model 2 used crude cancer rates as the response variable, and a non-standardized deprivation index, but included the age structure of the population.

Model goodness of fit, measured using the deviance information criterion (DIC), was very similar or slightly better

[5] Study of socioeconomic and environmental inequalities in small areas of cities in Spain and other European countries (http://www.proyectomedea.org/eng/medea.html).

TABLE 1. Results obtained from two ecological regression models relating crude rates of cancer incidence and a deprivation index, Girona Health Region, Catalonia, Spain, 1993–2006

Variable	Model 1[a]		Model 2[b]
Lung, tracheal, and bronchial cancer[c]			
Men			
$RR^{d}_{deprivation}$ (95% CI[e])			
Quintile 2	1.1633 (0.9468, 1.4304)		1.1140 (0.9097, 1.3652)
Quintile 3	1.5304 (1.2444, 1.8860)		1.1063 (0.9003, 1.3610)
Quintile 4	1.2678 (1.0380, 1.5499)		1.1783 (0.9629, 1.4442)
Quintile 5	1.4287 (1.1664, 1.7514)		1.2081 (0.9721, 1.5057)
Age group (years)			
45–64	—[f]		0.0468 (0.0065, 0.3384)
≥ 65	—		317.3329 (75.8287, 1326.0866)
Random effects (SD[g])			
Spatial	0.2033 (0.0539)		0.1658 (0.0564)
Unstructured	0.4449 (0.0299)		0.4313 (0.0286)
DIC[h]	2508.3046		2487.6556
Zero values (%)		13.33	
Women			
$RR_{deprivation}$ (95% CI)			
Quintile 2	1.3048 (0.7279, 2.4031)		1.1954 (0.6779, 2.1679)
Quintile 3	0.9726 (0.5245, 1.8380)		0.7962 (0.4350, 1.4862)
Quintile 4	1.1273 (0.6412, 2.0601)		0.8654 (0.4902, 1.5854)
Quintile 5	1.3829 (0.7879, 2.5171)		1.0533 (0.6050, 1.9060)
Age group (years)			
45–64	—		0.0089 (0.0000, 1.8985)
≥ 65	—		20.6843 (1.6346, 257.6425)
Random effects (SD)			
Spatial	0.2956 (0.1490)		0.1899 (0.1607)
Unstructured	0.3626 (0.3581)		0.4420 (0.3970)
DIC	1190.1819		1190.6651
Zero values (%)		60.0	
Larynx cancer[i]			
Men			
$RR_{deprivation}$ (95% CI)	1.2933 (0.9306, 1.8064)		1.2207 (0.8757, 1.7097)
Quintile 2	1.3956 (1.0088, 1.9425)		1.1980 (0.8584, 1.6810)
Quintile 3	1.4338 (1.0492, 1.9752)		1.3023 (0.9429, 1.8114)
Quintile 4	1.7018 (1.2505, 2.3348)		1.3810 (0.9864, 1.9453)
Quintile 5			
Age group (years)	—		0.0212 (0.0011, 0.3918)
45–64	—		537.7822 (65.3151, 4399.3923)
≥ 65			
Random effects (SD)			
Spatial	0.1612 (0.0647)		0.1544 (0.0580)
Unstructured	0.2063 (0.0859)		0.1799 (0.0828)
DIC	1321.5089		1319.5850
Zero values (%)		50.48	
Breast cancer[j]			
Women			
$RR_{deprivation}$ (CI 95%)			
Quintile 2	0.9222 (0.7741, 1.0984)		0.9259 (0.7870, 1.0890)
Quintile 3	1.1640 (0.9763, 1.3897)		0.8852 (0.7507, 1.0442)
Quintile 4	0.8900 (0.7492, 1.0573)		0.8652 (0.7335, 1.0211)
Quintile 5	0.8241 (0.6877, 0.9865)		0.8050 (0.6770, 0.9578)
Age group (years)			
45–64	—		1.3461 (0.2969, 6.1276)
≥ 65	—		51.7321 (23.0477, 116.3544)
Random effects (SD)			
Spatial	0.1262 (0.0400)		0.1202 (0.0390)
Unstructured	0.4385 (0.0256)		0.3847 (0.0240)
DIC	2604.262		2597.6874
Zero values (%)		10.00	
Melanoma skin cancer[k]			
Men			
$RR_{deprivation}$ (CI 95%)			
Quintile 2	0.9829 (0.6591, 1.4690)		0.9994 (0.6670, 1.5007)
Quintile 3	0.8300 (0.5519, 1.2493)		0.8477 (0.5563, 1.2935)

(Continued)

TABLE 1. *Continued*

Variable	Model 1[a]		Model 2[b]
Quintile 4	0.8108 (0.5428, 1.2134)		0.7946 (0.5250, 1.2066)
Quintile 5	0.8011 (0.5343, 1.2034)		0.7308 (0.4673, 1.1490)
Age group (years)			
45–64	—		3.1204 (0.0586, 158.1628)
≥ 65	—		12.4271 (0.6514, 235.3040)
Random effects (SD)			
Spatial	0.2047 (0.0768)		0.1970 (0.0653)
Unstructured	0.1275 (0.0601)		0.1305 (0.0635)
DIC	900.5892		904.2285
Zero values (%)		66.19	
Women			
$RR_{deprivation}$ (CI 95%)			
Quintile 2	0.8977 (0.5864, 1.3769)		0.9001 (0.5888, 1.3793)
Quintile 3	1.1733 (0.7837, 1.7661)		1.1417 (0.7613, 1.7228)
Quintile 4	0.9528 (0.6331, 1.4428)		0.9150 (0.6018, 1.4005)
Quintile 5	0.8567 (0.5534, 1.3252)		0.8221 (0.5260, 1.2859)
Age group (years)			
45–64	—		1.2185 (0.0265, 55.1081)
≥ 65	—		20.5916 (2.6945, 154.6464)
Random effects (SD)			
Spatial	0.4290 (0.1401)		0.3583 (0.1261)
Unstructured	0.3807 (0.1108)		0.3860 (0.1107)
DIC	1021.8240		1017.9371
Zero values (%)		61.90	
Non-Hodgkin's lymphoma[l]			
Men			
$RR_{deprivation}$ (CI 95%)			
Quintile 2	1.0941 (0.8005, 1.4975)		1.0514 (0.7691, 1.4395)
Quintile 3	1.3659 (1.0100, 1.8531)		1.1977 (0.8806, 1.6343)
Quintile 4	1.4154 (1.0577, 1.9032)		1.3262 (0.9865, 1.7969)
Quintile 5	1.0299 (0.7559, 1.4053)		0.8938 (0.6410, 1.2483)
Age group (years)			
45–64	—		0.0422 (0.0024, 0.7293)
≥ 65	—		819.3749 (101.6819, 6533.3545)
Random effects (SD)			
Spatial	0.1755 (0.0441)		0.1760 (0.0532)
Unstructured	0.3838 (0.0672)		0.3614 (0.0697)
DIC	1440.7600		1442.0572
Zero values (%)		42.38	
Women			
$RR_{deprivation}$ (CI 95%)			
Quintile 2	1.0192 (0.7185, 1.4497)		0.9887 (0.6971, 1.4059)
Quintile 3	1.1029 (0.7833, 1.5595)		0.9783 (0.6921, 1.3888)
Quintile 4	1.2299 (0.8879, 1.7150)		1.0686 (0.7631, 1.5063)
Quintile 5	1.0497 (0.7448, 1.4857)		0.8925 (0.6242, 1.2815)
Age group (years)			
45–64	—		0.0253 (0.0010, 0.6163)
≥ 65	—		158.6682 (30.3626, 838.8250)
Random effects (SD)			
Spatial	0.2091 (0.0776)		0.2091 (0.0752)
Unstructured	0.3585 (0.0835)		0.3468 (0.0868)
DIC	1275.0970		1283.1420
Zero values (%)		50.48	

[a] Response variable: standardized incidence ratio (SIR); predictor: non-standardized deprivation index.
[b] Response variable: crude incidence rates; predictors: non-standardized deprivation index and age.
[c] International Classification of Diseases, 10th Revision [ICD-10] codes C33–C34.
[d] RR: relative risk.
[e] CI: Bayesian confidence interval.
[f] Not applicable.
[g] SD: standard deviation.
[h] DIC: deviance information criterion.
[i] ICD-10 code C32.
[j] ICD-10 code C50.
[k] ICD-10 code C43.
[l] ICD-10 codes C82–85 and C96.

in Model 2. However, the standard errors of the random effect that captures the spatial variability and the standard deviations of the random effect that captures the nonstructured variability were lower when using Model 2 for all types of cancers except melanoma skin cancer in men and women, and tracheal, bronchial, and lung cancer in women (where they were higher), and non-Hodgkin's lymphoma (where they were almost equal).

The most significant differences were found in the statistical significance of the relative risk associated with the quintiles of the deprivation index. Thus, while the results of Model 1 indicate an association between the quintiles of the deprivation index and male cases of tracheal, bronchial, and lung cancer; larynx cancer; and non-Hodgkin's lymphoma, the statistical significances disappeared completely when using Model 2. Moreover, the relative risks of Model 2 were much lower (albeit without statistical significance) than those obtained in Model 1.

The only case where both methods of estimation provided a statistically significant association with the deprivation index (albeit only in the fifth quintile) was for incidence of breast cancer. However, the relative risk obtained in Model 2 was lower than that obtained in Model 1.

In the above-mentioned three cases where Model 1 revealed a significant association between the deprivation index and incidence of cancer (tracheal, bronchial, and lung; larynx; and non-Hodgkin's lymphoma—all in men), being in either of the two age groups included in Model 2 (45–64 years, and 65 years or older) appeared to be a significant predictor. Statistical significance was the same in all three cases: negative for the group aged 45–64 years, and positive for the group aged 65 years or older.

TABLE 2. Results of simulations testing the properties of the estimators and goodness of fit of two ecological regression models, Girona Health Region, Catalonia, Spain, 1993–2006

	Scenario 1		Scenario 2	
	Sub-scenario A	Sub-scenario B	Sub-scenario A	Sub-scenario B
Model 1				
Log of RRI[a]				
Median	0.991	0.008	0.976	–0.009
95% CI[b]	0.934, 1.049	–0.053, 0.069	0.928, 1.026	–0.065, 0.046
Coverage rate (%)[c]	92	84	56	68
Random effects (SD[d])				
Spatial	0.237	0.279	0.231	0.261
Unstructured	0.254	0.293	0.243	0.271
DIC[e]	2574.513	2570.395	2608.727	2606.805
n_{ef}[f]	237.012	235.763	236.251	243.139
Model 2				
Log RRI				
Median	1.001	–0.001	1.002	–0.001
95% CI	0.959, 1.041	–0.049, 0.048	0.953, 1.051	–0.055, 0.054
Coverage rate (%)	96	93	99	84
Random effects (SD)				
Spatial	0.196	0.214	0.208	0.236
Unstructured	0.206	0.223	0.220	0.248
DIC	2501.478	2495.402	2517.502	2505.003
n_{ef}[f]	188.836	189.473	197.276	194.670

[a] RRI: relative risk index.
[b] CI: Bayesian confidence interval.
[c] Percentage of times 95% CI contains the actual value of the parameter.
[d] SD: standard deviation.
[e] Deviance information criterion.
[f] Effective number of parameters.

Simulation

Even in the most favorable scenario (Scenario 1), the results for coverage (Table 2) and probability density for the parameter of interest (Figure 1) provide evidence that Model 1, which used the standard practice in ecological spatial regression (considering SIR as the response variable and a non-standardized deprivation index as an explanatory variable), provided biased estimates.

In all scenarios, Model 2 had a better fit than Model 1 in terms of lower DIC, effective number of parameters, and standard deviations of the random effects.

DISCUSSION

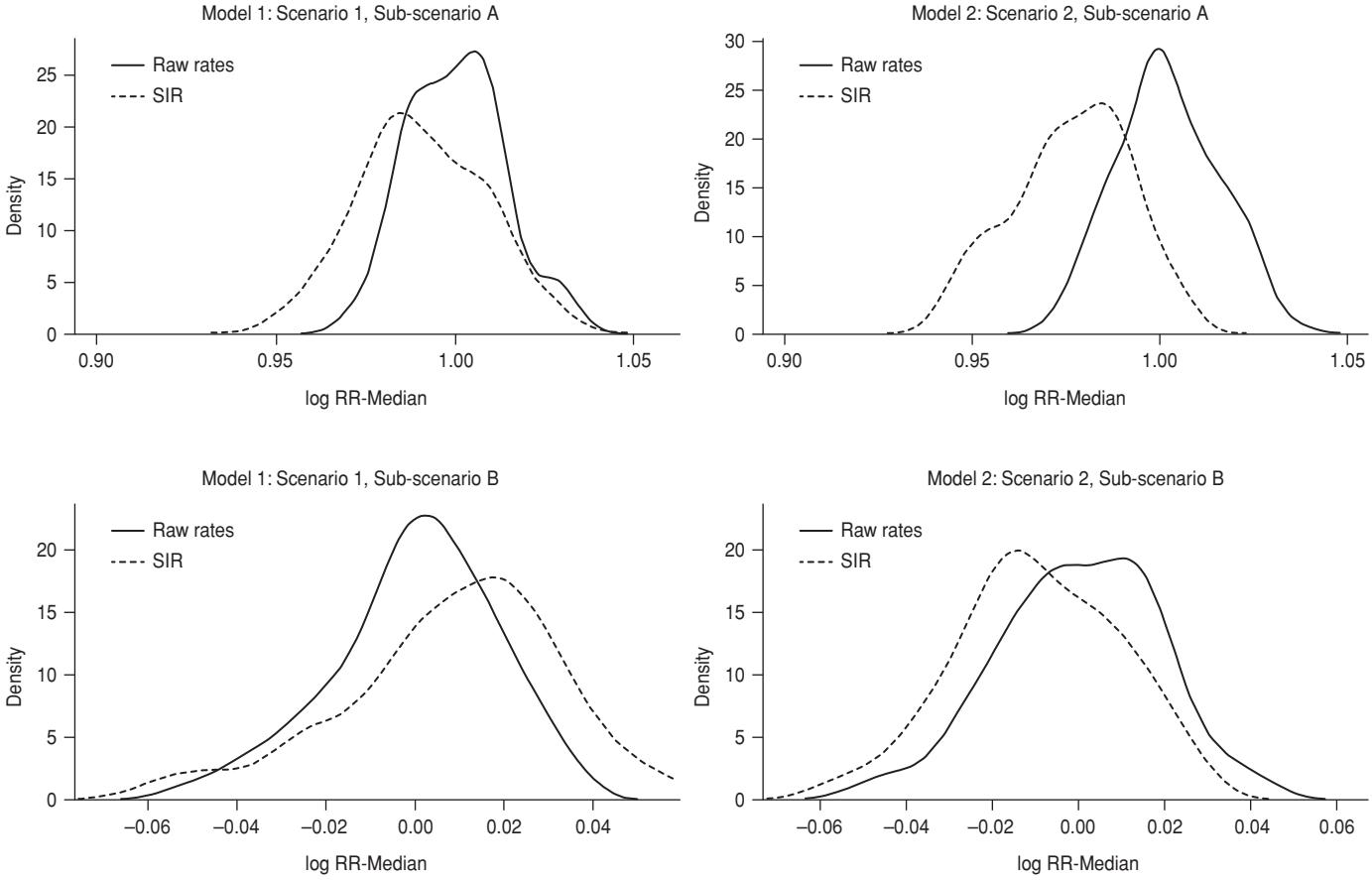
Technological progress and the availability of geographic information have allowed for the application of spatial epidemiology in the area of public health to identify areas with a higher risk of given health problems, using the SMR or the SIR, where incidence serves as an indicator for comparing risks in the different geographic areas studied. Various indices have been introduced in the models to explain geographic variability (12, 13, 19).

The method of indirect standardization of rates is frequently used to study the space–time distribution of incidence and mortality in small areas. However, various limitations exist that suggest this method should not be used in studies with geographic correlation, time series analysis of morbidity, or risk comparison between areas (20).

The SIR, like the SMR, is essentially a quotient of rates, adjusted using the direct method, where the numerator is the weighted mean for the specific study areas, or target population, and the denominator is the weighted mean for the rates specific to the external region, or non-target population. In this calculation, the standard or reference population corresponds to the area of study itself (the target population). The weighted values used to discern the weighted mean of specific rates derive from said population, so the reference population is never the external area or non-target population, as is often asserted in scientific publications. For this reason, SIRs for different geographic regions always have different reference populations, and the confusion bias resulting from the different population pyramids is always present when comparisons are attempted. Therefore, it is incorrect to claim that geographic areas with elevated SIRs show a higher incidence than areas with low SIRs. If the areas are not comparable, it is not possible to rank their values as synonyms of incidence frequency adjusted according to age and sex groupings.

In addition, calculating the SIR would only make sense when the specific rates of the target and non-target areas are

FIGURE 1. Simulations testing the properties of the estimators and goodness of fit of two ecological regression models (standard incidence ratio and crude cancer rates): probability density of the log of relative risk associated with the deprivation index, Girona Health Region, Catalonia, Spain, 1993–2006



proportional. This condition is especially difficult to verify when the indirect method is applied, because the specific target area (20) rates are not used. For this reason, studying the geographic distribution of incidence using the SIR or SMR may dilute important aspects of each stratum of the population and result in biased results (21).

These questions about the use of SIRs also affect the analysis of ecological associations. The majority of epidemiological studies model the logarithm of the mean number of cases observed according to the Napierian logarithm of the number of cases expected, which acts as an offset, plus a linear combination of explanatory variables. The parameters of the model are estimated through frequentist or Bayesian methods, and the exponential value of the linear combination of explanatory variables is risk, or the adjusted SIR. However, because the order of these values lacks epidemiological sense, it is inappropriate to equate a percentage increase or decrease in the SIR between two consecutive values of an explanatory variable to an increase or decrease in incidence adjusted by age or sex, because the compared areas have a different population pyramid (20).

Two scenarios with one sub-scenario each were simulated using two different models. Model 1 reproduced the standard practice, with the SIR as a response variable and the non-standardized deprivation index as an explanatory variable. In Model 2, crude rates were the response variable, and the non-standardized deprivation index (the explanatory variable) was adjusted by age.

The results of the simulation provide evidence that, even supposing that the effect of the explanatory variable (the deprivation index, in this case) is constant between strata of the confounding variable (age, in this study), which would theoretically be the most favorable scenario for standard approximation, the estimators used in Model 1 prove to be biased. In addition, goodness of fit, measured as both DIC and the size of the standard deviations of the random effects, proved to be much better in Model 2, as confirmed by the cancer incidence estimates from the various study sites.

Limitations

This study had some limitations. First, the duration of the study period was 14 years, during which time changes may have occurred in the geographic distribution of cancer incidence, the population pyramid, and the indicators that constitute the deprivation index. This could have produced bias in the results, which were obtained using models that do not consider dynamic behaviors because annually updated sources of census information are not available. Second, the deprivation index may not truly measure deprivation among women. However, even if it does not,

the study results indicate that the variance in statistical significance across age groups and index quintiles seems to function differently among women versus men. Third, the fact that the statistical significance of the breast cancer index does not disappear may be explained by the age cutoff points (45 and 64 years respectively) for incidence of two types of breast cancer (onset before age 45–50 and onset after age 45–50) and the fact that incidence among older age groups is better explained by the deprivation index. None of these limitations are likely to have rendered the main results of this work invalid, however, because the evidence suggests that both geographic variability of incidence and the deprivation index depend on age.

Conclusion

This study found that when attempting to explain the relative risk of incidence of cancer using ecological models that control geographic variability (meaning both spatial and nonstructural extra variability), crude rates should be used as a response variable and age included as another explanatory variable. The adjustments obtained would appear to be in line with Rosenbaum and Rubin (2), Anselin (3, 4), and Grisotto et al. (5), suggesting that the parameter estimators using SIRs as a response variable without standardizing the deprivation index (the standard focus) are biased and thus providing more evidence of the "mutual standardization problem" (biased results stemming from the use of standardized rates as a response variable in regression models). Introducing age as another explanatory variable in an ecological regression model relating crude rates of cancer incidence and the deprivation index provides better results because there is no reason to assume that the effect of each deprivation index quintile is constant across the different strata of the age variable (5). The current study results also suggest that geographic variability of incidence and the deprivation index depend on age. Therefore, the standard (SIR) focus should probably not be used, as its effect would not be independent of the variable associated with risk.

Acknowledgments. The authors appreciate the comments of the attendees at the XXVIII Reunión Anual de la Sociedad Española de Epidemiología (Valencia, Spain, October 27–29, 2010), where a preliminary version of this work was presented.

Funding. This work was partially supported by the Health Research Fund (*Fondo de Investigación Sanitaria*, FIS) of the Spanish Ministry of Science and Innovation, through projects ETS-06/90553, ETS-07/90453, and FIS-08/0142; the Agency for Technology Evaluation and Medical Research (*Agencia de Evaluación de Tecnología e Investigación Médicas*, AATRM); the Catalan Health Service of the Catalan Government (*Generalitat de Catalunya*), through project AATRM-006/01/2006; and the Second Programme of Community action in the field of Health (2008–2013) of the European Commission's Directorate General for Health and Consumer Protection (DG-SANCO), through project A/101156.

Conflicts of interest. None. All authors disclose any financial and personal relationships with other people or organizations that could inappropriately influence and/or bias their work.

REFERENCES

1. Morgenstern H. Ecologic studies. In: Rothman KJ, Greenland S, Lash TI, editors. Modern epidemiology. 3rd ed. Philadelphia: Lippincott-Williams & Wilkins; 2008.
2. Rosenbaum P, Rubin D. Difficulties with regression analyses of age-adjusted rates. Biometrics. 1984;40:437–43.
3. Anselin L, Lozano N, Koschinsky J. Rate transformations and smoothing. Working paper. Urbana-Champaign, Illinois: Spatial Analysis Laboratory, Department of Geography, University of Illinois; 2006.
4. Anselin L. How (not) to lie with spatial statistics. Am J Prev Med. 2006;30(2 Suppl):S3–6.
5. Grisotto L, Catelan D, Accetta G, Biggeri A. Material deprivation as marker of health needs. Statistica. 2010;70(3):343–52.
6. Borrell C, Marí-Dell'olmo M, Serral G, Martínez-Beneito M, Gotsens M; MEDEA members. Inequalities in mortality in small areas of eleven Spanish cities (the multicenter MEDEA project). Health Place. 2010;16(4):703–11.
7. Barceló MA, Saez M, Cano-Serral G, Martínez-Beneito MA, Martínez JM, Borrell C, et al. Métodos para la suavización de indicadores de mortalidad: aplicación al análisis de desigualdades en mortalidad en ciudades del Estado español (Proyecto MEDEA). Gac Sanit. 2008;22(6):596–608.
8. Unitat d'Epidemiologia i Registre de Càncer de Girona (ES). Cancer in Girona 2003–2004 [in Catalan]. Girona: UERCG-Pla Director d'Oncologia; 2009. (CanGir No. 2, March 2009).
9. Curado MP, Edwards B, Shin HR, Storm H, Ferlay J, Heanue M, et al., editors. Cancer incidence in five continents, vol. IX. Lyon: International Agency for Research on Cancer; 2007. (Scientific Publication No. 160).
10. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks. Biometrics. 1987;43(3):671–81.
11. Lawson AB, Browne WJ, Vidal-Rodeiro CL. Disease mapping with WinBUGS and MLwiN. Chichester, West Sussex: John Wiley & Sons; 2003.
12. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. Ann Inst Statist Math. 1991;43(1):1–59.
13. Mollié A. Bayesian mapping of disease. In: Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov chain Monte Carlo in practice. London: Chapman & Hall; 1996. Pp. 359–79.
14. Domínguez-Berjón MF, Borrell C, Cano-Serral G, Esnaola S, Nolasco A, Pasarín MI, et al. Construcción de un índice de privación a partir de datos censales en grandes ciudades españolas (Proyecto MEDEA). Gac Sanit. 2008; 22(3):179–87.
15. Besag J. Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc B. 1974;36(2):192–236.
16. Clayton DG, Bernadinelli L, Montomoli C. Spatial correlation in ecological analysis. Int J Epidemiol. 1993;22(6):1193–202.
17. Kelsall J, Wakefield J. Modeling spatial variation in disease risk: a geostatistical approach. J Am Stat Assoc. 2002;97(459):692–701.
18. Saez M, Saurina C. Estadística y epidemiología espacial. Girona: Edicions a Petició; 2007.
19. Elliot P, Best NG. Geographical patterns of disease. In: Armitage P, Colton T, editors. Encyclopedia of biostatistics. 2nd ed. Chichester: John Wiley & Sons; 2005. Available from: http://eu.wiley.com/legacy/wileychi/eob/articles.html
20. Ocaña-Riola R. Common errors in disease mapping. Geospat Health. 2010;4(2):139–54.
21. Ocaña-Riola R, Mayoral-Cortés JM. Spatiotemporal trends of mortality in small areas of Southern Spain. BMC Public Health. 2010; 10:26.

Manuscript received on 21 June 2012. Revised version accepted for publication on 16 July 2013.

ANNEX 1. Method used to estimate parameters for ecological regression models relating crude rates of cancer incidence and a deprivation index, Girona Health Region, Catalonia, Spain, 1993–2006

Spatial models were built as Bayesian hierarchical models with two stages (1). The first stage was the observational model $p(y \mid x)$, where y denotes the vector of observations and x denotes the unknown parameters following a Gaussian Markov random field (GMRF) denoted as $p(x \mid \theta)$. The second stage was given by the hyperparameters θ and their respective prior distribution $p(\theta)$. The desired posterior marginals

$$p(x_i|y) = \int_\theta p(x_i|\theta, y)p(\theta|y)d\theta$$

of the GMRF were approximated using the finite sum

$$\tilde{p}(x_i|y) = \sum_k \tilde{p}(x_i|\theta_k, y)\tilde{p}(\theta_k|y)\Delta_k \quad \text{(A1)}$$

where $\tilde{p}(x_i \mid \theta,y)$ and $\tilde{p}(\theta \mid y)$ denote approximations of $p(x_i \mid \theta,y)$ and $p(\theta \mid y)$, respectively. The finite sum was evaluated at support points θ_k using appropriate weights Δ_k.

The posterior marginal $p(\theta \mid y)$ of the hyperparameters is approximated using a Laplace approximation (2)

$$\tilde{p}(\theta|y) \propto \frac{p(x, \theta, y)}{\tilde{p}_G(x|\theta, y)} \mid x = x \times (\theta)$$

where the denominator $\tilde{p}_G(x_i \mid \theta,y)$ denotes the Gaussian approximation of $p(x \mid \theta,y)$ and $x \times (\theta)$ is the mode of the full conditional $p(x \mid \theta,y)$ (3).

According to Rue et al. (4), it is sufficient to "numerically explore" this approximate posterior density using suitable support points θ_k in the finite sum (A1). In this report, these points were defined in the H-dimensional space, using the central composite design (CCD) strategy. Here, center points were augmented with a group of star points, which allowed for estimating the curvature of $\tilde{p}(\theta \mid y)$ (4).

Here, to approximate the first component of the finite sum (A1), a simplified Laplace approximation (less expensive from a computational point of view, with only a slight loss of accuracy) was used (1, 4, 5).

The models were compared using the DIC (6)

$$DIC = goodness\ of\ fit + complexity = D(\bar{\theta}) + 2p_D$$

where $D(\bar{\theta})$ denotes the deviance evaluated at the posterior mean of the parameters and p_D denotes the "effective number of parameters," which measures the complexity of the model (6). The lower the DIC, the better the model.

The standard deviations of the spatial and the unstructured random effects, which reflect the extra variability captured by the model, were also observed.

(1) Schrödle B, Held L. A primer on disease mapping and ecological regression using INLA. Comput Stat. 2011:26(2): 241–58. Available from: http://www.r-inla.org/papers
(2) Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc. 1986;81(393):82–6.
(3) Rue H, Held L. Gaussian Markov random fields: theory and applications. Boca Raton: Chapman & Hall/CRC; 2005.
(4) Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Statist Soc B. 2009;71 (Pt 2):319–92. Available from: http://www.r-inla.org/papers
(5) Martino S, Rue H. Case studies in Bayesian computation using INLA. In: Mantovan P, Secchi P, editors. Complex data modeling and computationally intensive statistical methods. Milan: Springer Milan; 2010. Pp. 99–114. Available from: http://www.r-inla.org/papers
(6) Gilks WR, Richardson S, Spiegelhalter DJ, editors. Markov chain Monte Carlo in practice. New York: Chapman & Hall/CRC; 1996

ANNEX 2. Simulation testing properties of estimators and goodness of fit for ecological regression models relating crude rates of cancer incidence and a deprivation index, Girona Health Region, Catalonia, Spain, 1993–2006

To test the properties of the estimators and goodness of fit, two scenarios and two sub-scenarios were simulated for each of the models. In Scenario 1 it was assumed that the effect of the index was constant across age levels, whereas in Scenario 2 it was assumed that the effect of the index varied across age levels. For both scenarios, two sub-scenarios were simulated—one with the relative risk associated with the index as statistically significant (Sub-scenario A) and one in which it was not (Sub-scenario B). The GHR was used as the study area, with the same age and sex structure described above, and the previously constructed deprivation index was used as an explanatory variable, although in this case it was a continuous variable:

Scenario 1	Scenario 2
Sub-scenario A	Sub-scenario A
$\log(\mu_i) = -5.25 + Index + \log(E_i) + S_i + \upsilon_i$	$\log(\mu_i) = -5.25 + Index + \log(Pob_i) - 1.40P4565_i + 7.30P65M_i + S_i + \upsilon_i$
Sub-scenario B	Sub-scenario B
$\log(\mu_i) = -5.25 + \log(E_i) + S_i + \upsilon_i$	$\log(\mu_i) = -5.25 + \log(Pob_i) - 1.40P4565_i + 7.30P65M_i + S_i + \upsilon_i$

Index, *Pob*, *P*4565, *P*65*M*, *E*, *S*, and υ are defined above; the variable $\sigma_s = 0.20$ and the variable $\sigma_\upsilon = 0.25$ (i.e., they are chosen values, like those of the parameters, with the exception of the one associated with the deprivation index, based on estimates from a preliminary model).

A total of 100 Poisson variables were simulated, corresponding to $O_i \sim Poisson\ (\mu_i)$, for each of the four cases (two scenarios and two sub-scenarios).

RESUMEN

Un error frecuente en los modelos de regresión ecológica de la incidencia de cáncer con respecto al índice de carencia

Objetivo. Determinar si la introducción de la edad como otra variable independiente en un modelo de regresión ecológica que relaciona las tasas brutas de incidencia de cáncer con un índice de carencia, ofrece mejores resultados que la práctica corriente del uso de la razón de incidencia normalizada como criterio de valoración, con introducción del índice sin normalización y sin incluir la edad en el modelo.

Métodos. Se calcularon los riesgos relativos asociados con el índice de carencia de algunos tipos de cáncer en la Región Sanitaria de Girona en España, mediante dos modelos diferentes. En el modelo 1 se calcularon los riesgos relativos con el método indirecto, usando la razón de incidencia normalizada como criterio de valoración. En el modelo 2 se calcularon los riesgos relativos introduciendo la edad como una variable independiente y las tasas brutas de cáncer como criterio de valoración. Se simularon dos hipótesis y dos subhipótesis con el fin de verificar las propiedades de los estimadores y la bondad del ajuste de ambos modelos.

Resultados. Los resultados obtenidos a partir de las estimaciones con el modelo 2 fueron un poco mejores (menos sesgados) que los resultados obtenidos con el modelo 1. Los resultados de la simulación indicaron que en todos los casos (las dos hipótesis y las dos subhipótesis) el modelo 2 exhibió un mejor ajuste que el modelo 1. La función de densidad del parámetro de interés puso en evidencia que el modelo 1 da lugar a estimaciones sesgadas.

Conclusiones. Cuando se intenta explicar el riesgo relativo de incidencia de cáncer mediante modelos de regresión ecológica que tienen en cuenta la variabilidad geográfica, se obtienen resultados menos sesgados cuando se introduce la edad como una de las variables independientes y se utilizan las tasas brutas de incidencia como criterio de valoración.

Palabras clave Incidencia; análisis espacial; riesgo; España.