

Resources allocation in healthcare for cancer: a case study using generalised additive mixed models

Monica Musio¹, Erik A. Sauleau^{2,3}, Nicole H. Augustin⁴

¹*Dipartimento di Matematica ed Informatica, Università di Cagliari, Cagliari, Italy;* ²*Faculté de Médecine, Université de Strasbourg, Strasbourg, France;* ³*Registre du Cancer de Haut Rhin, Mulhouse, France;* ⁴*Department of Mathematics, University of Bath, Bath, UK*

Abstract. Our aim is to develop a method for helping resources re-allocation in healthcare linked to cancer, in order to replan the allocation of providers. Ageing of the population has a considerable impact on the use of health resources because aged people require more specialised medical care due notably to cancer. We propose a method useful to monitor changes of cancer incidence in space and time taking into account two age categories, according to healthcare general organisation. We use generalised additive mixed models with a Poisson response, according to the methodology presented in Wood, *Generalised additive models: an introduction* with R. Chapman and Hall/CRC, 2006. Besides one-dimensional smooth functions accounting for non-linear effects of covariates, the space-time interaction can be modelled using scale invariant smoothers. Incidence data collected by a general cancer registry between 1992 and 2007 in a specific area of France is studied. Our best model exhibits a strong increase of the incidence of cancer along time and an obvious spatial pattern for people more than 70 years with a higher incidence in the central band of the region. This is a strong argument for re-allocating resources for old people cancer care in this sub-region.

Keywords: health policy, generalised additive mixed models, resources allocation, cancer incidence, space-time models.

Introduction

Ageing of the population has a considerable impact on the use of health resources and has become an increasingly important health policy problem. Aged people require more specialised medical care due notably to cancer. The general principle of equity (Gillon, 1985) aims to guarantee allocation of healthcare resources on the basis of need. But when resources are limited and demand exceeds supply, allocation becomes a problem. Indeed if the financial global amount for healthcare is not increasing (or very little) whereas the need does, the question remains as to whether resources should be re-allocated in different healthcare fields (with some fields cut back so that others can expand). In most countries, health care is managed and administered by partially centralised health organisations. However, these decision makers may not be well equipped to make explicit rationing decisions and as such often rely on historical or political resource allocation processes, which can lead to sub-optimal use of the limited resources (Birch and Chambers, 1993). For decades, the problem of how to allocate healthcare resources in an equitable manner has

been the subject of concerted discussion and analysis (Petrou and Wolstenholme, 2000).

The question underlying the work presented here came from a healthcare authority which asked in 2009 the cancer registry of Haut-Rhin (a region in the north-east of France, covering approximately 700,000 inhabitants) for some predictions of cancer activity in the six regional main healthcare providers to help a financial reallocation between them and potentially between cancer care and other care activities. Indeed the general aim of a cancer registry is to routinely gather individual data on new cases of cancer. The real need of the healthcare authority was thus to determine the number of cancer cases to treat in the near future years, that is a need for prevalence prediction of cases. However, knowledge on disease prevalence supposes an epidemiological cross-sectional study in the general population which was not available. Moreover no additional epidemiological study was planned or funded. Hence, we assumed that the resources were allocated “ideally” to provide better care with respect to the burden of patients on health care providers. Then cancer incidence can be used to help to gauge the potential need in increasing resources.

We analyse yearly data on all cancer incidence collected between 1992 and 2007 from the cancer registry of Haut-Rhin. Children were not considered here since they are generally cared for in specific paediatric structures. It is important to be able to compare the

Corresponding author:
Monica Musio
Dipartimento di Matematica ed Informatica
Università di Cagliari, via Ospedale 72, 09124 Cagliari, Italy
Tel. +39 070 675 8521; Fax +39 070 675 8504
E-mail: mmusio@unica.it

evolution in time of incidence rates, across geographical boundaries and at as fine a spatial scale as possible, notably to decrease the risk of ecological bias. Furthermore major gains in life expectancy in France and improved health and living standards has made it necessary to re-plan the financial allocation taking into account patients age. Due to different healthcare organisation (different types of hospital wards for example) adult patients under study were divided into two categories according to their age: adults if the age ranges between 20 and 69 years (younger age category), and elderly if they are equal or more than 70 years old (older age category).

Statistical methods used to display the geographical patterns of mortality and disease incidence, are usually based on maps of estimates of relative risks, standardised incidence ratios (SIR) or standardised mortality ratios (SMR) under Poisson likelihood. The most common approaches involve hierarchical Bayesian models with random effects for each region. For examples, Clayton and Kaldor (1987) introduced hierarchical models and associated empirical Bayesian inference for region specific SMRs which allows spatial correlation between neighbouring regions. On the other hand, Besag et al. (1991) provided a widely applied expansion of the basic structure above, allowing very general applications of Poisson regression with correlated errors. The full advantage of this approach arises when when one wishes to include additional covariates and consider various spatial correlation structures to the relative risk parameters. Most studies have considered data aggregated over a period of time. Due to the availability of historical high quality data cancer recorded during the last 20 years, in recent years, interest relies on extending these spatial models to incorporate time trends and spatio-temporal interactions. The challenge is to incorporate adequately both temporal and spatial information to find out how the spatial patterns of the diseases evolve in time. Extension of hierarchical spatial models to space-time modelling of one or several diseases has been discussed by a number of authors (see for example Lagazio et al., 2001; Ugarte et al., 2009).

Our primary aim is to monitor changes of cancer incidence in space and time taking into account two age group categories. We assume that the repartition of the new cases of cancer in the two different age categories is not the same in all of the geographical units of the region and changes in time. Beyond these practical considerations, we would like to present a general approach which allows to account for possible effects of covariates and to include space-time interactions.

For this purpose, we use generalised additive mixed models combined with tensor product of the space-time dimension (Wood, 2004, 2006). Tensor products allow smoothness parameter selection to be independent of the different scales of the covariates. They provide a straightforward and simple generalisation of the uni-dimensional spline functions with the advantage to be scale invariant, so that they permit to model interactions between two or more variables which have different scales of measure. To investigate the effect of age categories in time and space we consider varying coefficient models (Hastie and Tibshirani, 1993). These models allow the smooth functions to be multiplied by some known covariate, continuous or not.

Several alternatives are considered here, to include time and space effects, and possible space-time interactions. Estimation for parameters uses maximum likelihood. To take into account the strong variability inside geographical units across time and age categories, adding a random effect acts as a random intercept for each unit. The choice of models and the necessity of including random effects can be assessed using stepwise addition/deletion of model terms as estimated using a generalised version of the Akaike Information Criterion (AIC) (Akaike, 1974) obtained by treating smooth functions as penalised fixed effects (for more details see Augustin et al., 2009).

Materials and methods

Data description and exploratory analysis

The Haut-Rhin cancer registry covers a “department” located in the north-east of France, adjacent to Germany and Switzerland. It is a region of 3,525 km² divided in a very dense irregular lattice of 377 municipalities (“communes”). Registry data used in this research were collected and validated between years 1992 and 2007, which we coded from 1 to 16.

Each case in the dataset is characterised by the patient’s geographical unit of residence, sex, date of birth and the date of diagnosis. We extracted the counts of cancer by age at the diagnosis, sex, year of diagnosis and geographical unit. We use the following notation. Suppose O is the number of observed cases and E the number of expected cases. The quantity E is fixed is calculated by internal standardisation method. The counts of cases, as initially gathered, O_{sctr} are indexed by the covariates sex $s \in \{1, 2\}$, five-years category of age $c \in \{1, \dots, 14\}$ (patients less than 20 years old have be excluded), year of diagnosis $t \in \{1, \dots, 16\}$, geographical unit $r \in \{1, \dots, 377\}$. The distribution of the

number of cases aggregated over age categories and years across the 377 geographical units, is very heterogeneous, varying between 3 and 11,470 cases, with a mean of 207.9 cases while the median is 72. Population counts are known by age, sex and geographical unit for 1990, 1999 and 2004 (national census). We used the 1990 population for 1992, the 1999 population for 1998, 1999 and 2000, and the population of 2004 for 2004 to 2007. Linear interpolations of the census populations were used to estimate the populations of the other years. The sex is not, with respect to planning for cancer healthcare resources, a relevant covariate even if it plays a role in the distribution of the cancer incidences. Thus we calculate expected counts using a sex-adjusted risk: letting R_{sctr} denotes the population counts corresponding to O_{sctr} , the corresponding expected counts is obtained as $E_{sctr} = \hat{p}_s O_{sctr}$ where \hat{p}_s is the sex-adjusted risk, estimated in the whole investigated area, for all age categories and for all years as:

$$\hat{p}_s = \frac{\sum_c \sum_t \sum_r O_{sctr}}{\sum_c \sum_t \sum_r R_{sctr}}$$

Observed and expected cases are then aggregated on sex and on two age categories: between 20 and 69 years for the first one and ≥ 70 years for the second. We denote such aggregated variables as O_{atr} for observed cases and E_{atr} for expected cases, where a is 1 for 20-69 years and 2 for ≥ 70 years, while the indexes t and r vary as before. From the 78,366 registered cases between 1992 and 2007, 44,817 are in the younger age category and 33,549 in the older. Due to covariates, the data set counts are spread over 12,064 cells with 2,346 zeros (19% of cells have zeros). The cohort dimension is important when analysing time trend (Dreassi et al., 2005; Catelan et al., 2006; Biggeri et al., 2009). However, the inclusion of the cohort in the analysis will increase the size of the dataset and consequently the number of zero-cells. For this reason we consider the age-time interaction as a proxy for the cohort effect.

The raw SIRs have a standard deviation in each geographical unit which varies between 0.98 and 6.77.

An exploratory study of such data reveals a temporal variation in the standardised incidence ratios (Fig. 1). The plot shows different levels of risk for age categories and clear evidence of an increasing but moderate trend in time for both categories of age. To visualise a possible spatial effect we then consider the SIRs aggregated over age a and year t . The resulting descriptive plot is shown in Fig. 2. The left panel is a choropleth map based on quartiles of SIRs whereas the

right panel is based on the significance (at 5%) of these SIRs, under the Poisson assumption. Both of these plots exhibit a higher risk in the north and in the centre parts of the region and lower risk in the south.

Statistical model and statistical analysis

Generalised additive mixed models (GAMMs) combine the flexibility of generalised additive models (Hastie and Tibshirani, 1990) for modelling the relationship between covariates and the response, with the ability of mixed models to model random effects. Since the work of Lin and Zhang (1999), GAMMs have been used to model overdispersed and spatial correlated data. This class of models uses additive non parametric functions to model covariate effects while the

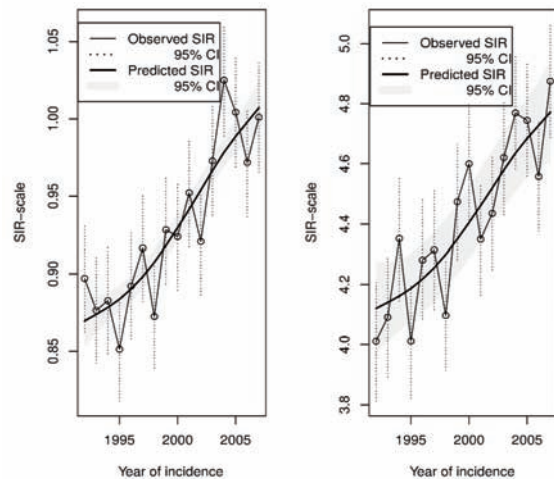


Fig. 1. Raw SIRs aggregated over geographical units versus year according to age categories (younger on the left and older on the right) and predicted SIRs using the best model (Model 4).

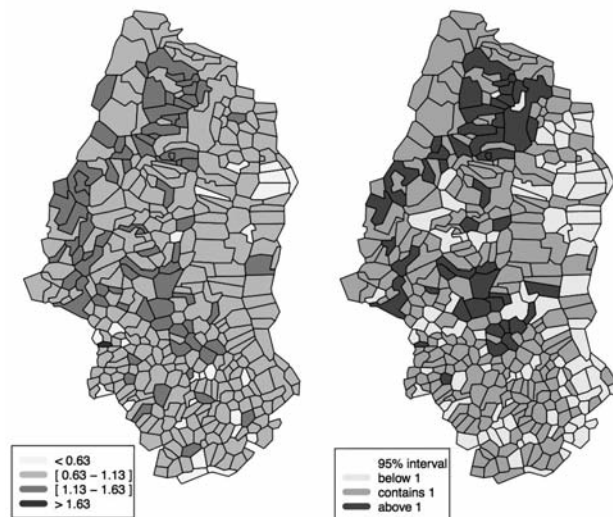


Fig. 2. Raw SIRs aggregated over time and age categories.

introduction of random effects make it possible to account for overdispersion caused by unobserved heterogeneity. In our specific case, the response variable O_{atr} (random variable representing the observed number of cancer in region r , year t and age category a) is supposed to follow a Poisson distribution with mean $E_{atr}e^{\mu_{atr}}$, where $e^{\mu_{atr}}$ represents the relative risk associated with people living in region r , in year t and for age category a . We thus assume the following model:

$$O_{atr} \sim P(E_{atr}e^{\mu_{atr}})$$

implying that $\log O_{atr} = \log E_{atr} + \mu_{atr}$, where $\log E_{atr}$ acts as an offset. To complete the model we have to specify the functional form of the predictor μ_{atr} . In all our models, we include fixed age effect $\omega_a \beta_a$ and an overall constant β_0 . Considering the main smoothed effects:

- (i) The time effect is modelled by a cubic regression spline on the year of diagnosis with 14 knots.
- (ii) For the spatial effect, we use a smoother with a thin plate regression spline basis on the centroid coordinates of the geographical unit each case belongs to. The thin plate regression splines are constructed by a simple truncation of the basis that arises from the thin plate smoothing problem (Wood, 2004). The basis is isotropic, hence a rotation of the coordinates system will not affect the results and for this reason it is often used for smoothing spatial effect on geographical coordinates.

Considering the interactions:

- (i) Since the units of time (year) and space (km) are different, the smoother needs to be invariant to their relative scaling. For this purpose we use tensor products allowing smoothness parameter selection to be independent of the different scales of the covariates.
- (ii) To model interactions between age and space or age and time we use varying coefficient models. The implementation of varying coefficient models is straightforward: each row of the model matrix for the smooth is multiplied by a dummy variable for each age category. We add a random effect to account for the variability inside each geographical unit. Among the models tested we have the following two:

$$\mu_{atr} = \beta_0 + \omega_a \beta_a + \omega f_1(east_r, north_r, year_t)$$

$$\mu_{atr} = \beta'_0 + \omega'_a \beta'_a + \omega f_2(east_r, north_r) + f_a(year_t) + Z_r \cdot b_r$$

(Eq. 1)

where ω_a is the dummy variable associated with the older age category (the younger age category acts as reference), ω is the matrix of the two dummy variables for age, β_0 and β'_0 are estimated overall means, β_a and β'_a are fixed parameters for age, f_1 is a multidimensional smooth functions of easting and northing and year (one per each age category), the function f_2 is a bi-dimensional smooth function of the geographical coordinates easting (east) and northing (north), f_a is an arbitrary smooth function of covariate year while Z_r is a row of a spatial random effect model matrix. It is further assumed that random effects b are independent and have a normal distribution with mean 0. The distribution of the random effects b are then completely characterised by their variance-covariance matrices Φ , $b \sim N(0, \Phi)$. These random effects act as a spatial centred random intercept.

For estimation, we use the fact that the smoothed model terms can be represented as random effects, allowing their estimation via standard mixed modelling software (Lin and Zhang, 1999; Wood, 2004).

We build and compare different models. We include in all the models a fixed effect for age β_a and a grand mean. We then add different effects involving space, time and age with or without interactions. Into these models we add further the spatial random component, assuming independent and identically normal distributed effects with 0 mean and variance matrix $\Phi = \sigma^2 I$. We use AIC for comparing all the models tested. The analysis has been performed using the R packages mgcv, gamm4 (Wood, 2006, 2011) and geoR (Ribeiro and Diggle, 2001).

Results

As shown in Table 1, which summarises the results of our model selection analysis, random effects have to be included (models with random effects have a lower AIC). The best selected model is model 4. Its predictor is as equation 1. A more complex model, including

Table 1. Values of the AIC for each model fitted.

Model	Without random effect	With random effect
$\mu_{atr} = b_0 + \omega_a \beta_a$... $Z_r b, b \sim N(0, \sigma^2 I)$
1. $f_1(east_r, north_r, year_t)$	14.608,27	14.458,85
2. $f_2(east_r, north_r) + (f_3 year_t)$	14.645,70	14.471,96
3. $f_2(east_r, north_r) + (\omega f_3 year_t)$	14.648,02	14.474,38
4. $\omega f_2(east_r, north_r) + (f_3 year_t)$	14.602,13	14.416,93
5. $\omega f_2(east_r, north_r) + (\omega f_3 year_t)$	14.604,49	14.419,41
6. $\omega f_1(east_r, north_r, year_t)$	14.594,72	14,436,66

age-year interaction (model 5) results in a very similar fit than model 4 (AIC is less than 3 higher), while modelling space-time interaction with a three-dimensional tensor product smooth (model 1 or model 6) gives rise to a worst fit (in terms of AIC). Finally the three models including an age-space interaction (models 4, 5 and 6) are better than the three models including a main spatial effect only (models 1 to 3) with some great improvements in AIC (about 30). To visualise the effects of time estimated in model 4, we consider as an estimate of the log-SIR for younger age at year t , in the whole of region, the quantity $\log \mu_{at\bullet} = \hat{\beta}_0 + f_1(\text{year}_t)$, where β_0 is the constant in the model (estimated as -0.437) and for older age $\log \hat{\mu}_{at\bullet} = \hat{\beta}_0 + \hat{\beta}_a + f_1(\text{year}_t)$. Here β_a is the fixed effect parameter for age (the younger age category acts as reference). This estimate is 1.560 (with standard error 0.012). In both cases, the \bullet in $\mu_{at\bullet}$ indicates that the SIR is calculated whatever the geographical unit is. Fig. 1 plots such estimates of -SIRs and allows to compare predicted with observed -SIRs. Fig. 3 plots the age-space interaction for the model 4 and it exhibits a very different spatial pattern according to age. Younger new patients are more concentrated on the north and west part of the region and older new patients more in a central vertical band of the region, where two main cities and two main healthcare providers are. More precisely, this plot corresponds to the estimation of the smoother on a very fine grid (and not at the centroids location as estimated in the model). To model checking we use deviance residuals. We check zero mean of residuals and plot fitted values against observed cases. To assess a possible presence of a residual spatial structure we draw an omnidirectional semi-variogram. Finally the Poisson assumption is verified by plotting the deviance residuals against their Poisson theoretical quantiles (Garcia Ben and Yohai, 2004; Augustin et al., 2011). The model diagnostic plots (unreported), suggest that the model assumptions are generally valid.

Concerning the random effect in b , its range is -0.168, 0.209. The standard deviation of this effect among geographical units is estimated to be 0.092, much lower than the standard deviation within geographical units which is estimated as 1.033, indicating that the variability inside each unit is higher (because of replications along time and age) than the variability between.

The plot of the random effect on Fig. 4 shows a similar spatial repartition than observed SIRs (left panel of Fig. 2) indicating that a constant amount of variability is explained by the smoothing model and that remains a little part of variation for each geographical unit, taken into account by the random effect.

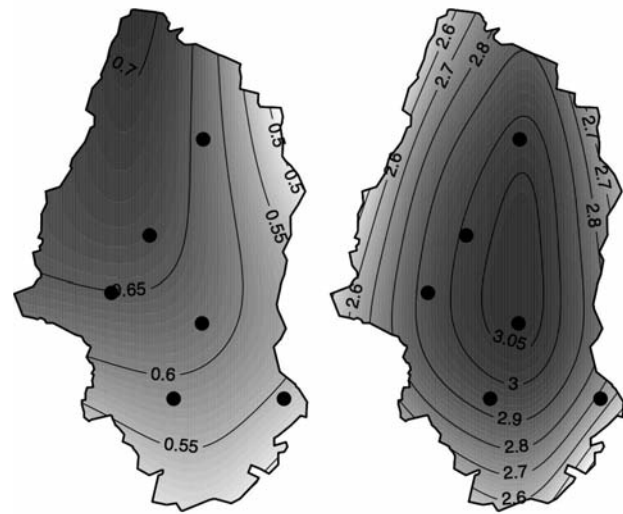


Fig. 3. Model 4: predicted spatial effect (exponential of the values) for both of age categories (black dots are location of public health care providers). The left plot shows predictions for the younger age category, the right plot shows predictions for the older age category.

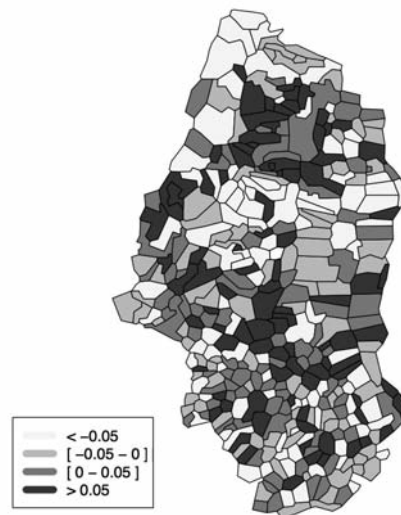


Fig. 4. Model 4: estimated random effect.

Discussion

The main issue of our work is to tackle the problem of potential resources re-allocation between six health-care providers in a given region, concerning cancer cares. This organisation of care splits hospital wards for younger adults (20-69 years) and for older adults (≥ 70 years). We do not address here the particular issue of children cancer cares, because of its specificity. Our modelling approach allows to exhibit different trends and effects. Whatever the age is, the incidence of cancer increases along time. There is no interaction between space and time but the geographical pattern

of new cases is different from 20-69 years and for ≥ 70 years. For the younger category, the incidence tends to be higher in a large north-west part of the region, even if this effect is relatively slight. To the contrary the incidence is obviously higher for older category of age in the central band of the region where the two main healthcare providers are. This is a strong argument for re-allocating in these two hospitals resources for old people cancer care. All checks on the best model we selected are reassuring (residuals, plot of observed against predicted effects, etc.), prove its correct fit to the data and improve our self-confidence in our results. However, the straightforward way to achieve the main goal of re-allocation is to represent on maps the burden of needed cancer cares for each future year. But this supposes that the prevalence in geographical units is available which is not the case: the knowledge of the cancer prevalence implies a permanent gathering of cancer cases with health status on all the population. Thus we would have to collect these morbidity data coming for all healthcare providers and at least all general practitioners in the region and to clean the datasets for n-uplons (because in France there is not really a unique identifier). Finally a prediction model would have to be chosen in addition to the smoothing model. An alternative is to estimate this prevalence, using incidence and survival (Feldman et al., 1986; Gail et al., 1999; Phillips et al., 2001). For example, incidence rate, ζ and prevalence proportion, π , are related by:

$$\pi = \frac{\zeta \bar{D}}{1 + \zeta \bar{D}}$$

where \bar{D} is the mean duration of disease. In our case, the problem is that the incidence is distributed in two categories of age, 14 years and 377 geographical units and so does the survival. This splitting implies a model for stabilisation of the estimates, which would make more difficult the answer to the initial problem. Thus, if the resources are supposed to be allocated right now, ideally to the healthcare providers, the estimation of incidence may be sufficient to decide how to distribute the excess or the decrease of additional patients. A further assumption in our approach is that we use the geographical unit of residence of the new cases for guessing where they will be cared for. In fact, the different healthcare providers in the region have different technical levels and different skills. It might be possible that certain patients (for example patients with the high severity of illness or extended tumour) is cared for not in the most proximate hospital but in another one. In the same spirit we do not take into account the

accessibility to the provider but implicitly only the distance from the residence.

The general method using GAMMs allows to include complex models for effects with their interactions but can also deal with more complex correlation structures to model cancer incidence data in which a strong space and/or time effect is present. For example, instead of smoothing time effect with a spline, assuming a temporal autoregression on errors could be an interesting alternative.

References

- Akaike H, 1974. A new look at the statistical model identification. *IEEE TRANS AUTOMAT CONTR* 19, 716-723.
- Augustin NH, Musio M, von Wilpert K, Kublin E, Wood SN, Schumacher M, 2009. Modelling spatio-temporal forest health monitoring data. *J Am Stat Assoc* 104, 899-911.
- Augustin NH, Sauleau EA, Wood SN, 2011. On quantile-quantile plots for generalised linear models. *Comput Stat Data Anal* 56, 2404-2409.
- Besag J, York J, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math* 43, 1-59.
- Biggeri A, Catelan D, Dreassi E, 2009. The epidemic of lung cancer in Tuscany (Italy): a joint analysis of male and female mortality by birth cohort. *Spat Stat Epidemiol* 1, 31-40.
- Birch S, Chambers S, 1993. To each according to need: a community based approach to allocating health care resources. *Can Med Assoc J* 149, 607-612.
- Catelan D, Biggeri A, Dreassi E, Lagazio C, 2006. Space-cohort Bayesian models in ecological studies. *Stat Model* 6, 159-173.
- Clayton D, Kaldor J, 1987. Empirical Bayes estimates of age-standardised relative risks for use of disease mapping. *Biometrics* 43, 671-681.
- Dreassi E, Biggeri A, Catelan D, 2005. Space-time models with time-dependent covariates for the analysis of the temporal lag between socio-economic factors and lung cancer mortality. *Stat Med* 24, 1919-1932.
- Feldman AR, Kessler L, Myers MH, Naughton MD, 1986. The prevalence of cancer: estimates based on the connecticut tumor registry. *N Engl J Med* 315, 1395-1397.
- Gail MH, Kessler L, Midthune D, Scoppa S, 1999. Two approaches for estimating disease prevalence from population based registries of incidence and total mortality. *Biometrics* 55, 1137-1144.
- Garcia Ben M, Yohai VJ, 2004. Quantile-quantile plot for deviance residuals in the generalised linear model. *J Comput Graph Stat* 13, 36-47.
- Gillon R, 1985. Justice and allocation of medical resources. *Br Med J* 291, 266-268.
- Hastie T, Tibshirani R, 1990. Generalised additive models.

- Chapman and Hall, London, UK.
- Hastie T, Tibshirani R, 1993. Varying-coefficient models. *J Roy Stat Soc B* 55, 757-796.
- Lagazio C, Dreassi E, Biggeri A, 2001. A hierarchical Bayesian model for space-time variation of disease risk. *Stat Model* 1, 17-29.
- Lin X, Zhang D, 1999. Inference in generalised additive mixed models by using smoothing splines. *J Roy Stat Soc B* 61, 381-400.
- Petrou S, Wolstenholme J, 2000. A review of alternative approaches to healthcare resource allocation. *Pharmacoeconomics* 18, 33-43.
- Phillips N, Coldman A, McBride ML, 2001. Estimating cancer prevalence using mixture models for cancer survival. *Stat Med* 21, 1257-1270.
- Ribeiro Jr PJ, Diggle PJ, 2011. GeoR: a package for geostatistical analysis. *R-NEWS*, 1, 14-18. Available at: <http://CRAN.Rproject.org/doc/Rnews/> (accessed on June 2001).
- Ugarte M, Goicoa T, Ibanez B, Militino A, 2009. Evaluating the performance of spatio-temporal Bayesian models in disease mapping. *Environmetrics* 20, 647-665.
- Wood SN, 2004. Stable and efficient multiple smoothing parameter estimation for generalised additive models. *J Am Stat Assoc* 99, 673-686.
- Wood SN, 2006. Generalised additive models: an introduction with R. Chapman and Hall/CRC.
- Wood SN, 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalised linear models. *J Roy Stat Soc B* 73, 3-36.