Radiat Environ Biophys (2007) 46:205–213
DOI 10.1007/s00411-007-0109-0

ORIGINAL PAPER

A short review of model selection techniques for radiation epidemiology

Linda Walsh

Received: 23 November 2006 / Accepted: 24 March 2007 / Published online: 28 April 2007

Abstract A common type of statistical challenge, widespread across many areas of research, involves the selection of a preferred model to describe the main features and trends in a particular data set. The objective of model selection is to balance the quality of fit to data against the complexity and predictive ability of the model achieving that fit. Several model selection techniques, including two information criteria, which aim to determine which set of model parameters the data best support, are reviewed here. The techniques rely on computing the probabilities of the different models, given the data, rather than considering the allowed values of the fitted parameters. Such information criteria have only been applied to the field of radiation epidemiology recently, even though they have longer traditions of application in other areas of research. The purpose of this review is to make two information criteria more accessible by fully detailing how to calculate them in a practical way and how to interpret the resulting values. This aim is supported with the aid of some examples involving the computation of risk models for radiation-induced solid cancer mortality fitted to the epidemiological data from the Japanese A-bomb survivors. These examples illustrate that the Bayesian information criterion is particularly useful in concluding that the weight of evidence is in favour of excess relative risk models that depend on age-at-exposure and excess relative risk models that depend on age-attained.

L. Walsh (✉)
Institute of Radiation Protection, GSF National Center
for Environment and Health, Ingolstädter Landstraße 1,
85764 Neuherberg, Germany
e-mail: Linda.Walsh@gsf.de

Introduction

Poisson regression, involving a multivariate analysis of numbers of uncommon events (e.g., the incidence of cancer or the mortality from cancer) in cohort studies, is often applied to the field of radiation epidemiology. With this method it is possible to determine several theoretical models given relevant epidemiological data, for example, for the relative risk of dying from lung cancer for smokers who live in areas associated with high radon levels. The decision concerning which of these models is the most plausible, without necessarily considering the preferred values of the model parameters, can be made with model selection techniques. Within each model, the parameters indicate the importance of particular effects, for example, the age dependence of the spontaneous death rate in the absence of radon and smoking, or the change in the death rate with radon exposure levels and/or smoking levels. Such parameters are not usually predicted by prior knowledge, but need to be estimated from the data in order to determine which combination of (explanatory) covariables, if any, is capable of adequately describing the total detrimental health risks. Current data may not be statistically powerful enough to constrain the parameters of the model at the required level. Alternatively, although less common in epidemiology than in other fields, the presence of good data may lead to the very different problem of determining when to stop adding extra useful parameters, or when to stop re-parameterisation procedures. In this case, it is possible then to arrive at several competing models that seem to fit the data approximately equally well. Occam's razor (also known as the principle of parsimony) provides a solution to model selection here—the simpler model should be preferred. A complicated model that explains the data slightly better than a simpler model needs to be penalised for the extra parameters which tend to decrease the overall power of a model in making predictions. In contrast, a model that is too simple and unable to fit the data well needs to be discarded. Such considerations and problems associated with preferred model selection are widespread across many areas of research and form a common type of statistical challenge.

The standard approach to model fitting usually involves choosing one initial set of parameters to be varied and then using a likelihood method to determine the best-fit model and associated parameter confidence intervals. Eventually, the initial parameter set may be replaced by another set chosen ad hoc and the whole process repeated an ad hoc number of times. Typically, the introduction of extra parameters will often improve the fit to the data set, regardless of the relevance of these new parameters, and so a simple comparison of maximum likelihoods will generally tend to favour the model with the most parameters[1]. A less commonly adopted approach, which compensates for this effect by penalising models which have more parameters, and therefore counterbalances the improvement in maximum likelihood that the extra parameters may allow, is that of model selection.

A considerable wealth of the statistical literature is devoted to model selection (excellent text book accounts have recently been given [1–3]) and its use is widespread in many branches of science. In model selection, the data are involved in allowing the determination of which combination of parameters gives the preferred fit. Here, the emphasis is placed on the application of information criteria to aid in the elimination of parameters that do not play a sufficient role in improving the fit to the available data. These information criteria have led to considerable advances in the understanding of how statistical inference is related to information theory.

The model selection techniques reviewed here aim to determine which set of parameters the data support by computing the probabilities of the different models, given the data, rather than considering the allowed values of the fit parameters. Choice of the technique depends on the nesting properties of the competing models. Nested models are those where the more complicated model has additional parameters to those in the simpler model and where the latter may be interpreted as a particular case of the former with the additional parameters kept fixed at some fiducial values. Several techniques are reviewed that apply to linear and non-linear models including: ‘‘Likelihood ratio’’ tests, which *require the models to be nested* and were originally proposed by Neyman and Pearson [4] (see, however, [5] for a modern textbook explanation); and two likelihood based information criteria [1–3] which *do not require the models to be nested*. These information criteria due to Akaike [6, 7] and Schwarz [8] arise from extending the likelihood-based methods by information theoretical and Bayesian considerations, respectively. These criteria have only recently been applied to the field of radiation epidemiology [9, 10], even though they have longer traditions of application in other areas of research (e.g. [11, 12]). Although the underlying theoretical considerations associated with information criteria are very involved (and are not covered in detail here—but just described and cited), the actual criteria have very simple expressions and are easy to derive from the standard output of most optimisation software. The purpose of this review is to promote the application of these techniques in the field of radiation epidemiology by aiming to increase their accessibility and by fully describing how to calculate them and how to interpret the resulting values. This is done with the aid of some practical examples involving the epidemiological data from the Japanese A-bomb survivors.

Model selection statistics

In Poisson regression, it is possible to specifically model rate functions for grouped survival data. Let d_i, P_i and x_i denote the number of deaths (or cases), the total number of person–years at risk and covariates (e.g., age and dose) for the ith data cell, respectively. Then the model for the expected number of deaths $E(d_i)$ in the cell can be written as:

$$E(d_i) = P_i\lambda(\beta, x_i),$$

where $\lambda(\beta, \mathbf{x})$ is the rate function model and β is the chosen set of fit parameters.

If $\hat{\beta}$ represents the computed optimised values of the fit parameter set, then the contribution of the ith data cell to the log likelihood is

$$L_i = d_i \ln(P_i\lambda(\hat{\beta}, x_i)) - P_i\lambda(\hat{\beta}, x_i)$$

and the log likelihood is simply the sum of L_i over the i data cells, ΣL_i.

The overall quality of a model fit to the data in Poisson regression is often quantified by the deviance, dev. The deviance contribution from the ith data cell is computed as twice the difference between the likelihood contribution, when d_i is used as the estimate of the cell mean, and the value of L_i for the current model. Thus, the total deviance is minus twice the natural logarithm of the maximum likelihood, $M = \max(\Sigma L_i)$.

[1] However, this should not give the impression that the standard model selection approach involving maximum likelihoods pays no attention to the number of fit parameters, which, in fact, determines the number of degrees of freedom, as explained below.

As indicated above, the general problem of choice of the procedure to use for selection of a preferred model in Poisson regression usually depends on whether the competing models are ‘‘nested’’. Model A is nested within model B, if model A is a special case of model B, i.e., if both contain the same model parameters and model B has at least one additional parameter.

Nested models

When two models are nested, it is known that the *difference* between their deviances, dev(B) – dev(A) is chi-square (χ^2) distributed [4, 5]. In this case, *the degrees of freedom for the difference* is equal to the difference in degrees of freedom for the two original test statistics, $df(B)$ – $df(A)$. This suggests the most commonly used method for comparing the fit of two nested models to see if a particular parameter can be dropped from a model without substantially reducing the explanatory power of the model: one tests whether the resulting difference in deviance, dev(B) – dev(A) is significant or not, for the given degrees of freedom and a chosen level of statistical significance. If the difference is significant, then the extra parameters associated with model B are retained. This method is also known variously as partitioning the deviance and applying likelihood ratio tests [4, 5] and is not strictly applicable to non-nested models. Correspondingly, model B with one additional fit parameter than model A is considered to be an improvement over model A with 95% probability, if the deviance is reduced by more than 3.84 points. This is because the χ^2-probability distribution with one degree of freedom leaves 5% of the total probability excluded, and consigned to the tail of the total distribution, at χ^2 = 3.84. This and a few other examples are given in Table 1.

Non-nested models

It is sometimes possible, with a little ingenuity, to create nested models from non-nested models in order to test whether a particular parameter can be dropped from a model without substantially reducing the explanatory power of the model. However, in the situation of fitting different types of models to the same data set, for example, when fitting biologically based mechanistic models, empirical excess relative risk models and excess absolute risk models all to the same A-bomb data set, this is *often not* possible. The AIC and BIC information criteria (as explained below) allow many more inter-comparisons between totally different model types and provide guidance, for example, on whether biologically based models fit the data more economically, i.e., with fewer parameters, than the empirical models, or whether the ERR model fits better than the EAR model.

Table 1 Probability and evidence ratio (ER) values connected with various model-to-model changes in deviance (i.e. ΔDeviance)

ΔDeviance (P = 0.05)	Δnumber of parameters
3.84	1
5.99	2
7.81	3
9.49	4

In general, if the models are not nested and cannot be reformulated as nested models, there is a tendency, in the field of radiation epidemiology, to just quote the change in deviance without interpretation (e.g. [13, p. 390]). This approach can be improved on by the application of information criteria.

The more general problem of choosing among non-nested models, with different numbers of parameters, can be approached with an information theoretic extension of the maximum likelihood principle, as originally suggested by Akaike [6, 7] and fully described in a dedicated textbook to Akaike information criterion statistics [14] and in [1]. Another information criterion involves evaluating the leading term in the asymptotic expansion of the Bayes solution as suggested by Schwarz [8]. An informative description of both methods has recently been given [15].

Akaike’s [6, 7] suggestion amounts to maximising the likelihood function separately for each model j, obtaining the likelihood M_j and then choosing the model that minimises the Akaike information criterion (AIC),

$$\mathrm{AIC} = -2\ln(M_j) + 2k_j, \tag{1}$$

where k_j is the number of fit parameters in the model (i.e., the number of values that are estimated from the data) and the first term on the right-hand side of Eq. 1 is just the familiar deviance.

The AIC is derived by an approximate minimisation of the Kullback–Leibler information entropy, which measures the difference between the true data distribution and the model distribution. The full statistical justification is given in the original Akaike papers [6, 7] and in [1].

Adopting this formulation of AIC, the probability P for a model improvement can then be computed by the following equation [16]:

$$P = 1 - \exp(-0.5\ \Delta\mathrm{AIC})/(1 + \exp(-0.5\ \Delta\mathrm{AIC})) \tag{2}$$

where ΔAIC is the change in AIC between two competing models.

Thus, an arbitrary model A is considered to be an improvement of another model B with 95% probability, if the AIC for model A is smaller than the AIC for model B by 5.9 points, i.e. ΔAIC = –5.9 (see Table 2 for this and other examples).

When comparing two models A and B, the probability that model A fits the data better than model B can be divided by the probability that model B fits better than model A (by invoking complementary probabilities) to obtain the evidence ratio, ER as given in Table 2, where

$$\mathrm{ER} = 1/\exp(-0.5\ \Delta\mathrm{AIC}). \tag{3}$$

The other criterion for model selection, mentioned above, is a later product of early work on a Bayesian approach for comparing predictions made by two competing scientific theories [17, 18] and involves Bayes factors. If the prior probabilities of two competing models are equal, then the Bayes factor is just the posterior probability of one of these models. It is possible to avoid the introduction of the prior probabilities, and the associated numerical integrations associated with the full Bayesian method (as in [19] for example), by using a rough asymptotic approximation to the Bayes factors developed by Schwarz [8]. Then the relevant procedure for model selection involves choosing the model that minimises the Bayes Information Criterion (BIC), where the BIC is often defined to be minus twice the Schwarz criterion [8]:

$$\mathrm{BIC} = -2\ln(M_j) + k_j \ln(n), \tag{4}$$

where n is either the number of data points (for individual data) or the number of data groups or cells (for binned data).

In contrast to the AIC, the BIC involves an asymptotic approximation and does not have an information-theoretic justification—despite the name. The factor of two, just mentioned, has the function of putting the BIC on the same scale as the familiar deviance and likelihood ratio test statistic [4, 5] and so here again, the first term on the right-hand side of Eq. 4 is just the deviance. The evidence for model improvement is positive, strong or very strong, if the difference in the BIC values, between two competing models, lies in the ranges of 2–6, 6–10, and 10 and above, respectively [20] (Table 3).

Although approximate minimum t values for the different grades of evidence and sample size have been given in Table 2 of [20], the basic idea presented here is to rely on the BIC ranges for grades of Bayesian evidence for model selection among non-nested models, rather than on P or t values.

The presence of different information criteria in the literature naturally leads to the question of which one is best. Monte Carlo tests have indicated that the AIC has a tendency to favour models which have more parameters than the true model [20]. A formal proof [21] has shown the AIC to be ''dimensionally inconsistent''. This means that the probability of AIC favouring an over-parameterised model does not tend to zero even as the data set size tends to infinity. Nevertheless, the AIC has been considered here in addition to the dimensionally consistent BIC, which penalises over-parameterised models more harshly than AIC, as the data set size increases (due to the second term in its definition, Eq. 4).

Table 2 Probability and evidence ratio (ER) values connected with various model-to-model changes in AIC (i.e. ΔAIC)

ΔAIC	Probability	Evidence ratio
–1.0	0.622	1.65
–2.0	0.731	2.72
–3.0	0.818	4.48
–4.0	0.881	7.39
–5.0	0.924	12.18
–5.9	**0.950**	**19.11**
–6.0	0.953	20.09
–7.0	0.971	33.12
–8.0	0.982	54.60

The bold values are for the 95% probability of model improvement

Table 3 Probability and evidence ratio (ER) values connected with various model-to-model changes in BIC (i.e. ΔBIC)

|ΔBIC|	Evidence
0–2	Weak
2–6	Positive
6–10	Strong
>10	Very strong

Source [20]

Other statistics for model selection that are of general interest, but not applied to the examples of the next section, include: Mallows C_p [22]; the shortest length description principle [23, 24]; stochastic complexity (of a data string relative to a class of probabilistic models) [25]; the shortest data description [26]; and the deviance information criterion [27].

An example of applications of model selection: the A-bomb survivors

Data on cancer mortality

The cohort of the atomic bomb survivors from Hiroshima and Nagasaki is unique due to the large number of cohort members; the long follow-up period of more than 50 years; a composition that includes males and females, children and adults; whole-body exposures (which are more typical for radiation protection situations than the partial-body exposures associated with many medically exposed cohorts); a large dose range from natural to lethal levels; and an internal control group with negligible doses, i.e. those who survived at large distances (>3 km) from the hypocentres. The most recent data set on cancer mortality for the follow-up time periods from 1950 to 2000 with the new dosimetry system DS02 [28, 29] (data file: DS02CAN.DAT from http://www.rerf.or.jp) has been selected for the analysis here. DS02 was developed by a large international team of scientists and included the calculation of the neutron and gamma radiation transport from the point of A-bomb explosion through the atmosphere, accounting for shielding due to buildings and the human body. Validation of these calculations involved neutron activation measurements performed on environmental samples from Hiroshima (e.g. [28–33]). The mortality data are in a grouped form and are categorised by sex, city, age-at-exposure, age-attained, the calendar time period during which the health checks were made and weighted survivor colon dose. This data set provides an opportunity for conducting analyses of the data with various risk models, e.g., for radiation induced all-solid-cancer mortality, as applied in the next section.

Weighted doses

Weighted organ doses are defined by

$$d = d_\gamma + \mathrm{RBE}\ d_n, \tag{5}$$

where d_γ and d_n are organ absorbed doses from γ-rays and neutrons, respectively. For RBE, the relative biological effectiveness of neutrons, the value 10 has been used.

Only the data groups with mean weighted colon dose categories corresponding to < 2 S_v were used. The two data subsets chosen for the modelling, the associated number of cancer deaths and the number n of data cells, are given in Table 4.

Since this analysis involves all types of solid cancers grouped together, weighted organ-averaged doses [34] are used in a place of the weighted colon dose. The organ-averaged doses are calculated with weighting factors accounting for the risk contribution of individual tumour sites. The weighted organ-averaged doses are larger than the colon doses (which are used in the radiation effects research foundation analyses) by factors of 1.085 and 2 for the gamma and neutron contributions, respectively [34].

The risk models

The risk models applied here, for radiation-induced solid cancer mortality, are very similar to those already considered and explained in detail [9, 13]. In the present work, all analyses are sex-specific in order to facilitate the model-to-model comparisons here and to explore different functional forms for the age-related parameters, which may be different for males and females (an aspect to be included in a future paper). This approach deviates slightly from that in [13], where the analysis pertains to both sexes together but where the baseline model contains fit parameter values that are all sex-specific, with the only fit parameters that are really treated as common to both males and females, relating to the explanatory covariables of age-attained and age-at-exposure. Use is made of a general rate (hazard) model of the form

Table 4 Some characteristics of the data sets of atomic bomb survivors with mean weighted colon doses <2 Sv: number of cancer deaths from all types of solid cancer and number of data cells (n, required in the calculation of BIC using Eq. 4) in the grouped mortality data which covers the time from 1950 to 2000

Data set	Number of deaths	Number of data cells (n)
Male, all solid, DS02	4,779	14,803
Female, all solid, DS02	5,234	15,139

$$\lambda(d,a,e) = \lambda_0(a,e)[1 + \mathrm{ERR}(d,a,e)], \tag{6}$$

for the excess relative risk (ERR) and

$$\lambda(d,a,e) = \lambda_0(a,e) + \mathrm{EAR}(d,a,e) \tag{7}$$

for the excess absolute risk (EAR), where $\lambda_0(a, e)$ is the baseline cancer death rate, a is age-attained and e is age-at-exposure.

The ERR is factorised into a linear function of dose and a modifying function that depends either in terms of the age-attained model, ERR*(d, a)*, [35, 36] or in terms of the traditionally applied age-at-exposure model, ERR*(d, e)*, (which postulates an ERR that does not decrease in time). A more complicated mixed model which includes both age variables, ERR*(d, a, e)*, can also be considered as a third alternative. The functional form is exponential for age-at-exposure in ERR*(d, e)* or a power function for age-attained in ERR*(d, a)* and the modifying factors (see, Eq. 6) have been modelled as

$$\mathrm{ERR}(d,a,e) = k_{\mathrm{d}} d \exp(-g_e(e-30) + g_a \ln(a/70)), \tag{8}$$

where k_d is the ERR per unit dose for an age-at-exposure of 30 years and an age-attained of 70 years, and g_e, g_a are fit parameters.

The model centering at age-at-exposure of 30 years and an age-attained of 70 years was chosen to match that adopted in previous analyses, e.g. [13]. Note that here ERR*(d, e)* and ERR*(d, a) are* nested within ERR*(d, a, e)*; however, ERR*(d, e)* and ERR*(d, a)* are not nested models.

Similarly, the EAR is also factorised into a linear function of dose and a modifying function that depends either exponentially on age-at-exposure or on the natural logarithm of age-attained or on both age variables:

$$\mathrm{EAR}(d,a,e) = k_d d \exp[-g_e(e-30) + g_a \ln(a/70)] \tag{9}$$

where k_d, g_e and g_a are fit parameters. However k_d is now the EAR in units of number of excess cases per 10,000 person years per S_v, for an age-at-exposure of 30 years and an age-attained of 70 years.

The nesting properties of the EAR models are also analogous to those of the ERR models.

Although the baseline rates can be dealt with by stratification, the main calculations in the next section adopt a fully parametric model:

$$\begin{aligned}\lambda_0(a,e) = \exp\{&\beta_0 + \beta_1 \ln(a/70) + \beta_2 \ln^2(a/70)\\ &+ \beta_3 \max^2(0, \ln(a/40)) + \beta_4 \max^2(0, \ln(a/70))\\ &+ \beta_5(e-30) + \beta_6(e-30)^2\},\end{aligned} \tag{10}$$

where $\beta_0,\ldots,\beta_6$ are fit parameters.

This is a simplified version of the model of Preston et al. [13]. Some terms, including a city parameter relating to differences in baseline cancer rates between Hiroshima and Nagasaki, were dropped from the full model of Preston et al. [13] in arriving at Eq. 10. This was because an application of the likelihood ratio test for nested models [4, 5], as described above, indicated that the extra terms did not significantly improve the fit in the current analysis.

Table 5 Preferred models

Data set	Model	Number of parameters	Deviance	BIC	AIC
Male	**ERR(d, e)**	11	6,419	**6,525**	6,441
	ERR(d, a)	11	6,420	6,526	6,442
	ERR(d, a, e)	12	6,416	6,531	6,440
	EAR(d, e)	11	*6,447*	*6,553*	*6,469*
	EAR(d, a)	11	6,422	**6,528**	6,444
	EAR(d, a, e)	12	6,417	6,532	6,441
Female	**ERR(d, e)**	11	6,697	**6,803**	6,719
	ERR(d, a)	11	*6,704*	6,810	*6,726*
	ERR(d, a, e)	12	6,695	6,811	6,719
	EAR(d, e)	11	*6,742*	*6,848*	*6,764*
	EAR(d, a)	11	6,695	**6,801**	6,717
	EAR(d, a, e)	12	6,693	6,809	6,717

The bold text indicates the preferred models (i.e., minimum BIC value for each of the data sets). The models and numbers that are italicised indicate the models that are a particularly bad choice in terms of model-to-model changes in deviance or AIC or BIC

Estimation of fit parameters and statistical analysis

The maximum likelihood technique is used to fit the models, as described in [37, 38]. Best estimates uncertainty ranges and correlations of the fit parameters were determined by minimising the deviance using the MIGRAD minimisation subroutine from the CERN LIBRARY MINUIT software for optimisation. MIGRAD implements a stable version of the Davidon–Fletcher–Powell variable-metric (a quasi-Newton method) [37]. The models were also computed in EPICURE/AMFIT [38] as a double check on the numerical methods, associated convergence properties, resulting parameter values and uncertainty ranges. No inconsistencies were found.

The number of parameters in the age-at-exposure model, for example, was assumed to be equal to the number of parameters actually optimised (9 parameters) plus the two spline joins in the β_3 and β_4 parameters at 40 and 70 years, respectively, in the baseline model (Eq. 10), thus a total of 11 parameters.

The quality of model fits and associated information criterion values

Full details of the properties of interest in radiation epidemiology, i.e., ERR dose response curves with age effect-modifications and central estimates for the ERR/S_v, have already been given for these types of models [9, 13] and are not discussed here. However, for completeness, the parameter sets for four preferred models are given in Table 6, in the Appendix. Since the purpose here is to illustrate model selection techniques, the main results of relevance are given in Table 5. All inferences made in this section come from an evaluation of model-to-model changes in the quantities given in Table 5 with the aid of Tables 1, 2, 3 for interpretation. Table 5 gives the values of Deviance, BIC and AIC associated with the two classes (ERR, EAR) of models considered here. The borderlines necessary for interpreting the model-to-model changes in these values can be seen from Tables 1, 2, 3. Among these models, comparisons can be made between two nested models in the same class (where the nesting properties have been explicitly given above) using the change in deviance, and between any two models using the model-to-model changes in AIC and BIC.

The full process of model selection would normally start with adding the explanatory variables one-by-one to the model i.e., add dose, then add one age related variable and then the other age related variable. However, the full process has not been described here since the aim is one of illustrations of model selection techniques rather than of detailing the complete model selection process. There are also intrinsic difficulties involving the evaluation of time-related effect-modification factors which are caused by collinearity (i.e. correlations) in the variables [39], but these are not considered here.

Considering the ERR age-at-exposure model, it can be seen from Table 5 that when the age-attained parameter is added to the model the deviance is reduced by 3 and 2 points for the male and female data sets, respectively. Here, the likelihood ratio test would indicate that inclusion of age-attained does not lead to a significant improvement in model fit. However, if one happened to start with the ERR age-attained model and then added the age-at-exposure parameter, the deviance is reduced by 4 and 9 points for the male and female data sets, respectively, which does lead to a significant improvement in an overall model fit. This indicates the main problem in this type of model fitting—which age covariable describes the data best? Is it the age-at-exposure or the age-attained? This clearly cannot be answered with the conventional method of just looking for the change in deviance (because non-nested models are involved) and it is exactly here where the information criteria are of greatest value. It is also important to reiterate here that the inability to distinguish between two models could also arise because the data are not intrinsically powerful enough to fulfil this purpose.

There are several cases of model-to-model comparisons in Table 5 where the changes in deviance and AIC are very small (and therefore do not indicate model preferences) but where the changes in BIC indicate strong Bayesian evidence in favour of one model. For example, the comparisons between the ERR(*d*, *e*) and ERR(*d*, *a*, *e*) models for the female data set yield ΔDeviance = 2, ΔAIC = 0 and ΔBIC = 8. Given the theoretical considerations of the dimensional consistency of BIC mentioned above, this seems to be the more credible measure here and indicates strong Bayesian evidence in favour of ERR(*d*, *e*).

Comparisons between the three ERR models or between the three EAR models, for the male data generally yielded changes in AIC of 4 or less—except in the case of the EAR*(d, e)* model which stands out as a particularly poor choice. This is also true for the female data set with the additional qualification that ERR*(d, a)* is also a poor choice because AIC, in this case, is seven points more than the other two models in this class.

The preferred models in terms of BIC for both sets of data are ERR*(d, e)* and EAR*(d, a)*. The female data supports the ERR*(d, e)* (ΔBIC = 7 and 8) and EAR*(d, a)* (ΔBIC = 8 and 47) models with strong to very strong Bayesian evidence (Table 3). However, the male data support the ERR*(d, e)* and EAR*(d, a)* models with Bayesian evidence that encompasses all four categories (in Table 3) for the various model-to-model comparisons that are possible in Table 5. The Bayesian evidence does not provide support for the mixed age models, ERR*(d, a, e)*, EAR*(d, a, e)* in either data set, since the addition of a second age-related fit parameter was penalised with positive and strong evidence for the male and female data, respectively.

It is also possible to determine the relative quality of fit between the two model types ERR and EAR using AIC and BIC. Considering the changes in AIC and BIC between the preferred models in each class, i.e. ERR*(d, e)* and EAR*(d, a)*, it can be seen from Table 5 that for males, ΔAIC = 3, indicating that ERR*(d, e)* is an improvement over EAR*(d, a)* with 82% probability (according to Table 2), and ΔBIC = 3, indicating positive Bayesian evidence in favour of ERR*(d, e)* (Table 3). For females, ΔAIC = 2 indicating that EAR*(d, e)* is an improvement over ERR*(d, a)* with 73% probability (Table 2) and ΔBIC = 2, indicating weak Bayesian evidence in favour of EAR*(d, e)* for the female data set (Table 3).

Conclusion

An effort here has concentrated on explaining, applying and interpreting the outcomes of several techniques in the area of ''goodness of fit evaluations'' so that main conclusions drawn from model selection do not depend on just one type of statistical test, which could be associated with stringent assumptions (e.g. nested models). The usual comparison of deviance values and number of model parameters has been applied along with two other measures: two information criteria (AIC and BIC), not usually applied to radioepidemiology. The BIC appears to be the best method from theoretical considerations of dimensional consistency.

As examples, to illustrate the application of theses techniques, several types of radiation risk models have been fitted to the most recent mortality data for all solid cancers occurring in the Japanese A-bomb survivors. Model-to-model changes in the BIC have been seen, from these examples, to display more decisive properties in model selection than changes in AIC or changes in deviance considerations. Considering the results from all techniques together, the weight of evidence was in favour of excess relative risk models that depend on age-at-exposure and excess absolute risk models that depend on age-attained. There was positive Bayesian evidence that the excess relative risk models that depend on age-at-exposure fitted the male data better than the excess absolute risk models that depend on age-attained. However, the reverse trend was found with weak evidence for the female data. It has been demonstrated here that application of the two information criteria allows *interpretable* comparisons between non-nested models and indeed between different model types, which are not allowed by standard methods of likelihood ratio testing for nested models. This feature renders the information criteria to be particularly useful in the field of radiation epidemiology. Finally, it is probably of some importance to follow Box [40] in believing that ''all models are wrong, but some are useful''; actually, some are more useful than others.

Acknowledgments The author would like to thank Dr. W. Rühm and Dr. J. R. Walsh for critically reading the manuscript, Prof. D. Pierce and Dr. P. Jacob for useful discussions and two anonymous reviewers for many valuable comments which lead to an improvement of the original manuscript. This work makes use of the data obtained from the Radiation Effects Research Foundation (RERF) in Hiroshima, Japan. RERF is a private foundation funded equally by the Japanese Ministry of Health and Welfare and the US Department of Energy through the US National Academy of Sciences. The conclusions in this work are those of the author and do not necessarily reflect the scientific judgement of RERF or its funding agencies.

Appendix

Table 6

Table 6 Fit parameters [with standard errors (SE)] for the four preferred models in Table 5 as defined by Eq. 6–10

Parameter	Males				Females			
	ERR(d, e)		EAR(d, a)		ERR(d, e)		EAR(d, a)	
	Fitted value	SE	Fitted value	SE	Fitted value	SE	Fitted value	SE
β_0	–3.50	0.52	–3.69	0.79	–6.90	0.64	–6.98	0.75
β_1	8.49	1.75	7.81	2.71	–1.72	2.18	–1.96	2.56
β_2	1.92	1.33	1.16	2.18	–4.05	1.74	–4.28	2.07
β_3	–2.98	1.66	–2.32	2.52	5.06	2.04	5.32	2.38
β_4	–6.96	1.90	–7.09	2.01	–0.61	1.54	–0.70	1.62
β_5	3.89×10^{-4}	1.28×10^{-3}	-9.25×10^{-4}	1.23×10^{-3}	1.38×10^{-2}	1.49×10^{-3}	1.26×10^{-2}	1.43×10^{-3}
β_6	-3.11×10^{-4}	5.28×10^{-5}	-3.00×10^{-4}	5.29×10^{-5}	-4.35×10^{-4}	5.55×10^{-5}	-4.38×10^{-4}	5.58×10^{-5}
k_d	0.3199	0.07185	26.32	6.054	0.4982	0.08104	24.03	3.484
g_e	0.0397	0.0138			0.04518	0.01135		
g_a			3.215	0.6272			2.76	0.443

k_d has units of ERR/S_v (at an age-at-exposure of 30 years) for the ERR models, and number of excess cases per 10,000 person years per S_v (at an age-attained of 70 years) in the EAR models. Note, however, that the baseline cancer rate parameter values for β_0 to β_7 will result in $\lambda_0(a, e)$ with units of number of cancer deaths per year

References

1. Burnham KP, Anderson DR (2002) Model selection and multimodel inference. 2nd edn. Springer, New York
2. MacKay DJC (2003) Information theory, inference and learning algorithms. Cambridge University Press, London
3. Gregory P (2005) Bayesian logical data analysis for the physical sciences. Cambridge University Press, London
4. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference, part II. Biometrika 20A:263–294
5. Harrell FE Jr (2001) Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. Springer Series in Statistics
6. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Caski F (eds) Proceedings of the 2nd international symposium on information theory. Budapest, Hungary, Akademiai Kiado, pp 267–281
7. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
8. Schwarz G (1978) Estimating the dimension of a model. Ann stat 6:461–464
9. Walsh L, Rühm W, Kellerer AM (2004) Cancer risk estimates for γ-rays with regard to organ specific doses, part I: All solid cancers combined. Radiat Environ Biophys 43:145–151
10. Izumi S, Ohtaki M (2004) Aspects of the Armitage–Doll gamma frailty model for cancer incidence data. Environmetrics 15:209–218
11. Tavecchia G, Pradel R, Boy V, Johnson AR, Cezilly F (2001) Sex- and age-related variation in survival and cost of reproduction in greater flamingos. Ecology 82(1):165–174
12. Mukherjee S, Feigelson ED, Babu GL, Murtagh F, Fraley C, Raftery A (1998) Three types of gamma-ray bursts. Ap J 508:314–325
13. Preston DL, Shimizu Y, Pierce DA, Suyama A, Mabuchi K (2003) Studies of the mortality of atomic bomb survivors. Report 13 solid cancer and noncancer disease mortality1950–1997. Radiat Res 160:381–407
14. Sakamoto Y, Ishiguro M, Kitagawa G (1986) Akaike information criterion statistics. Kluwer Academic, Dordrecht
15. Yang Y (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika 92:937–950
16. Motulsky H, Christopoulos A (2002) Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. GraphPad Software, Inc.
17. Jeffreys H (1935) Some tests of significance, treated by the theory of probability. Proc Camb Philo Soc 31:203–222
18. Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford
19. Radivoyevitch T, Hoel DG (2000) Biologically-based risk estimation for radiation-induced chronic myeloid leukemia. Radiat Environ Biophys 39:153–159
20. Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assn 90:773–795
21. Kashyap R (1980) Inconsistency of the AIC rule for estimating the order of autoregressive models. IEEE Trans Auto Control 25:996–998
22. Mallows CL (1973) Some Comments on C_p. Technometrics 15(4):661–675
23. Kolmogorov A (1968) Three approaches to the quantitative definition of information. Probl Inf Transmission 1:1–12
24. Ramos AA (2006) The minimum description length principle and model selection in spectropolarimetry. Online under arXiv:astro-ph/0606516 v1 21 June 2006
25. Rissanen J (1986) Stochastic complexity and modeling. Ann Stat 14(3):1080–1100
26. Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471
27. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. J R Stat Soc B 64, part 4, 583–639
28. Bennett B (2003) DS02: The new dosimetry system DS02. Hiroshima Igaku (Japanese). J Hiroshima Med Assoc 56:386

29. Young R, Kerr GD (eds) (2005) DS02: Reassessment of the atomic bomb radiation dosimetry for Hiroshima and Nagasaki, Dosimetry System 2002, DS02, vols 1, 2, Radiation Effects Research Foundation, Hiroshima
30. Straume T, Rugel G, Marchetti AA, Rühm W, Korschinek G, McAninch JE, Carroll K, Egbert S, Faestermann T, Knie K, Martinelli R, Wallner A, Wallner C, Fujita S, Shizuma K, Hoshi M, Hasai H (2003) Measuring fast neutrons in Hiroshima at distances relevant to atomic-bomb survivors. Nature 424:539–541
31. Straume T, Rugel G, Marchetti AA, Rühm W, Korschinek G, McAninch JE, Carroll K, Egbert S, Faestermann T, Knie K, Martinelli R, Wallner A, Wallner C, Fujita S, Shizuma K, Hoshi M, Hasai H (2004) Measuring fast neutrons in Hiroshima at distances relevant to atomic-bomb survivors. Nature 430:483
32. Huber T, Rühm W, Hoshi M, Egbert SD, Nolte E (2003) ^{36}Cl measurements in Hiroshima granite samples as part of an international intercomparison study: results from the Munich group. Radiat Environ Biophys 42:27–32
33. Huber T, Rühm W, Kato K, Egbert S, Kubo F, Lazarev V, Nolte E (2005) The Hiroshima thermal neutron discrepancy for ^{36}Cl at large distances; Part I: New ^{36}Cl measurements in granite samples exposed to a-bomb neutrons. Radiat Environ Biophys 44:75–86
34. Kellerer AM, Walsh L (2001) Risk estimation for fast neutrons with regard to solid cancer. Radiat Res 156:708–717
35. Kellerer AM, Barclay D (1992) Age dependences in the modelling of radiation carcinogenesis: age-dependent factors in the biokinetics and dosimetry of radionuclides. Radiat Prot Dosim 41:273–281
36. Pierce DA, Mendelsohn ML (1999) A model for radiation related cancer suggested by atomic bomb survivor data. Radiat Res 152:642–654
37. James F (1994) Minuit function minimization and error analysis, Version 94.1. Technical report, CERN
38. Preston DL, Lubin JH, Pierce DA (1993) Epicure User‘s Guide. HiroSoft International Corp., Seattle
39. Lagarde F (2006) Understanding estimation of time and age effect-modification of radiation-induced cancer risk among atomic-bomb survivors. Health Phys 91(6):608–618
40. Box GEP (1976) Science and statistics. J Am Stat Assoc 71:791–799