Luque-Fernandez *et al. BMC Medical Research Methodology* (2016) 16:129
DOI 10.1186/s12874-016-0234-z

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Adjusting for overdispersion in piecewise exponential regression models to estimate excess mortality rate in population-based research

Miguel Angel Luque-Fernandez*, Aurélien Belot, Manuela Quaresma, Camille Maringe, Michel P. Coleman and Bernard Rachet

Abstract

Background: In population-based cancer research, piecewise exponential regression models are used to derive adjusted estimates of excess mortality due to cancer using the Poisson generalized linear modelling framework. However, the assumption that the conditional mean and variance of the rate parameter given the set of covariates x_i are equal is strong and may fail to account for overdispersion given the variability of the rate parameter (the variance exceeds the mean). Using an empirical example, we aimed to describe simple methods to test and correct for overdispersion.

Methods: We used a regression-based score test for overdispersion under the relative survival framework and proposed different approaches to correct for overdispersion including a quasi-likelihood, robust standard errors estimation, negative binomial regression and flexible piecewise modelling.

Results: All piecewise exponential regression models showed the presence of significant inherent overdispersion (*p*-value <0.001). However, the flexible piecewise exponential model showed the smallest overdispersion parameter (3.2 versus 21.3) for non-flexible piecewise exponential models.

Conclusion: We showed that there were no major differences between methods. However, using a flexible piecewise regression modelling, with either a quasi-likelihood or robust standard errors, was the best approach as it deals with both, overdispersion due to model misspecification and true or inherent overdispersion.

Keywords: Epidemiologic methods, Regression analysis, Survival analysis, Proportional hazard models, Cancer

Background

In population-based cancer research, the relative survival setting is used because the cause of death is often not available or considered to be unreliable [1]. Therefore, the survival and the mortality associated with cancer are estimated by incorporating the information of the expected mortality from the general population (i.e. background mortality) obtained from national or regional life tables [1, 2]. The main advantage of the relative survival setting is that it provides a measure of patients survival and mortality associated with cancer without the need for information on the specific cause of death [1]. These measures of survival and mortality are known as the net survival and the excess mortality respectively [2–4]. When multivariable adjustment is of interest, the excess mortality can be modelled using piecewise exponential regression models [3, 5]. Piecewise exponential regression excess mortality (PEREM) models derive adjusted excess mortality rates accounting for the expected mortality of the background population [5, 6].

It has been shown that PEREM models can be fitted in the Generalized Linear Modelling (GLM) framework [3]. Using the GLM framework it is relatively easy to

*Correspondence: miguel-angel.luque@lshtm.ac.uk
[1] Department of Non-Communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, Cancer Survival Group, Keppel Street, WC1E 7HT London, UK

extend the models to deal with clustering, through either a random-effects model or by utilizing sandwich-type estimators for the standard errors (SE) [6–8]. To fit PEREM models follow-up time is split into k intervals (e.g., yearly, monthly) and person-times of follow-up y_k is introduced as an offset in the model, assuming that the excess mortality rate is constant within each interval but, it can vary arbitrarily between the intervals. Moreover, the usual assumption that the number of deaths (d_k) observed in interval k can be described by a Poisson distribution with rate parameter $\lambda_k = \frac{d_k}{y_k}$ has been adapted to the relative survival setting [3].

The rate parameter λ_k is adapted to include the expected mortality of the general population under the relative survival setting

$$\lambda_k^+ = \frac{d_k - d_k^*}{y_k} = \frac{d_k^+}{y_k}, \tag{1}$$

where (d_k) and (d_k^*) are the observed and expected number of deaths from the general population and $\left(d_k^+\right)$, the excess number of deaths.

Thus, the Log-likelihood for the PEREM model includes the updated rate parameter:

$$ln\left(\lambda_k^+\right) = ln\left(\lambda_{0k}^+\right) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{2}$$

where $ln(\lambda_k^+)$ is the logarithm of the excess mortality and $\mathbf{x}^{\mathrm{T}}$ denotes the transpose of the vector of covariates $\mathbf{x}$ and $\boldsymbol{\beta}$ represent the corresponding parameter estimates.

Using (1), we can rewrite the rate parameter defined in (2) as:

$$ln\left(d_k - d_k^*\right) = ln(y_k) + ln\left(\lambda_{0k}^+\right) + \mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}, \tag{3}$$

where $ln(y_k)$ is the logarithm of the person-time at risk for the k^{th} interval incorporated in the model as an offset and $ln(\lambda_{0k}^+)$, is the log of the baseline excess mortality rate [3, 6].

Using (1), we can rewrite the PEREM model in (3):

$$\frac{d_k}{y_k} = \frac{d_k^*}{y_k} + \lambda_{0k}^+ \exp\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right).$$

The model in (3) assumes constant rates over intervals of time and it may lead to overdispersion due to extra variability in the rate parameter (i.e. the variance is greater than the mean). The assumption that the conditional mean and variance of the rate parameter given covariates $\mathbf{x}$ are equal is strong and may fail to account for inherent or genuine overdispersion. The variance exceeds the mean generally because of positive correlation between variables or excess variation between rates [9].

Overdispersion in PEREM models is typically due to extra variability in the rate parameter (genuine overdispersion). However, other forms of non-genuine overdispersion may appear when the model omits important explanatory predictors; the data contain outliers, or the model fails to introduce enough interaction terms or non-linear functional form between predictors and outcome. By contrast, no external remedy can be applied in the case of inherent or genuine overdispersion [10].

Fitting an overdispersed PEREM model leads to underestimating standard errors (SE) and therefore to the inappropriate interpretation of the conditional estimates of the covariates introduced in the model (i.e. a variable may be wrongly considered as a significant predictor).

Using an empirical example, we aim to take advantage of the relationship between the GLM framework and the PEREM model to apply a simple method to test and correct for overdispersion that could be easily implemented and used by population-based cancer researchers.

Methods

The presence of overdispersion can be recognized when the value of the ratio between the Pearson χ^2 (or deviance statistics) over the degrees of freedom is larger than one. However, a more formal statistical approach is required to test the presence of inherent overdispersion, then to correct for it [11].

Testing overdispersion in PEREM models

A regression-based score test enables us to evaluate whether the variance is equal to the mean $(Var(\lambda^+) = E(\lambda^+))$ or proportional to the square mean [11]:

$$Var\left(\lambda^+\right) = E\left(\lambda^+\right) + \alpha E\left(\lambda^+\right)^2, \tag{4}$$

We first calculate the score statistic (Z) to test H0: $\alpha = 0$ against H1: $\alpha > 0$, using the fitted values of the excess mortality rate $\widehat{\lambda^+}$ [11–13]:

$$Z = \sum_{i=1}^{N}\sum_{k=1}^{M}\left(\frac{\left(\lambda_{ik}^+ - \widehat{\lambda_{ik}^+}\right)^2 - \lambda_{ik}^+}{\widehat{\lambda_{ik}^+}}\right),$$

where $\lambda_{ik}^+ = \frac{d_{ik}^+}{y_{ik}}$ and substituting λ_{ik} and $\widehat{\lambda_{ik}^+}$ gives:

$$Z = \sum_{i=1}^{N}\sum_{k=1}^{M}\left(\frac{\left(\frac{d_{ik}^+}{y_{ik}} - \frac{\widehat{d_{ik}^+}}{y_{ik}}\right)^2 - \frac{d_{ik}^+}{y_{ik}}}{\frac{\widehat{d_{ik}^+}}{y_{ik}}}\right). \tag{5}$$

The test is implemented by a linear regression of the generated dependent variable Z on $\widehat{\lambda_{ik}^+}$ (independent variable), without including an intercept term. Hence the output can be interpreted as a T-test of whether the coefficient of $\widehat{\lambda_{ik}^+}$ is zero testing whether the variance of the rate parameter is equal to the mean [12].

Correcting for overdispersion

The most commonly used approaches to correct for inherent overdispersion are relatively straightforward to implement in common statistical software.

Quasi-likelihood approach

Inherent overdispersion in PEREM modeling may be due to extra variability in the parameter $\lambda^+_{ik} = \frac{d_{ik}-d^*_{ik}}{y_{ik}}$. Including an extra parameter ϕ in the model allows the variance to vary freely from the mean [14]. There are several options to compute the extra parameter ϕ. The simplest is to take $f(\lambda^+, \phi) = \phi \times \lambda^+$, which specifies a constant proportional overdispersion ϕ across all individuals. Using a PEREM modeling approach, we assume that the distribution of λ^+_{ik} is Poisson. Hence the Pearson Chi-squared statistic can be computed as a criteria of goodness of fit using the observed (O) and expected values (E) from the model:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

Substituting O and E by λ^+_{ik} and $\widehat{\lambda^+_{ik}}$ gives:

$$\chi^2 = \sum_{i=1}^{N}\sum_{k=1}^{M} \left(\frac{\left(\frac{d^+_{ik}}{y_{ik}} - \frac{\widehat{d^+_{ik}}}{y_{ik}} \right)^2}{\frac{\widehat{d^+_{ik}}}{y_{ik}}} \right).$$

The ratio between the Pearson χ^2 or deviance statistic, and the degrees of freedom should be closed to one as we expect that the variance of the model is equal to that of the assumed Poisson distribution. We can estimate the overdispersion parameter ϕ multiplying the inverse of the degrees of freedom (df) of the model times the Pearson χ^2 statistic.

Scaling the SE with $\sqrt{\hat{\phi}} = \sqrt{\chi^2/df}$ will correct the estimated SE of $\hat{\beta}$, which was estimated under the model of constant overdispersion [15, 16]. The estimated $\hat{\phi}$ is integrated as a scalar updating the variance-covariance matrix of the PEREM model estimated under the GLM framework and thus correcting for overdispersion [17]. Under the GLM framework $\hat{\beta}$ and the SE of $\hat{\beta}$ is optimized via an iteratively reweighted least squares procedure [11, 17]. Therefore, scaling the SE of $\widehat{\beta}$ in terms of matrix notation is given by [17]

$$\text{Variance}(\hat{\beta}) = \hat{\phi} \left(\mathbf{X}^T \mathbf{W} \mathbf{X} \right)^{-1},$$

where $\mathbf{X}$ represents the n×p design matrix of the observed data and, $\mathbf{W}$ is a diagonal n×n matrix with the values of $\widehat{\lambda^+} = \exp(\mathbf{x}^{\mathrm{T}} \boldsymbol{\beta})$ on the diagonal. Thus, the variance is updated with the new values of the weighted matrix under the assumption of no specific probability distribution [14].

Robust standard errors of parameters estimates

In maximum likelihood estimation, the standard errors of the estimated parameters are derived from the Hessian (matrix of second derivatives on the parameters) of the likelihood. However, theses standard errors are correct only if the likelihood is the true likelihood of the data [14]. In cases where we consider that overdispersion might be due to unobserved covariates and the link function or the probability distribution function are misspecified, the assumption about the true likelihood of the data does not hold. Under these scenarios, we can still use robust estimates of the standard error known as Huber, White, or sandwich variance estimates to correct for overdispersion Additional file 1 [18–20].

$$\text{Variance}(\hat{\beta}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} (\mathbf{X}^T \Sigma \mathbf{X}) (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1},$$

where Σ is a n×n matrix with the values of $(\lambda^+ - \widehat{\lambda^+})^2$ on the diagonal.

Negative binomial regression model

Given the presence of heterogeneity of subject-specific rates leads naturally to the question of whether we can model subject-specific rates using a random effects framework. The simplest random effects model assumes a person-to-person heterogeneity can be expressed by a model for the mean along with a log-Gamma distribution of the random intercept term. The random intercept follows a log-Gamma distribution, and the marginal distribution of the outcome followed a negative binomial distribution which has two parameters, shape ($E(\lambda^+)$) and scale ($Var(\lambda^+)$), but importantly its variance and mean are related by (4), where the parameter α must be positive allowing the variance of λ^+ to be greater than the mean [11]. The Poisson distribution is a special case of the negative binomial distribution where $\alpha = 0$ [11]. We can estimate α using the coefficient from a linear regression with Z (5) as dependent variable on $\widehat{\lambda^+}$ (independent variable), without including an intercept term, as described above [12].

Flexible PEREM model

The piecewise exponential regression model under the GLM and relative survival frameworks could be extended by finely splitting the time scale and using a flexible function of time such as splines [21, 22]. The flexible PEREM models allow modelling the baseline hazard and any time-dependent effects as smooth and continuous functions [6]. A time-dependent effect is easily modelled by including an interaction term between the smooth function of time and the covariate [23]. Cubic regression splines is a very popular choice for modelling flexible functions. In a truncated power basis, a cubic regression spline s(t) of time t, with K knots located in different places of the distribution of the smooth function of time can be written as [5]:

$$s(t) = \sum_{j=0}^{3} \beta_{0j} t^j + \sum_{i=1}^{K} \beta_{i3} (t - k_i)_+^3,$$

where

$$(\mathrm{t} - k_i)_+^3 = \begin{cases} (\mathrm{t} - k_i)^3 & \text{if t} > k_i, \\ 0 & \text{otherwise.} \end{cases}$$

In order to deal with high variance at the outer range of the predictors, they may be forced (restricted) to be linear before the first knot and after the last knot leading to a natural or restricted cubic spline [23]. The first and the last knots are known as boundary knots [24, 25]. If we define m interior knots, $k_1, \ldots, k_m$, and also two boundary knots, $k_{min}, \ldots, k_{max}$, we can now write $s(t)$ as a function of parameters γ and some newly created variables $z_1, \ldots, z_{m+1}$, giving [5]:

$$s(t) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \ldots + \gamma_{m+1} z_{m+1},$$

The basic functions $z_j (j = 2, \ldots, m + 1)$ are derived as follows:

$$z_1 = t,$$

$$z_j = (x - k_j)_+^3 - \lambda_j (x - k_{min})_+^3 - (1 - \lambda_j)(x - k_{max})_+^3,$$

$$\lambda_j = \frac{k_{max} - k_j}{k_{max} - k_{min}}.$$

These functions can be easily implemented using various Stata commands (e.g., rcsgen) [5, 6]. The flexible PEREM approach using splines allows modelling easily non-proportional excess mortality rate ratios including time-dependent effects of the covariates. Thus, we can achieve a better model specification which should minimize non-genuine overdispersion [25]. However, we can still scale the SE estimates in case of inherent overdispersion previously detected using the suggested regression based Score test.

Illustration

Data were obtained from the Office for National Statistics (ONS), comprising 376,791 women diagnosed with breast cancer in England between 1997 and 2005, with a follow-up to the end of 2012. The event of interest is death from any cause, with follow-up restricted to 7 years after diagnosis though we estimated up to 5 years excess mortality [21]. We built life tables from England to derive the expected mortality in the background population, by sex, single year of age, calendar year, and deprivation quintile. We aimed to estimate excess mortality hazard rate for age and deprivation in the first five years after the diagnosis of a breast cancer. Legal authority to hold the cancer data derives from a contract with the ONS to produce the official national statistics on cancer survival.

Statistical methods

First, we split the times-to-event to merge the cancer data with the estimated expected number of deaths for all causes using life tables from England [26]. Then, we fitted four types of PEREM models: in model A, we did not correct for overdispersion, in model B we scaled the SEs by the $\sqrt{\widehat{\phi}}$, in model C we used the Sandwich estimates of the SEs and in model D we fitted a NBR assuming a log-gamma distribution. All models were within the GLM framework with the Poisson family and the modified link $\left(ln\left(d_{ik} - d_{ik}^*\right)\right)$. The modified link log was used to incorporate in the maximum likelihood estimation the expected number of deaths (d^*) from the background population [3, 5].

Deprivation was included in all PEREM models as a categorical variable, with Q1, the least deprived group, as the reference category. Age was included as a categorical variable with five levels (<50, 50-59, 60-69, 70-79, ≥80) using <50 as reference. Follow-up time was parameterized as a categorical variable in PEREM models and as a smooth function of time for the flexible PEREM models. We reported $\widehat{\beta}$, var($\hat{\beta}$), and the relative loss in efficiency (RLE) of var($\widehat{\beta}$). To estimate RLE for each PEREM model corrected for overdispersion, the model not corrected for overdispersion was the reference [27]. The RLE was computed as the ratio between the variance of the estimates from the models adjusted for overdispersion and the variance from the uncorrected model

$$\mathrm{RLE}(\mathrm{var}(\hat{\beta}_2), \mathrm{var}(\hat{\beta}_1)) = \frac{\mathrm{var}(\hat{\beta}_2)}{\mathrm{var}(\hat{\beta}_1)}, \qquad (6)$$

where var($\hat{\beta}_2$) refers to the corrected estimate of the variance for overdispersion (scaling the SE or using the sandwich robust estimates) and var($\hat{\beta}_1$) to the uncorrected. The RLE was interpreted as the percentage of efficiency loss (% of increase in the variance estimate) for PEREM models needed to reduce bias after correction for overdispersion.

Finally, we fitted a flexible PEREM model which included an interaction between deprivation quintiles and follow-up time to allow the effect of deprivation to vary over time. Hence, the baseline rate was defined as a restricted cubic spline, with one-month intervals and five knots placed at the minimum and the maximum and at the 25th, 50th, and 75th centiles of the event times. For this flexible PEREM model, we plotted the excess mortality rate ratios and 95 % CI for each quintile of deprivation in categories with corrected and uncorrected SE for overdispersion [6, 23, 28]. All analysis were performed using Stata v.14 (StataCorp LP, College Station, Texas, US) Additional file 2.

Results

The Pearson χ^2 deviance residuals were non-normally distributed for the uncorrected PEREM model (Shapiro-Wilk test for normality p-value = 0.01) [29], and the overdispersion parameter (ϕ) was 21.3 % times higher than expected suggesting the presence of overdispersion. The Score test for overdispersion rejected the H_0 (p-value $<$ 0.001) indicating the presence of truly overdispersion in the rate parameter and, the scatter plot of the standardized Pearson's χ^2 residuals against the excess mortality rates suggested the presence of heteroscedasticity and hence, potential overdispersion (Fig. 1).

Table 1 contrasts the exponentiated coefficients, SE, and RLE for the four different PEREM models, uncorrected (model A) or corrected for the presence of inherent overdispersion (models B, C with ϕ parameter = 21.3) and model D adjusted for overdispersion using the NBR approach.

The model with the less conservative SE, model A, showed significant excess mortality rate ratio for each of the four deprivation quintiles (compared with the first quintile). Models corrected for overdispersion (B, C, and D) provided more conservative estimates of the SEs. After accounting for oversdispersion, deprivation showed a significant excess mortality, compared to Q1, only for the deprivation quintiles Q3-Q5 for models B and D, and for the deprivation quintiles Q4 and Q5 for model C (Table 1). Compared with the unadjusted model A, all corrected models showed a non-significant effect for age groups 60-69 and 50-59. Overall, the RLE ranged between 12 and 46 percent for corrected models compared with the model uncorrected for overdispersion. The RLE was, however, larger for model C (robust SE). The model D (NBR), compared with model B (scaled SE) and C, showed the smallest RLE. The loss of precision in the models corrected for overdispersion was reflected by the loss of statistical significance for the age groups 50-59, 60-69 and the deprivation quintiles Q1 and Q2. However, scaling the SE to control for overdispersion (model B) showed better efficiency (smaller RLE) compared with the robust SE estimation (model C)(Table 1).

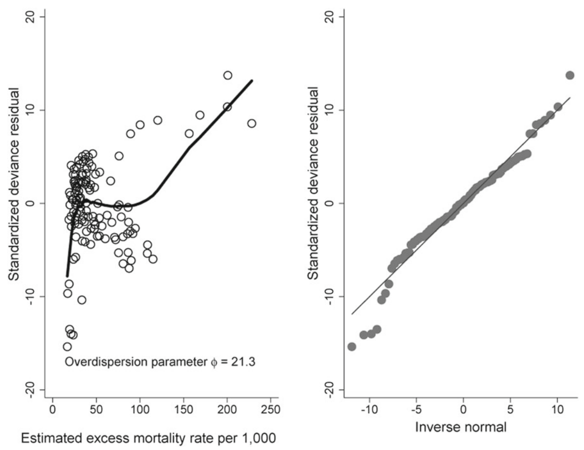

Fig. 1 Piecewise exponential regression excess mortality model: standardized Pearson χ^2 residual analysis, n= 376,791 women diagnosed with breast cancer in England between 1997 and the end of 2005

Finally, the flexible PEREM model showed smaller overdispersion parameter (ϕ = 3.2). The test for overdispersion showed the presence of significant inherent overdispersion (p-value <0.001). The flexible PEREM model reduced significantly the overdispersion parameter compared with the models without the smooth functions of time (21.3 vs. 3.2). Allowing for the time dependent effect of deprivation, revealed a decreasing trend of the excess mortality over time during the first five years after the diagnosis of breast cancer. Furthermore, the interaction between the smooth function of time with deprivation showed a stronger effect of deprivation over time, illustrated with 8 to 4 times higher excess mortality rate ratios for the most deprived group compared with the least deprived (Fig. 2).

Discussion

We have shown that under the relative survival and GLM frameworks, the modified link to fit PEREM models, allows the inclusion in the maximum likelihood estimation of the information regarding the background mortality of the reference population [3]. However, data analysts may expect to find inherent overdispersion as a characteristic of this modelling approach [30].

We have shown, that inappropriate imposition of the Poisson restriction may produce spuriously small SEs of the estimated coefficients $\hat{\beta}$. Fitting an overdisperse PEREM model under the relative survival and GLM frameworks, may lead to underestimate SEs and, therefore, to inappropriate statistical interpretation of the significance of the conditional estimates from the effects of the covariates introduced in the model (i.e., a variable or the levels of a categorical variable, may appear to be significant predictors of the outcome, when in fact it is not).

We encourage epidemiologist and applied statistician using PEREM models under the relative survival framework, to consider to test the Poisson restriction and to relax it, if appropriate, using the methodological approaches described in this article. However, in addition to cancer, the same advice may apply to any other chronic disease or condition for which estimates of disease-specific population-based survival time controlling for competing risk are of interest. We have shown, that using

Table 1 Piecewise exponential regression excess mortality models with and without correcting for overdispesion, $n = 376{,}791$ women diagnosed with breast cancer in England between 1997 and the end of 2005

	PEREM A		PEREM B (scaled SE)			PEREM C (Robust SE)			PEREM D (NBR)		
Age at diagnosis	EMRR	SE	EMRR	SE	RLE (%)	EMRR	SE	RLE (%)	EMRR	SE	RLE (%)
$50 - 59$*vs.* < 50	0.75	0.0107	0.75*	0.0493	21.3579	0.75*	0.0576	29.2222	0.75*	0.0380	12.6944
60-69 vs. <50	0.88	0.0130	0.88*	0.0600	21.3580	0.88*	0.0599	21.2823	0.87*	0.0486	14.0296
70-79 vs. <50	1.71	0.0235	1.71	0.1086	21.3578	1.71	0.1324	31.7953	1.65	0.1005	18.3181
≥80 vs. <50	3.39	0.0465	3.39	0.2150	21.3579	3.39	0.3159	46.1198	3.15	0.2222	22.8188
Quintiles of deprivation											
Q2 vs. Q1	1.05	0.0153	1.05*	0.0705	21.3659	1.05*	0.0747	24.0197	1.05*	0.0626	16.8745
Q3 vs. Q1	1.16	0.0166	1.16	0.0767	21.3711	1.16*	0.0873	27.6612	1.15	0.0687	17.1404
Q4 vs. Q1	1.27	0.0182	1.27	0.0839	21.2723	1.27	0.0934	26.3313	1.27	0.0762	17.5240
Q5 vs. Q1	1.48	0.0218	1.48	0.1007	21.3249	1.48	0.1046	23.0039	1.47	0.0885	16.4928

EMRR Excess mortality rate ratio, *NBR* Negative binomial regression, *PEREM* Piecewise exponential regression excess mortality model, *RLE* Relative loss in efficiency, *SE* Standard error, *p-value >0.05

a simple test for overdispersion we can identify significant inherent overdispersion, and applying a pseudo-likelihood estimation, fitting an NBR or a more advanced flexible PEREM modeling approach we can correct for it. These simple approaches may allow applied researchers in population-based cancer registries to infer correct conclusions from the analysis of their data in the presence of significant overdispersion. Applied researchers will have to consider the trade-off between modelling complexity and model interpretation as it might happen that

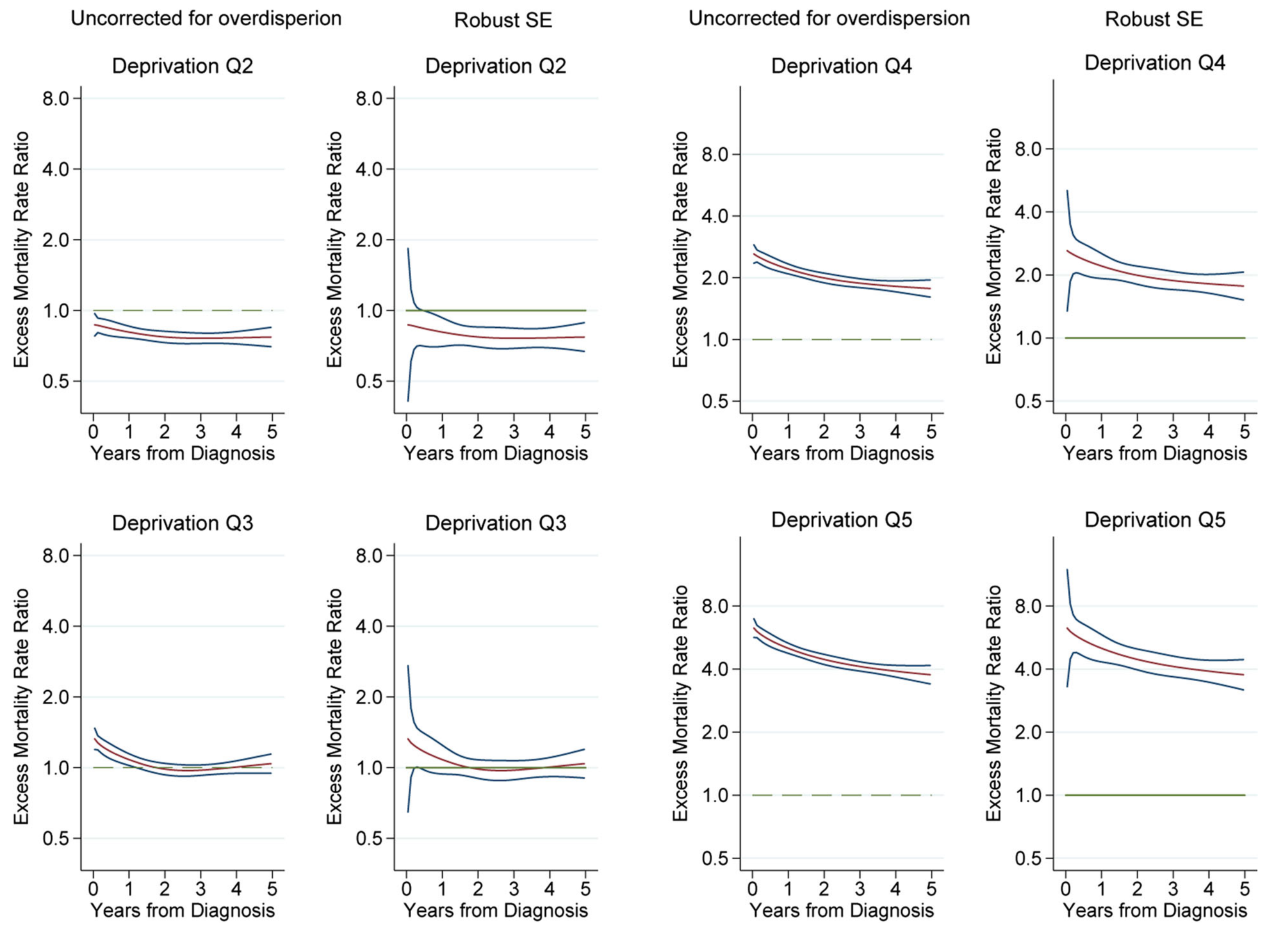

Fig. 2 Flexible piecewise exponential regression model: A (non-scaled SE) B (robust SE), $n = 376{,}791$ women diagnosed with breast cancer in England between 1997 and the end of 2005

there is no reason for applying a more advanced flexible PEREM modelling given a non-significant overdispersion test. However, it rarely happens, as under the relative survival framework we may expect the presence of overdispersion due to the variability of the rate parameter. Furthermore, in case of a significant overdispersion test, applied researchers will have to consider the compromise of model efficiency (i.e., the precision of the SE) while deciding which method or approach to use to deal with overdispersion. As suggested in our results, the flexible PEREM model showed the smaller loss in precision.

We suggest scaling the SE to correct for overdispersion due to the variability of the rate parameter with a significant overdispersion test and a small overdispersed parameter ϕ. However, it should be noticed that our results regarding the RLE are based on only one empirical data. Hence, further investigation is warranted using simulations.

The maximum likelihood methods are based on strong distributional assumptions, while quasi-likelihood or maximum likelihood methods with robust SEs rely on weaker assumptions. Furthermore, using a flexible parametric approach including time-dependent effects allows for a better model specification decreasing overdispersion. We suggest testing for the presence of inherent overdispersion in the data and correct for it using any of the approaches presented in this article. Given that there are no major differences between the above-described methods, the question is not which method to use (robust SE, scaled SE or NBR) in the presence of inherent overdispersed data, but to use any of them to correct for overdispersion and to infer correct conclusions from the models. However, we have shown the benefits of using the flexible PEREM modelling approach with either scaled, robust SE or NBR, under the GLM and relative survival frameworks. Flexible PEREM modelling benefits are double as it deals with model misspecification and overdispersion. The introduction of smooth functions of time and time-dependent effects in the flexible PEREM models may improve the model specification reducing significantly the overdispersion parameter.

Conclusion

In population-based cancer research, PEREM models are used to estimate the excess mortality rate from cancer under the relative survival framework. We have shown the impact of overdispersion on the excess mortality rate estimates by deprivation among women diagnosed with breast cancer in England between 1997 and the end of 2005. PEREM models are fitted under the assumption of a Poisson distribution leading to overdispersion. We have shown that inappropriate imposition of the Poisson restriction may produce spuriously small estimated standard errors, and thus, wrong interpretation of the model estimates. Given the public health relevance of population-based data analyses for policy and decision making, it is desirable to test for overdispersion and to correct it if appropriate.

Additional files

Additional file 1: Robust standard error estimation for generalized linear models. (2.84 KB)

Additional file 2: Stata do file with commented syntax. (8.81 KB)

Abbreviations

E: Expected value; GLM: Generalized linear model; NBR: Negative binomial regression; O: Observed value; ONS: Office of national statistics; PEREM: Piecewise exponential regression excess mortality model; RLE: Relative loos of efficiency; SE: Standard error

Acknowledgements

The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. We appreciate Dr. Francisco Rubio's comments and thorough review of the algebraic expressions.

Funding

This work was supported by Cancer Research UK grant number C7923/A18525. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of Cancer Research UK.

Availability of data and materials

Stata code is provided as a supplement of the article.

Authors' contributions

MALF developed the concept and design of the study, analyzed the data and, wrote the manuscript. All authors interpreted the data, drafted and revised the manuscript critically. All authors read and approved the final version of the manuscript. MALF is the guarantor of the paper.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Approval to analyse the data including the consent to participate was obtained from the ONS Medical Research Service (MR1101, Nov 20, 2007) and from the statutory Patient Information Advisory Group (PIAG; now the Ethics and Confidentiality Committee of the National Information Governance Board) under Section 61 of the Health and Social Care Act 2001 (PIAG 1-05(c)/2007,July 31, 2007).

Received: 6 July 2016 Accepted: 20 September 2016
Published online: 01 October 2016

References

1. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: elements for further discussion. Stat Med. 1990;9(5):529–38.
2. Perme MP, Stare J, Estève J. On estimation in relative survival. Biometrics. 2012;68(1):113–20. doi:10.1111/j.1541-0420.2011.01640.x.
3. Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. Stat Med. 2004;23(1):51–64. doi:10.1002/sim.1597.
4. Mariotto AB, Noone AM, Howlader N, Cho H, Keel GE, Garshell J, Woloshin S, Schwartz LM. Cancer survival: an overview of measures, uses, and interpretation. J Natl Cancer Inst Monographs. 2014;2014(49):145–86.
5. Dickman PW, Coviell E. Estimating and modelling relative survival. Stata J. 2015;15(1):186–215.

6. Royston P, Lambert PC, et al. Flexible parametric survival analysis using stata: beyond the cox model. College Station, Texas: Stata Press Books; 2011, p. 347.
7. Hardin JW. The sandwich estimate of variance. Adv Econ. 2003;17:45–74.
8. Dupont C, Bossard N, Remontet L, Belot A. Description of an approach based on maximum likelihood to adjust an excess hazard model with a random effect. Cancer Epidemiol. 2013;37(4):449–56. doi:10.1016/j.canep.2013.04.001.
9. Hardin JW, et al. The robust variance estimator for two-stage models. Stata J. 2002;2(3):253–66.
10. Rao CR, Miller JP, Rao DC, Vol. 27. Handbook of statistics: epidemiology and medical statistics. Amsterdam: North Holland; 2007, p. 870.
11. Hardin JW, Hilbe JM, Hilbe J. Generalized linear models and extensions. College Station, Texas: Stata Press Books; 2007, p. 387.
12. Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the poisson model. J Econ. 1990;46(3):347–64.
13. Cameron AC, Trivedi PK. Econometric models based on count data. comparisons and applications of some estimators and tests. J Appl Econ. 1986;1(1):29–53.
14. Rabe-Hesketh S, Everitt B. Handbook of statistical analyses using stata, Fourth edition. USA: Chapman and Hall/CRC; 2007, p. 342.
15. Aitkin M. A general maximum likelihood analysis of overdispersion in generalized linear models. Stat Comput. 1996;6(3):251–62.
16. Guisan A, Edwards TC, Hastie T. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol Model. 2002;157(2):89–100.
17. Faraway JJ. Extending the linear model with r: generalized linear, mixed effects and nonparametric regression models, Second edition. USA: Chapman and Hall/CRC; 2016, p. 399.
18. Rabe-Hesketh S, Skrondal A. Multilevel and longitudinal modeling using stata. College Station, Texas: Stata Press Books; 2008, p. 562.
19. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley: University of California Press; 1967, pp. 221–33. http://projecteuclid.org/euclid.bsmsp/1200512988.
20. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: J Econ Soc. 1980;48(4): 817–38.
21. Remontet L, Bossard N, Belot A, Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. Stat Med. 2007;26(10): 2214–28.
22. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med. 1989;8(5):551–61.
23. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata J. 2009;9(2):265.
24. De Boor C. A practical guide to splines. Math Comput. 1978;27:348.
25. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013. p. 426.
26. Coleman MP, Rachet B, Woods LM, Mitry E, Riga M, Cooper N, Quinn MJ, Brenner H, Estève J. Trends and socioeconomic inequalities in cancer survival in england and wales up to 2001. Br J Cancer. 2004;90(7):1367–73. doi:10.1038/sj.bjc.6601696.
27. Rao CR. Efficient estimates and optimum inference procedures in large samples. J R Stat Soc Ser B Methodol. 1962;24(1):46–72.
28. Royston P, Sauerbrei W. Multivariable modeling with cubic regression splines: a principled approach. Stata J. 2007;7(1):45.
29. Royston P. A simple method for evaluating the shapiro-francia w' test for non-normality. Statistician. 1983;32:297–300.
30. McCullagh P, Nelder JA, Vol. 37. Generalized linear models, Second edition. USA: Chapman and Hall/CRC. Monographs on Statistics and Applied Probability; 1989, p. 532.

Additional file One

Robust standard error estimation for generalized linear models

Let $\theta \in R^p$ be a p x 1 parameter vector and f_i be a positive density function $y_i \sim, f_i(y_i|\theta)$. The data (y_i) are modeled as observed values of Y_i for i = 1,...,n and the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f_i(y_i|\theta),$$

and the log-likelihood function by

$$LL(\theta) = \sum_{i=1}^{n} ln\, f_i(y_i|\theta).$$

First derivative (gradient) of the log-likelihood function

$$L'(\theta) = \sum_{i=1}^{n} g_i(y_i|\theta) = \sum_{i=1}^{n} \frac{\partial ln f_i(y_i|\theta)}{\partial\theta}.$$

Second derivative (hessian) of the log-likelihood function

$$L''(\theta) = \sum_{i=1}^{n} h_i(y_i|\theta) = \sum_{i=1}^{n} \frac{\partial^2 ln f_i(y_i|\theta)}{\partial\theta^2}.$$

Assuming that the model is correct, there is a true value θ_0 for θ. Then, we can use the Taylor approximation of second order for the log-likelihood function to estimate θ.

$$L(\theta) = L(\theta_0) + L'(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T L''(\theta_0)(\theta - \theta_0) + \ldots$$

Therefore to derive the variance covariance matrix from the maximum likelihood estimation we can differentiate this expression to get

$$L'(\theta_0) + (\theta - \theta_0)^T L''(\theta_0) = 0.$$

$$\text{So, } \widehat{\theta} - \theta_0 = [-L''(\theta_0)]^{-1} L'(\theta_0)^T,$$

$$Cov(\widehat{\theta}) = [-L''(\theta_0)]^{-1} [Cov(L'(\theta_0)][-L''(\theta_0)]^{-1}$$

Additional file two

```
************************************************************************
*Controlling for Overdispersion in Piecewise Exponential Excess Mortality
*Regression Models in Population-based Cancer Research
************************************************************************
**********************************************************************************************************
*A) Piecewise Exponential Regression Excess Mortality Models with a without accounting for overdispersion
**********************************************************************************************************
use England_breast_1997_2007.dta,clear
stset time, fail(dead=1) exit(time 10) id(ID2)
//Using strs to merge data with background expected number of deaths and mortality rate from England life tables
    strs using popmort_England_2, ///
breaks(0(1)5) noshow ///
diagage(agediag) diagyear(yeardiag) ///
attage(age) attyear(year) ///
mergeby(ctry dep region sex year  age) ///
by(agegroup dep) ///
survprob(survprob) ///
notables savgroup(group.dta, replace)
//Proportional excess hazards model: piecewise exponential regression excess mortality model
use group.dta, clear
quietly tabulate agegroup, generate(agep)
quietly tabulate dep, gen(depr)
glm d ibn.end agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nocons nolog baselevels
*Pearson statistic
predict double pred, mu
gen  per = (d-pred)^2/(pred)
egen pearson=sum(per)
sum pearson
// Assessing overdispersion: Analysis of Pearson and Deviance residual & Figure 1
*Expected excess mortality
sum nu
gen double erate= ((d-d_star)/y)
corr nu erate
replace erate=erate*1000
tw line erate midtime if agep5==1 & depr5==1 || line erate midtime if agep5==1 & depr1==1
*Standardized Deviance
predict dev, deviance
*Deviance residuals
predict stdev, standardized deviance
*Goodness of fit and checking model assumptions
lowess stdev erate, scheme(s1manual) xtitle("Expected excess mortality rate per 1,000") plotregion(style(none)), name(Figure1A)
qnorm stdev, scheme(s1manual) xtitle("Inverse normal") plotregion(style(none)), name(Figure1B)
swilk dev
graph combine Figure1A.gph Figure1B.gph, name(Figure1)
//Testing for overdispersion: Score based t-test
glm d ibn.end agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nocons nolog baselevels
predict double mu, mu
//Score based t-test: H0 E(x|y) = Var(x|y)
gen double zrs =((d-mu)^2-d)/(mu)
reg zrs mu, nocons nohead
//Negative binomial model Var(x|y) = E(x|y) + E(x|y)^2 x alpha. Alpha can be approximate using the score based t-test
reg zrs mu, nocons nohead //Beta is an approximation of alpha = .0237247
//Table 1
//Model Unadjusted for overdispersion
glm d ibn.end agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nocons nolog baselevels
eststo table1
//Model adjusted for overdispersion scaling SE
glm d ibn.end agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nocons nolog baselevels scale(x2)
eststo table2
//Model adjusted for overdispersion robust SE
glm d ibn.end agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nocons nolog baselevels vce(robust)
eststo table3
//Model adjusted for overdispersion using NBR (modelling mean and variance)
glm d ibn.end agep2-agep5 depr2-depr5, family(nbin .0237247) link(rs d_star) lnoffset(y) eform nocons nolog baselevels
eststo table4
estimates table table1 table2 table3 table4, stat(aic deviance df dispers_ps) b(%7.3f) se(%6.3f) p(%4.3f) eform
return list
esttab using table.tex, label ci nostar nodepvars brackets scalars(aic deviance df dispers_ps) ///
booktabs nomtitles ///
title("Piecewise exponential regression excess mortality models goodness of fit and point estimates (95$\%$CI)") eform replace
*******************************************************************************************************************************
*B) Flexible Poisson piecewise relative survival model with restricted cubic splines and time dependent effect of deprivation
*******************************************************************************************************************************
use England_breast_1997_2007.dta,clear
stset time, fail(dead=1) exit(time 10) id(ID2)
strs using popmort_England_2, ///
breaks(0(‘=1/12’)5) noshow ///
diagage(agediag) diagyear(yeardiag) ///
attage(age) attyear(year) ///
```

```
mergeby(ctry dep region sex year age) ///
by(agegroup dep) notables ///
survprob(survprob) ///
savgroup(group2.dta, replace)
//FLEXIBLE parametric piecewise exponential regression excess mortality model
use group2.dta, clear
gen midtime=(start + end)/2
gen lnmidtime=ln(midtime)
rcsgen lnmidtime, gen(rcs) df(3) fw(d) orthog
rcsgen lnmidtime, gen(rcstvc) df(3) fw(d) orthog
quietly tab agegroup, gen(agep)
quietly tabulate dep, gen(depr)
forvalues i= 2/5{
forvalues j= 1/3{
gen dep‘i’rcs‘j’ = depr‘i’*rcstvc‘j’
}
}
//Testing overdispersion
glm d rcs1-rcs3 agep2-agep5 dep?rcs? depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nolog baselevels
predict double mu, mu
//Score based t-test: H0 E(x|y) = Var(x|y)
gen double zrs =((d-mu)^2-d)/(mu)
reg zrs mu, nocons nohead
//Figure 2
//Flexible PEREM model allowing for the time dependent effect of deprivation without correcting for overdispersion
glm d rcs1-rcs3 dep?rcs? agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nolog baselevels
forvalues i = 2/5{
predictnl lhr‘i’ = _b[depr‘i’] + _b[agep‘i’] + _b[dep‘i’rcs1]*dep‘i’rcs1 + ///
    _b[dep‘i’rcs2]*dep‘i’rcs2 + _b[dep‘i’rcs3]*dep‘i’rcs3, ///
ci(lhr‘i’_lci lhr‘i’_uci)
gen hr‘i’=exp(lhr‘i’)
gen hr‘i’_lci=exp(lhr‘i’_lci)
gen hr‘i’_uci=exp(lhr‘i’_uci)
}
local title2 "Q2"
local title3 "Q3"
local title4 "Q4"
local title5 "Q5"
forvalues i = 2/5 {
twoway (rline hr‘i’_lci hr‘i’_uci midtime if depr‘i’ == 1 & agep‘i’==1) ///
(line hr‘i’ midtime if depr‘i’ == 1 & agep‘i’==1, lpattern(solid)) ///
(function y = 1, lpattern(dash) lwidth(thin) range(0 5)) ///
,yscale(log) name(hr‘i’, replace) ///
ylabel(0.5 1 2 4 8, format(%3.1f) angle(h)) ///
xtitle("Years from Diagnosis") ///
ytitle("Excess Mortality Rate Ratio") ///
legend(off) nodraw ///
title("Deprivation ‘title‘i’’") ///
plotregion(style(none))
}
graph combine hr2 hr3 hr4 hr5, ycommon nocopies name(Figure2A)
drop lhr2 lhr2_lci lhr2_uci hr2 hr2_lci hr2_uci lhr3 lhr3_lci lhr3_uci hr3 hr3_lci hr3_uci ///
lhr4 lhr4_lci lhr4_uci hr4 hr4_lci hr4_uci ///
lhr5 lhr5_lci lhr5_uci hr5 hr5_lci hr5_uci
//Flexible PEREM model allowing for the time dependent effect of deprivation and correcting for overdispersion using robust SE
glm d rcs1-rcs3 dep?rcs? agep2-agep5 depr2-depr5, family(poisson) link(rs d_star) lnoffset(y) eform nolog baselevels vce(robust)
forvalues i = 2/5{
predictnl lhr‘i’ = _b[depr‘i’] + _b[agep‘i’] + _b[dep‘i’rcs1]*dep‘i’rcs1 + ///
    _b[dep‘i’rcs2]*dep‘i’rcs2 + _b[dep‘i’rcs3]*dep‘i’rcs3, ///
ci(lhr‘i’_lci lhr‘i’_uci)
gen hr‘i’=exp(lhr‘i’)
gen hr‘i’_lci=exp(lhr‘i’_lci)
gen hr‘i’_uci=exp(lhr‘i’_uci)
}
local title2 "Q2"
local title3 "Q3"
local title4 "Q4"
local title5 "Q5"
forvalues i = 2/5 {
twoway (rline hr‘i’_lci hr‘i’_uci midtime if depr‘i’ == 1 & agep‘i’==1) ///
(line hr‘i’ midtime if depr‘i’ == 1 & agep‘i’==1, lpattern(solid)) ///
(function y = 1, lpattern(vsdash) lwidth(medium thick) range(0 5)) ///
,yscale(log) name(hr‘i’, replace) ///
ylabel(0.5 1 2 4 8, format(%3.1f) angle(h)) ///
xtitle("Years from Diagnosis") ///
ytitle("Excess Mortality Rate Ratio") ///
legend(off) nodraw ///
title("Deprivation ‘title‘i’’") ///
plotregion(style(none))
}
```

```
graph combine hr2 hr3 hr4 hr5, ycommon nocopies name(Figure2B)
graph combine Figure2A.gph Figure2B.gph
```