SUPPLEMENTARY METHODS

Article Title: Characterising and quantifying the effects of breast cancer therapy using mathematical modelling

Journal: Breast Cancer Research and Treatment

Authors: Walter M Gregory[1], Christopher J Twelves[2], Richard Bell[4], Stephen W Smye[3], Dena R Howard[1], Robert E Coleman[5], David A. Cameron[6]

Corresponding author affiliation and email address: Walter Gregory, University of Leeds
Email: w.m.gregory@leeds.ac.uk

Mathematical derivation of the model.

The model assumes exponential growth of an initial residual tumour volume, v and is described in its original form in (Gregory et al 1991[1]) . Let the random variables V and G be normally distributed and represent the log of the resistant tumour burden, with mean μ_v and SD σ_v, and the log of the tumour doubling time, with mean μ_g and SD σ_g, Thus;

$$V \sim N(\mu_v,\sigma_v) \text{ and } G \sim N(\mu_g,\sigma_g).$$

Consider an individual tumour, i, reduced to a log volume v below the log relapse threshold, V_r, and having a log tumour-doubling time g (shown diagrammatically in supplementary figure 1). Then the probability, P_i of relapse before a given time t for this patient is given by integrating over all values of g which result in relapse before t;,

$$P_i = \int_{-\infty}^{\log_e\left[\frac{t}{V_r - v}\right] + \log_e[\log_e(2)]} \left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_g - g)^2}{2{\sigma_g}^2}\right) dg \qquad (1)$$

The upper limit of integration in the preceding equation is found by considering the growth of the log resistant tumour volume v, with a doubling time, say DT, such that relapse occurs at volume V_r at time t. Thus

$$\alpha t = V_r - v$$

where α is the exponential growth parameter; α is thus related to the doubling time by the following equation:

$$\alpha = \frac{log_e(2)}{DT}$$

and therefore $DT = \left[\frac{t}{V_r - v}\right] log_e(2)$. Thus

$$log_e(DT) = \log_e\left[\frac{t}{V_r - v}\right] + \log_e\left[\log_e(2)\right] =, say, U_t$$

The resistant tumour is log-normally distributed, and not all tumours necessarily achieve CR (v is not always less than V_r). Let the probability of achieving CR be P_c. Then the probability, P, of relapse before a given time t for the whole population is:

$$P = \frac{\int_{V_0}^{V_r}\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v - v)^2}{2{\sigma_v}^2}\right)\int_{-\infty}^{U_t}\left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_g - g)^2}{2{\sigma_g}^2}\right) dg\ dv}{P_c} \qquad (2)$$

where

$$P_c = \int_{-\infty}^{V_r}\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v - v)^2}{2{\sigma_v}^2}\right) dv \qquad (3)$$

as described.

Then the probability density function (pdf)

$$p(t, \mu_v, \sigma_v, \mu_g, \sigma_g) = \frac{dP}{dt}$$

which quantifies the relative likelihood of relapsing at any at given time, can be derived by differentiating under the double integral sign from equation (2). The terms P_c and $\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v-v)^2}{2{\sigma_v}^2}\right)$ do not involve t, so it is merely necessary to differentiate

$$\int_{-\infty}^{U_t}\left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{\left(\mu_g-g\right)^2}{2{\sigma_g}^2}\right) dg$$

with respect to t. This probability, P_i, as described, is the probability of relapse before a given time t for this particular patient, i. Since $g = \log_e\left[\frac{t}{V_r-v}\right] + \log_e[\log_e(2)]$

$$dg = \frac{dt}{t}$$

Therefore

$$\int_{-\infty}^{U_t}\left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{\left(\mu_g-g\right)^2}{2{\sigma_g}^2}\right) dg = P_i = \int_0^t\left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{\left(\mu_g-g\right)^2}{2{\sigma_g}^2}\right)\left(\frac{1}{t}\right) dt$$

So

$$\frac{dP_i}{dt} = \left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{\left(\mu_g-g\right)^2}{2{\sigma_g}^2}\right)\left(\frac{1}{t}\right)$$

and therefore

$$\frac{dP}{dt} = p = \frac{\int_{V_0}^{V_r}\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v-v)^2}{2{\sigma_v}^2}\right)\left(\frac{1}{\sigma_g\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_g-g)^2}{2{\sigma_g}^2}\right)\left(\frac{1}{t}\right) dv}{P_c} \qquad (4)$$

The likelihood is the product of the pdf probabilities for each completed time to relapse, and the probability of remaining in remission longer than the time under consideration for the censored times (which includes the probability of being cured, as well as the probability of relapse occurring after this time), thus;.

$$L(\mu_v, \sigma_v, \mu_g, \sigma_g) = \prod_{i=1}^{m} p(t_i,\mu_v, \sigma_v, \mu_g, \sigma_g) \prod_{i=m+1}^{n} (1-P(t_i,\mu_v, \sigma_v, \mu_g, \sigma_g))$$

where t_i, i=1,....,m are the completed times to relapse, and t_i, i=m+1,.....,n are the censored times to relapse.

The maximum likelihood estimates for $\mu_v,\sigma_v,\mu_g,\sigma_g$ can then be derived using first, plus preferably second, derivatives (see, for example, Beale, 1988). Second derivatives make the convergence routines much more rapid, and this is important for this model since it is necessary to derive a number of normal distribution integrals, which involve numerical methods. The mathematics for the first and second partial derivatives can be derived by differentiating under the integral given in (2) for all the parameters, and are given in Gregory (1993)[5]. The analysis, involves a number of functions similar to (2), (3) and (4) which proved quick to evaluate numerically because for all the double integrals the inner integral is identical, namely, P_i. Otherwise, the fitting procedure would have been computationally too slow to enable a multivariate approach such as this to be feasible (see for example, Day et al (1985)[33] to appreciate the difficulties inherent in this task when such a solution is not available,).

Extension to the multivariate case can be achieved by allowing prognostic factors to influence the mean (log) resistant tumor volume and the mean (log) growth rate in a linear fashion, enabling a set of regression coefficients to be produced for these variables. This is a key step in the extension of the original model. The log resistant tumor volume and log tumor doubling time, μ_v and μ_g, are normally distributed, where μ_v, μ_g are now linear functions of n prognostic variables, as follows:

$$\mu_v = \log(\text{residual tumor volume}) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

$$\mu_g = \log(\text{tumour doubling time}) = \gamma_0 + \sum_{i=1}^{n} \gamma_i x_i$$

with $x_1,...,x_n$ being the values of the n prognostic variables for a given individual, $\beta_1,...,\beta_n$, $\gamma_1,...,\gamma_n$ being the sets of regression coefficients, and β_0, γ_0 being baseline log resistant tumor and log tumor doubling time values (potentially for a patient having values of 0 for all the prognostic factors).

For this multivariate model the same equations apply with $\beta_0 + \sum_{i=1}^{n} \beta_i x_i$ replacing μ_v and $\gamma_0 + \sum_{i=1}^{n} \gamma_i x_i$ replacing μ_g. The maximum likelihood routine needs to include the β's and γ's, and thus maximise on 2n+4 parameters, rather than just 4. Therefore the first and second partial derivatives involving the βs and γs must be derived. For β_0 and γ_0 these are identical to the values derived for μ_v and μ_g and given in Gregory (1993)[5]. The first partial derivatives for the βs and γs can be derived from differentiating under the integrals given in equations (2) and (4). For the βs, as an example:

$$\frac{\partial P}{\partial \beta_i} = \frac{\left[\int_{V_0}^{V_r} x_i \left\{\frac{(v-[\beta_0 + \sum_{i=1}^{n} \beta_i x_i])}{{\sigma_v}^2}\right\}\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v - v)^2}{2{\sigma_v}^2}\right) P_i dv\right]}{P_c}$$

$$- \frac{\frac{\partial P_c}{\partial \beta_i}\left[\int_{V_0}^{V_r}\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v - v)^2}{2{\sigma_v}^2}\right) P_i dv\right]}{{P_c}^2} \qquad (5)$$

and where, Pi is as in equation (1) but with $\gamma_0 + \sum_{i=1}^{n} \gamma_i x_i$ replacing μ_g, and again differentiating under the integral,

$$\frac{\partial P_c}{\partial \beta_i} = \int_{-\infty}^{V_r} x_i \left\{\frac{(v-[\beta_0 + \sum_{i=1}^{n} \beta_i x_i])}{{\sigma_v}^2}\right\}\left(\frac{1}{\sigma_v\sqrt{2\pi}}\right) exp\left(-\frac{(\mu_v - v)^2}{2{\sigma_v}^2}\right) dv \qquad (6)$$

The x_i can be taken outside the integral in equations (5) and (6) and the first derivative ultimately reduces to:

$$\frac{\partial P}{\partial \beta_i} = x_i \frac{\partial P}{\partial \mu_v} \qquad (7)$$

The form of all the first and second partial derivatives of the β's and γ's is identical to that of μ_v and μ_g respectively but with the additional multiplication by x_i, x_j, $x_i x_j$, x_i^2, or x_j^2, and these x_i's etc. can be taken outside the integrals as in equation (7). Thus Newton's method can be used to fit the model, and derive the maximum likelihood estimates for the βs and γs.

Confidence intervals for the 4 + n model parameters are obtained from the variance for each of these parameters, which is found from the diagonal elements of the inverted information matrix of second partial derivatives, $I(\hat{\theta})$, where

$$I(\hat{\theta}) = (-\partial^2 \log L(\hat{\theta})/\partial\hat{\theta}_i \, \partial\hat{\theta}_j)_{(4+n)x(4+n)}$$

where $\theta = \theta_i$, i=1,…4+n are the model parameters, i.e. $\theta_1=\beta_0$, $\theta_2=\sigma_v$, $\theta_3=\gamma_0$, $\theta_4=\sigma_g$, and ($\theta i=\eta_{i-4}$, i=5,…n) where η_i is whichever β or γ is entered at step i.

Although the mathematics involved in these kinds of models can be very complex[30] this analytically tractable solution has enabled this model to be fitted on large numbers of patients, in a few seconds.

Note that, in a model having a full analytic solution, such as this one, model sensitivities to parameters can in principle be determined from the closed form solutions, without the necessity for computational simulation. For instance, model outputs include variance estimates for all the parameters, which can be used to provide confidence intervals for these estimates. Likelihood contour plots can also be used to examine the uniqueness of the maxima in the maximum likelihood estimates. As with other multivariate models, we have observed that model fits tend to improve with the addition of more clinical covariates, and for this model it can be observed that the standard deviations of both resistant disease and tumour doubling times gradually reduce as more clinical covariates are included.

Stepwise approach to fitting the model

A forward stepwise approach is used. Each treatment/prognostic factor is considered for inclusion both for its effect on resistant disease and growth rates at each step. Thus, initially the model is applied with just the four baseline parameters (the means and standard deviations of the logs of resistant disease and of growth rates) giving a baseline likelihood. Then each factor is considered in turn both for its resistant disease effect and growth rate effect, giving a series of new likelihoods and AICs. A χ^2 difference test/statistic can then be calculated on the resulting AICs (see, for example, Hatcher, 1994) to yield a χ^2 statistic and associated p-value for inclusion of the factor in the model. The factor giving the largest χ^2 is then included, giving a new, baseline likelihood and the process repeated (see table 1) until no factors have significance levels < 0.05.

Comparison with the Cox model, and comparisons of goodness of fit, for the AZURE data

Firstly, note that Histological grade was included as two variables, namely grade 3 versus grades 1 and 2, and grade 2 versus grades 1 and 3 since the differences between the grades are not uniform. T3 and T4 tumors were combined, having nearly identical IDFS curves.

As each factor is included in the model the fit improves, and the measure of how much it improves is the improvement in AIC (ΔAIC), with an associated χ^2 difference test/statistic. Improvements in the Cox model fit can be similarly evaluated through improvements in the partial likelihood[6]. The fits of the two models cannot be compared directly because the Cox model approach is based on a partial likelihood, but the difference between these two AICs, which incorporate allowance for the 4 extra parameters in the mathematical model as compared to the Cox model, provides some indication of the difference

in model fit. AIC differences > 10 are considered large (see Hilbe, 2011) and differences in this paper are considerably greater than this suggesting that such differences are robust.

Overall, the ΔAIC for the biologically-based model (505.2) is vastly better than for the Cox model (361.4) suggesting greatly improved model fits (ΔAIC (difference) = 126.8 >>10). Similarly, the significance of ER status was considerably greater with the biologically-based model than when derived from the Cox model. The univariate ΔAICs for inclusion for the biologically-based and Cox models were 197.4 and 102.1 respectively, with multivariate ΔAICs of 109.0 and 40.9 respectively (ΔAIC (differences) = 95.3 and 68.1 respectively, again suggesting greatly improved fits in both cases). Finally, when considering whether the difference in the ER status DFS curves was more likely to result from a difference in doubling times or a difference in residual disease, the improvement in AIC (ΔAIC) for the biologically-based model was 111.4 for attributing this difference to a difference in doubling times compared to only 58.4 for attributing this to a difference in residual disease (such an analysis is obviously not possible for the Cox model). Note that for ER status higher relapse rates are seen in ER-negative women in early years, whereas in later years this is reversed so that ER-positive women have higher relapse rates.

Neither model accounts for missing data, which can also significantly impact DFS/survival curves, but which is a complicated topic in its own right, and is not covered in this manuscript.

Computer programs for fitting the model

A related interactive computer program for fitting the model is available on request. We are also working on a web-based version of this program which will be directly available on the flexsurv platform (see, for example: http://cran.r-project.org/web/packages/flexsurv/index.html).

SUPPLEMENTARY METHODS REFERENCES

Beale EM: Multi-dimensional Optimization. In Beale EM (ed): Introduction to Optimization. Chichester, John Wiley & Sons. 1988;25-36.

Day R, Shackney S, Peters W: The analysis of relapse-free survival curves: implications for evaluating intensive systemic adjuvant treatment regimens for breast cancer. Br J Cancer 2005;92:47-54.

Gregory WM, Richards MA, Slevin ML, Souhami RL (1991) A mathematical model relating response durations to amount of sub-clinical resistant disease. Cancer Res. 51:1210-1216

Gregory WM: The evaluation of durations of response to cancer treatments [dissertation]. University College London. 1993.

Hatcher L. A Step-By-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling (1st ed.). SAS Publishing. 1994 pp 201-224.

Hilbe JM. Negative Binomial Regression, Cambridge University Press, 2011.

Supplementary Table 1 Stepwise results for the multivariate model application in the AZURE trial.

	Step 0†			Step 1			Step 2		
	χ^2	p-value	coeff.	χ^2	p-value	coeff.	χ^2	p-value	coeff.
RD:T Stage	140.5	<.0001		124.0	<.0001		69.5	<.0001	
RD:ER status	121.9	<.0001		0.00	.97		0.2	.89	
RD:Lymph nodes	165.9	<.0001		150.7	<.0001		*150.7	<.0001	2.81
RD:Grade 1&2v3	86.2	<.0001		33.5	<.0001		27.7	<.0001	
RD:Grade 1v2&3	29.2	<.0001		6.4	.01		6.5	.01	
DT:T Stage	142.6	<.0001		121.3	<.0001		75.9	<.0001	
DT:ER status	197.4	<.0001		*197.5	<.0001	.48	*182.2	<.0001	.46
DT:Lymph nodes	166.6	<.0001		130.8	<.0001		4.7	.03	
DT:Grade 1&2v3	140.2	<.0001		56.1	<.0001		48.8	<.0001	
DT:Grade 1v2&3	60.8	<.0001		42.9	<.0001		31.7	<.0001	

	Step 3			Step 4			Step 5		
	χ^2	p-value	coeff.	χ^2	p-value	coeff.	χ^2	p-value	coeff.
RD:T Stage	6.6	.01		7.2	.007		6.4	.01	
RD:ER status	3.4	.07		1.9	.16		2.3	.13	
RD:Lymph nodes	*105.3	<.0001	2.41	*103.0	<.0001	2.34	*99.9	<.0001	2.35
RD:Grade 1&2v3	20.1	<.0001		3.9	.05		5.9	.02	
RD:Grade 1v2&3	5.2	.02		13.9	.0002		16.5	<.0001	
DT:T Stage	*75.9	<.0001	-.15	*69.7	<.0001	-.14	*65.7	<.0001	-.13
DT:ER status	*174.9	<.0001	.44	*111.1	<.0001	.36	*109.8	<.0001	.35
DT:Lymph nodes	1.5	.22		2.2	.14		1.8	.18	
DT:Grade 1&2v3	42.6	<.0001		*42.6	<.0001	-.19	*32.0	<.0001	-.16
DT:Grade 1v2&3	25.8	<.0001		15.2	.0001		*15.2	.0001	-.23

	Step 6			Step 7		
	χ^2	p-value	coeff.	χ^2	p-value	coeff.
RD:T Stage	6.8	.01		*6.8	.009	.20
RD:ER status	1.4	.23		0.2	.67	
RD:Lymph nodes	*103.5	<.0001	2.20	*92.6	<.0001	2.12
RD:Grade 1&2v3	3.2	.07		2.5	.11	
RD:Grade 1v2&3	*16.5	<.0001	12.4	*16.9	<.0001	12.4
DT:T Stage	*66.5	<.0001	-.13	*15.9	.0001	-.09
DT:ER status	*111.4	<.0001	.35	*110.7	<.0001	.34
DT:Lymph nodes	1.5	.21		2.9	.09	
DT:Grade 1&2v3	*32.8	<.0001	-.16	*33.9	<.0001	-.17
DT:Grade 1v2&3	*22.6	<.0001	-1.3	*21.9	<.0001	-1.3

RD: residual disease; DT: doubling time.

χ^2s and p-values are for inclusion of this variable in the model at this step, unless preceded with *, which indicates χ^2s and p-values are for removal of a variable which was previously included.

† Step 0 (no variables in the model) gives the univariate model χ^2s & p-values for each factor.

Supplementary methods figure caption

Supplementary figure 1. IDFS by ER-status with mathematical model fits and Cox model fits for the AZURE trial.

Supplementary figure 1