STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:581–595
Published online 5 April 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2550

A natural history model of stage progression applied to breast cancer

Sylvia K. Plevritis[1,*,†] Peter Salzman[1,2,‡] Bronislava M. Sigal[1,§] and Peter W. Glynn[2,¶]

[1]*Department of Radiology, Stanford University, CA, U.S.A.*
[2]*Department of Management Science and Engineering, Stanford University, CA, U.S.A.*

SUMMARY

Invasive breast cancer is commonly staged as local, regional or distant disease. We present a stochastic model of the natural history of invasive breast cancer that quantifies (1) the relative rate that the disease transitions from the local, regional to distant stages, (2) the tumour volume at the stage transitions and (3) the impact of symptom-prompted detection on the tumour size and stage of invasive breast cancer in a population not screened by mammography. By symptom-prompted detection, we refer to tumour detection that results when symptoms appear that prompt the patient to seek clinical care. The model assumes exponential tumour growth and volume-dependent hazard functions for the times to symptomatic detection and stage transitions. Maximum likelihood parameter estimates are obtained based on SEER data on the tumour size and stage of invasive breast cancer from patients who were symptomatically detected in the absence of screening mammography. Our results indicate that the rate of symptom-prompted detection is similar to the rate of transition from the local to regional stage and an order of magnitude larger than the rate of transition from the regional to distant stage. We demonstrate that, in the even absence of screening mammography, symptom-prompted detection has a large effect on reducing the occurrence of distant staged disease at initial diagnosis. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: natural history model; stage progression; breast cancer; SEER; tumour growth; symptomatic detection

*Correspondence to: Sylvia K. Plevritis, LUCAS Center Room P267, Department of Radiology, Stanford University, CA 94305-5488, U.S.A.
†E-mail: sylvia.plevritis@stanford.edu
‡E-mail: psalzman@bst.rochester.edu
§E-mail: slava@stanford.edu
¶E-mail: glynn@stanford.edu

Contract/grant sponsor: National Cancer Institute; contract/grant numbers: R01 CA82904, U01 CA097420

Received 4 April 2005
Accepted 2 January 2006

1. INTRODUCTION

Invasive breast cancer is commonly staged as local, regional or distant disease [1]. Larger tumours are more likely to be detected in advanced stages when compared to smaller tumours. Because of this observation, it is generally believed that the majority of invasive breast cancer begins in the local stage, and transitions from local to regional then to distant stages as the primary tumour increases in size. However, the rate of stage progression is not known. The primary tumour's size at stage transitions is not known. Even the effect of symptom-prompted detection on the observed stage distribution is not known. These quantities cannot be ethically observed but are biologically and clinically meaningful since they would provide insight into progression of the disease and the expected benefits attributable to cancer control programs in early detection. We propose a parametric, stochastic model of the natural history of invasive breast cancer that estimates: (1) the relative rates of stage transitions, (2) the primary tumour volume at stage transitions and (3) the effect of symptom-prompted detection on the observed stage distribution. By symptom-prompted detection, we refer to tumour detection that occurs when symptoms prompt the patient to seek clinical care.

Numerous models of the natural history of breast cancer have been proposed. Here we identify only those models that aim to better understand the rate that the disease progresses through the local, regional and distant stages, or the primary tumour size at these stage transitions or both. Koscielny *et al.* estimate the tumour volume at which remote metastases first occurs using survival data [2]. Kimmel *et al.* estimate the primary tumour size at transition from non-metastatic to metastatic states non-parametrically [3]. Atkinson *et al.* [4] and Bartoszynski [5, 6] explore various models for metastatic shedding, including ones where shedding is a function of tumour volume, with a systemic component. Shwartz [7, 8] proposes a comprehensive model that estimates the effect of screening mammography on breast cancer mortality and embedded in this model is a natural history submodel that estimates the tumour size at nodal involvement. In Section 7, we will compare these approaches with our approach. Duffy *et al.* [9] and their related publications [10, 11] estimate breast cancer stage transitions with a Markov chain model where parameter estimation relies on screening trial data. This approach is not directly comparable to ours. It estimates transitions between preclinical and clinical states of the disease, where the preclinical states are defined based on the probability of screen detection by mammography. In this work, we estimate the natural history of the breast cancer in a population which is not screened by mammography.

This work extends the basic ideas presented in References [4–6, 12, 13], which were primarily aimed at estimating functions of breast tumour growth and symptomatic detection, given only information on the primary tumour size at symptomatic detection. Our work differs in that we aim to elucidate the progression of the disease from local to regional stages and from regional to distant stages and the impact of symptom-prompted detection on the stage distribution in a population which is not undergoing mammographic screening.

2. MODEL ASSUMPTIONS

Our natural history model of invasive breast cancer is based on the five assumptions itemized below that describe the growth of the primary tumour, stage progression and symptomatic detection in the absence of screening. Assumptions 1–3 were adopted from References [4–6, 12, 13], but explored here together with Assumptions 4–5.

Assumption 1
The tumour volume grows exponentially from volume c_0. The volume at time t is represented as $V(t) = c_0 \exp^{t/R}$ where R is a random variable that represents the inverse growth rate. We assume spheroidal tumours. We assume the initial tumour volume is $c_0 = \pi/6\, d_0^3$ with d_0 is the tumour diameter, with $d_0 = 2\,\text{mm}$. We do not model the natural history of the tumour prior to c_0 and we do not consider an *in situ* stage of the disease. *Note*: R is proportional to the tumour volume doubling time (DT), $\text{DT} = \ln(2) * R$.

Assumption 2
The inverse growth rate R is a random variable that is gamma distributed with parameters α and β,

$$\Gamma_R(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, r^{\alpha-1} \exp(-\beta r)$$

Note: $E(R) = \alpha/\beta$

Assumption 3
T_{det} is a random variable that represents the time of symptomatic detection measured from the moment the tumour volume is c_0. The hazard function of T_{det} is $\gamma V(t)$, that is, $\Pr(T_{\text{det}} \in [t, t + \mathrm{d}t) | T_{\text{det}} > t) = \gamma V(t)\, \mathrm{d}t + o(\mathrm{d}t)$. The larger the volume, the more likely it will be detected.

Assumption 4
The tumour at volume c_0 is in the local stage. The transition from the local to the regional stage is defined to occur at the moment nodal involvement first becomes detectable by usual clinical care. Let T_{reg} be a random variable representing the time at which the disease transitions from the local to regional stage measured from the moment the tumour volume is c_0. The hazard function of T_{reg} is $\eta V(t)$, and expressed as

$$\Pr(T_{\text{reg}} \in [t, t + \mathrm{d}t) | T_{\text{reg}} > t) = \eta V(t)\, \mathrm{d}t + o(\mathrm{d}t)$$

Assumption 5
The transition from the regional to distant stage is defined to occur at the moment distant metastatic disease first becomes detectable by usual clinical care. Let T_{dist} be a random variable representing the time of onset of the distant stage measured from the moment the tumour volume is c_0. The hazard function of T_{dist} is assumed to be zero until the onset of regional disease and then $\omega V(t)$, and expressed as

$$\Pr(T_{\text{dist}} \in [t, t + \mathrm{d}t) | T_{\text{dist}} > t, T_{\text{reg}} = t_{\text{reg}}) = \begin{cases} \omega V(t)\, \mathrm{d}t + o(\mathrm{d}t), & t > t_{\text{reg}} \\ 0, & t \leqslant t_{\text{reg}} \end{cases}$$

An implicit assumption made here is that all patients who are initially diagnosed in the distant stage also have clinically detectable nodal involvement.

3. MODEL PROPERTIES

Using the model assumptions described above, we derive closed form expressions for: (1) the tumour volume distribution at symptomatic detection, (2) the tumour volume

distribution at the transition from local to regional stages and (3) the tumour volume distribution at the transition from regional to distant stages.

3.1. Tumour volume distribution at symptomatic detection

Let $D = V(T_{\text{det}})$ represent the tumour volume at symptomatic detection. The conditional distribution of $D - c_0$ is

$$\begin{aligned}\Pr(D < d|R = r) &= \Pr(V(T_{\text{det}}) < d|R = r)\\ &= \Pr(T_{\text{det}} < V^{-1}(d)|R = r)\\ &= 1 - \exp\left(-\int_0^{V^{-1}(d)} \gamma V(t)\,\mathrm{d}t\right)\\ &= 1 - \exp(-\gamma r(d - c_0))\end{aligned}$$

Given $R = r$, D is exponentially distributed with mean $(\gamma r)^{-1}$. A faster growing tumour is more likely to be detected at a larger volume than a slower growing tumour. The unconditional density function of D is

$$f_D(d) = \alpha\beta^{\alpha}\gamma[\beta + \gamma(d - c_0)]^{-(\alpha+1)}, \quad d > c_0$$

3.2. Tumour volume distribution at the transition from local to regional stage

Let $N = V(T_{\text{reg}})$ represent the tumour volume at the transition from the local to regional stage. Given $R = r$, $N - c_0$ is exponentially distributed with mean $(\eta r)^{-1}$, and expressed as

$$\Pr(N < n|R = r) = 1 - \exp\{-\eta r(n - c_0)\}$$

which is similar to the distribution of the $D - c_0$, with η in place of γ. Once metastatic disease can be detected in the lymphnodes, a faster growing tumour is more likely to be larger compared to a slower growing tumour. The unconditional density function of N is

$$f_N(n) = \alpha\beta^{\alpha}\eta[\beta + \eta(n - c_0)]^{-(\alpha+1)}, \quad n > c_0$$

3.3. Tumour volume distribution at the transition from regional to distant stage

Let $M = V(T_{\text{dist}})$ represent the tumour volume at the transition from the regional to distant stage. The distribution of M, conditioned on $R = r$ and $V(T_{\text{dist}}) = n$, is

$$\begin{aligned}&\Pr[V(T_{\text{dist}}) < m|V(T_{\text{reg}}) = n, R = r]\\ &= \Pr[T_{\text{dist}} < V^{-1}(m)|T_{\text{reg}} = V^{-1}(n), R = r]\\ &= 1 - \exp\left(-\int_{V^{-1}(n)}^{V^{-1}(m)} \omega V(t)\,\mathrm{d}t\right)\\ &= \begin{cases} 1 - \exp(-\omega r(m - n)), & n \leqslant m\\ 0, & n > m\end{cases}\end{aligned}$$

The density of M, conditioned on $R=r$, is

$$f_{M|R}(m|r) = \int_{c_0}^{m} f_{M|N,R}(m|n,r) f_{N|R}(n|r)\,\mathrm{d}n$$
$$= \left(\frac{\eta\omega}{\eta+\omega}\right) r[\exp(-r\omega(m-c_0)) - \exp(-r\eta(m-c_0))]$$

The unconditional density of M is

$$f_M(m) = \alpha\beta^{\alpha}\left(\frac{\eta\omega}{\eta-\omega}\right)[(\beta+\omega(m-c_0))^{-(\alpha+1)} - (\beta+\eta(m-c_0))^{-(\alpha+1)}]$$

4. DATA

We estimate model parameters with data on the tumour size and stage of breast cancer for patients who were symptomatically-detected with invasive disease and recorded by the Surveillance, Epidemiology and End Results (SEER) program [14]. Our patient population was selected from SEER using the following criterion: (1) female only, (2) invasive disease only, (3) primary breast cancer diagnosed between the years 1975 and 1981, which corresponds to a period of negligible levels of screening mammography [15], (4) age of initial diagnosis between 40 years and 80 years; and (5) breast cancer as the first malignant primary. Special permission was obtained from the NCI SEER program to access information on tumour sizes before 1982 since this information is not reported in Public-Use files due to coding changes. Tumour sizes were recorded in 7 bins: <5 mm, 5–9 mm, 1–1.9 cm, 2–2.9 cm, 3.0–3.9 cm, 4.0–4.9 cm and 5.0+ cm. We excluded records where the tumour size was recorded as 0 or 'diffuse'. SEER historic stage was used to define local, regional and distant disease. A total of 35 299 patient records were selected, as summarized in Table I.

Table I. The number of SEER breast cancer patients ages 40–80, diagnosed from 1975–1981, stratified by tumour size and stage.

	Stage		
Tumour size	Local	Regional	Distant
$<$ 5 mm	299	83	7
5–9 mm	1140	291	26
1–1.9 cm	6264	3385	218
2–2.9 cm	5276	4466	334
3–3.9 cm	2590	3157	320
4–4.9 cm	1086	1740	224
5 cm+	1021	2739	633

5. ESTIMATION PROCEDURE

We express the likelihood function for the data in Table I assuming a multinomial distribution where the observed outcome for an individual breast cancer patient is her tumour size and stage at symptomatic detection. Let I be an indicator function defined as 0, 1 or 2 if the patient is staged with local, regional or distant disease, respectively. The tumour size is binned according to the SEER data and the kth tumour size bin belongs to tumours in the range $[d_k, d_{k+1}]$. Let Q_k represent the total number of symptomatically detected breast cancer patients whose tumour falls in the kth bin. Let p_k, q_k and hence $Q_k - p_k - q_k$ be the number of patients with local, regional and distant stage disease, respectively. The likelihood function is

$$L = \prod_{k=1}^{K} \Pr(D \in (d_k, d_{k+1}), I = 0)^{p_k} \Pr(D \in (d_k, d_{k+1}), I = 1)^{q_k} \Pr(D \in (d_k, d_{k+1}), I = 2)^{Q_k - p_k - q_k}$$

where K is the total number of tumour size bins.

For simplicity in notation, we derive the joint density of $\{D, I\}$ for continuous-valued tumour sizes. The joint density for discrete tumour size bins is given in the Appendix.

5.1. Symptomatic detection in the local stage

A patient is staged with local disease if $N > d$, meaning that the patient's disease would transition to the regional stage after symptomatic detection. The joint density of $\{D, I = 0\}$, conditioned on $R = r$, is

$$\begin{aligned} f_{D,I|R}(d, 0|r) &= f_{D|R}(d|r) \times \Pr(N > d | R = r) \\ &= \gamma r \exp\{-(\gamma + \eta) r (d - c_0)\} \end{aligned}$$

The unconditional joint density is

$$f_{D,I}(d, 0) = \alpha \beta^{\alpha} \gamma (\beta + (\eta + \gamma)(d - c_0))^{-(\alpha+1)}$$

5.2. Symptomatic detection in the regional stage

A patient is staged with regional disease if $N < d < M$, meaning that the patient's disease is symptomatically detected after the transition to the regional stage but before the transition to the distant stage. The joint density of $\{D, I = 1\}$, conditioned on $R = r$, is

$$\begin{aligned} f_{D,I|R}(d, 1|r) &= f_{D|R}(d|r) \times \int_{c_0}^{d} \Pr(V(T_{\text{dist}}) > d | R = r, N = n) f_{N|R}(n|r)\, \mathrm{d}n \\ &= \frac{\eta\gamma}{(\eta - \omega)} r [\exp\{-r((\gamma + \omega)(d - c_0))\} - \exp\{-r((\gamma + \eta)(d - c_0))\}] \end{aligned}$$

The unconditional joint density is

$$f_{D,I}(d, 1) = \alpha \beta^{\alpha} \frac{\eta\gamma}{(\eta - \omega)} [(\beta + (\gamma + \omega)(d - c_0))^{-(\alpha+1)} - (\beta + (\gamma + \eta)(d - c_0))^{-(\alpha+1)}]$$

5.3. Symptomatic detection in the distant stage

The joint density of $\{D, I=2\}$ can be expressed as

$$f_{D,I}(d,2) = f_D(d) - f_{D,I}(d,0) - f_{D,I}(d,1)$$

where $f_D(d)$ is given in Section 3.1.

5.4. Non-identifability of absolute rate parameters

The model consists of five parameters $(\gamma, \eta, \omega, \alpha, \beta)$. Because the observed data (namely, the tumour size and stage at symptomatic detection) does not contain temporal information, the model cannot infer elements of time. Hence, the likelihood function does not change when the rate-related parameters $(\gamma, \eta, \omega, \alpha/\beta)$ are scaled by a constant. We maximize the likelihood by constraining the expected value of R, such that $\alpha = \beta$ (or equivalently, $E(R) = 1$), and report on the relative rates of stage transitions. The tumour size distributions at symptomatic detection and the stage transitions are not impacted since they involve the ratios of various combinations of the parameters $(\gamma, \eta, \omega, \alpha/\beta)$.

6. RESULTS

Optimization of the likelihood function for parameter estimation was performed in MATLAB using the Nelder–Mead simplex (direct search) method [16]. Table II gives maximum likelihood estimates, with asymptotic confidence intervals, conditioned on $\alpha = \beta$. The results indicate that the rate of symptomatic detection is similar to the rate of transition from the local to regional stage, and is an order of magnitude larger than the transition rate from the regional to distant stage.

Figures 1(a)–(c) compare the model fit to the data, where the data is summarized as 21 pairwise frequencies describing the observed joint distribution of tumour size and stage. The fit is close, but not exact. We found that it does not pass traditional quantitative measures for the goodness of fit, such as the Chi-square test. Many models would not pass such measures given the large amount of underlying data (35 299 observations are reported in Table I).

Table II. Maximum likelihood estimates for model parameters, with asymptotic confidence intervals conditioned on $\alpha = \beta$ or similarly $E(R) = \alpha/\beta = 1$. Without knowledge of $E(R)$, only the ratio of the parameters $\hat{\alpha}$, $\hat{\gamma}$, $\hat{\eta}$ and $\hat{\omega}$ can be inferred. If $E(R)$ were known, then $\hat{\gamma} = \exp(-9.602)/E(R)$, $\hat{\eta} = \exp(-9.636)/E(R)$, $\hat{\omega} = \exp(-11.765)/E(R)$, $\hat{\beta} = \exp(-0.165)/E(R)$ and $\hat{\alpha} = \hat{\beta} \times E(R)$.

	Estimate	95% CI
$\ln(\hat{\gamma})$	-9.602	$[-9.624, -9.580]$
$\ln(\hat{\eta})$	-9.636	$[-9.661, -9.610]$
$\ln(\hat{\omega})$	-11.765	$[-11.816, -11.713]$
$\ln(\hat{\beta})$	-0.165	$[-0.187, -0.143]$
$\hat{\alpha}$	$\hat{\beta}$	—

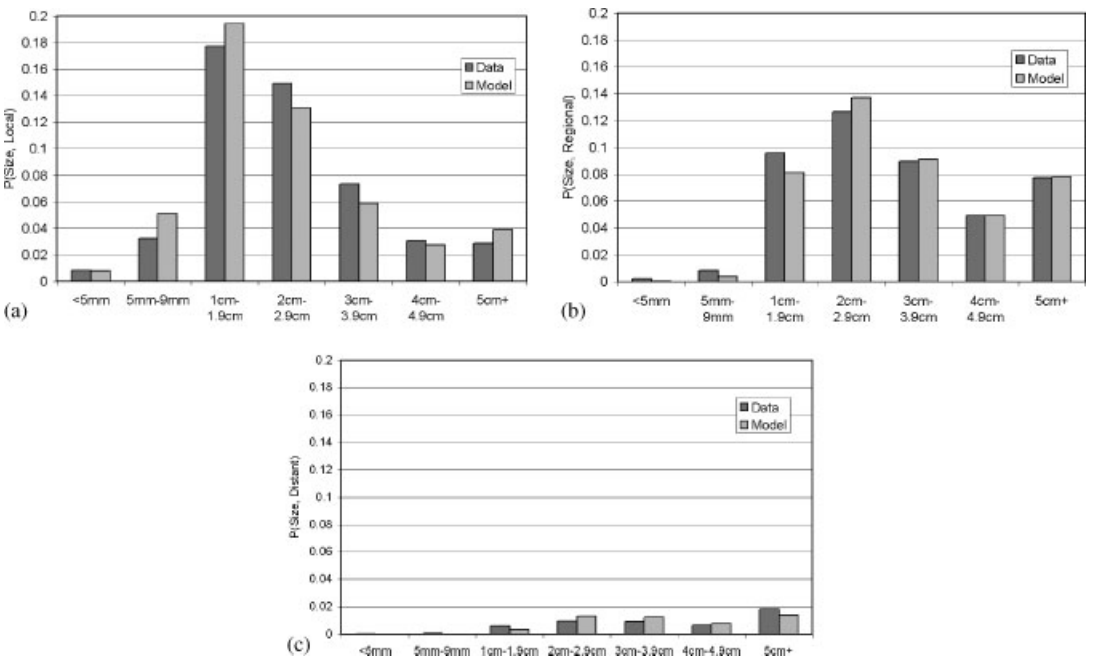

Figure 1. (a) Joint distribution of detected tumour size and local stage disease, comparing model and data; (b) Joint distribution of detected tumour size and regional stage disease, comparing model and data; and (c) Joint distribution of detected tumour size and distant stage disease, comparing model and data.

A linear regression of the 21 observed frequencies on those estimated yields an intercept of 0.00137 with bootstrap confidence interval (0.00101, 0.00172) and a slope of 0.971, bootstrap CI (0.964, 0.979). The correlation coefficient between the observations and estimates is 0.985, bootstrap CI (0.982, 0.987). Note that all of these CIs are conditioned on $\alpha = \beta$.

Figures 2(a) and (b) provide an alternative representation of the model fit. Figure 2(a) shows that the model produces a good fit to the tumour size distribution. Figure 2(b) shows the model fit to the stage distribution conditioned on tumour size and demonstrates that the model more closely reproduces the stage distribution for tumours above 1cm than below 1 cm. Tumours below 1 cm (which are approximately 5 per cent of the data) are less likely to be staged as local disease than predicted by the model; the model predicts that over 90 per cent will be staged local, whereas just less than 80 per cent are staged as local in the data. Among these small, advanced stage tumours, the model correctly predicts that the probability of distant disease is at least an order of magnitude lower than the probability of regional disease. With this understanding of the model fit, we proceed to analyse the model's properties.

Figure 3(a) and Figures 4(a)–(c) are generated from the model using parameter estimates in Table II, but with γ set to zero, in order to evaluate the hypothetical scenario that assumes no attention is given to breast cancer symptoms. Figure 3(a) shows the predicted stage distribution conditioned on tumour size. As the tumour size increases, the proportion of local disease decreases to zero and the proportion of distant disease increases to approximately 40 per cent

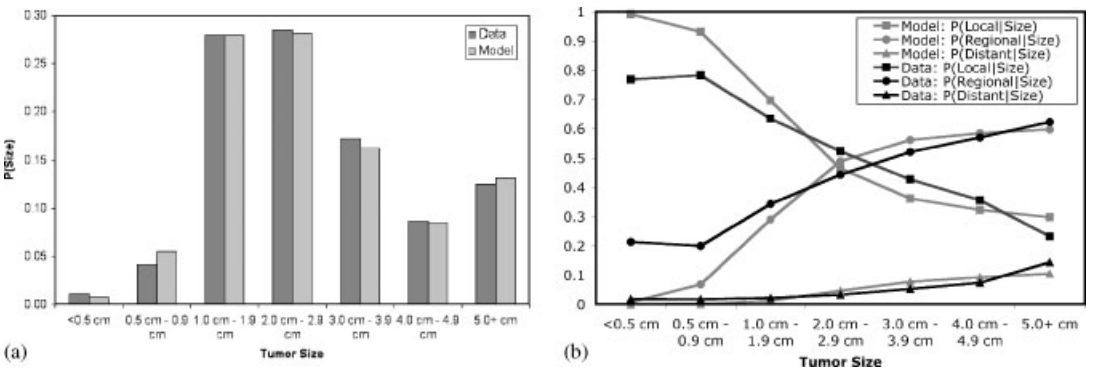

Figure 2. (a) Distribution of detected tumour size, comparing model and data; and (b) Distribution of detected stage (local, regional, distant) conditioned on detected tumour size, comparing model and data.

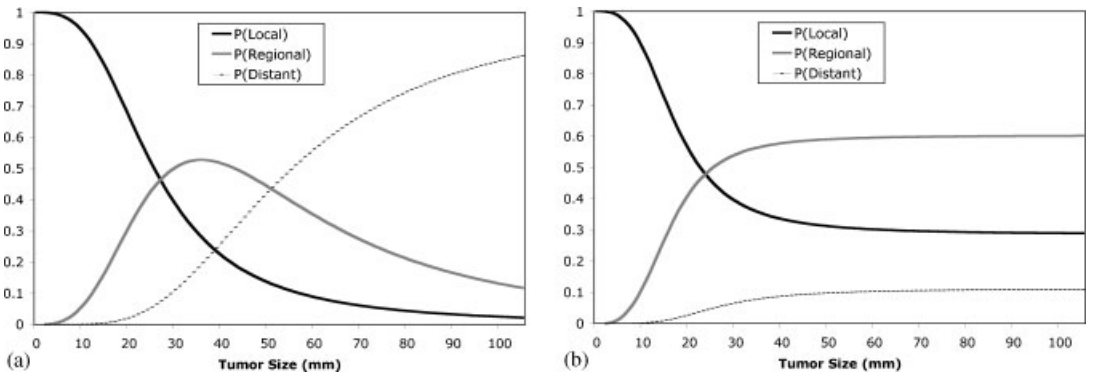

Figure 3. (a) Estimated breast cancer stage distribution conditioned on tumour size assuming no symptomatic detection mechanism, based on parameter estimates for η, ω, α, and β in Table II, with $\gamma = 0$; and (b) Estimated breast cancer stage distribution conditioned on the clinically detected tumour size, using the estimates for γ, η, ω, α, and β in Table II.

when the tumour reaches 5 cm. The median tumour size at the transition from the local to regional stage is 25.4 mm, and the median tumour size at the transition from the regional to distant stage is 55.2 mm. The model may be overestimating the tumour size at stage transition since it underestimates the proportion of advanced-staged tumours symptomatically detected below 1cm.

If we hold the assumption that no attention is given to clinical symptoms (i.e. $\gamma = 0$), Figure 4(a) shows the diameter of the primary tumour at time T, where T is measured from the moment that the tumour is 2 mm in diameter; Figure 4(b) shows the probability that the disease had transitioned to nodal involvement by time T; and Figure 4(c) shows the probability that the disease has transitioned to the distant stage by time T. Figures 4(a)–(c)

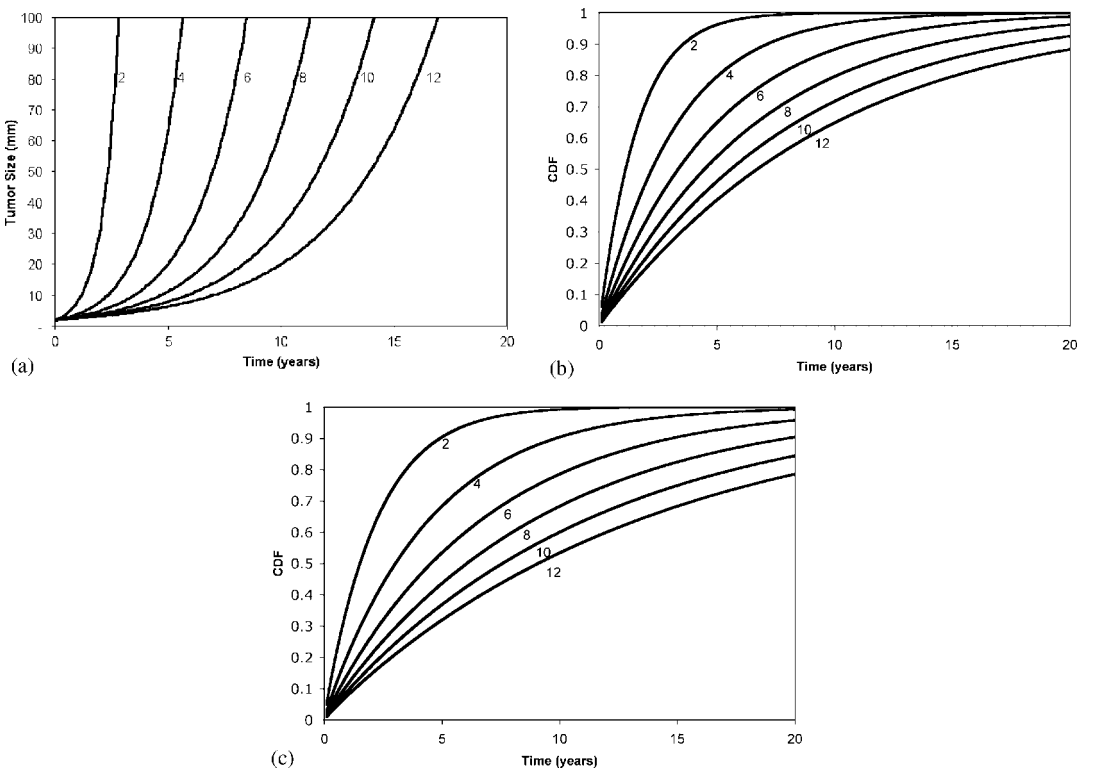

Figure 4. (a) Diameter of the primary tumour at time T; (b) probability of regional staged disease by time T; and (c) probability of distant staged disease by time T. $T = 0$ corresponds to a 2 mm tumour. Numbers to right of curves represent mean tumour volume doubling time, in months.

are generated for a range of tumour volume doubling times. For a mean DT of 8 months, by $T = 8$ years, an 3.2 cm tumour would have 72 per cent chance of nodal involvement and a 60 per cent chance of distant involvement.

With symptom-prompted detection, the estimated stage distribution conditioned on tumour size is shown in Figure 3(b) using the parameter estimates in Table II. Even though this distribution is observable, expressing it in terms of the model parameters provides several insights. These findings suggest that current practices of symptom-prompted detection are preventing a significant fraction of tumours from progressing to advanced stages, particularly the distant stage. Due to symptomatic detection, the proportion of distant disease is roughly 10 per cent, compared to over 40 per cent as tumour continues grows to 5 cm and greater in the absence of symptom-prompted detection. For the larger sized symptomatically detected tumours, the proportion of local tumours converges to $(\gamma/\eta + \gamma)^{(\alpha+1)}$ as the detected tumour size approaches infinity. The overall proportion of symptomatically detected local tumours is $\Pr(I = 0) = \gamma/(\gamma + \eta)$. As γ increases, the portion of local staged disease increases, as would be expected by our modelling assumptions.

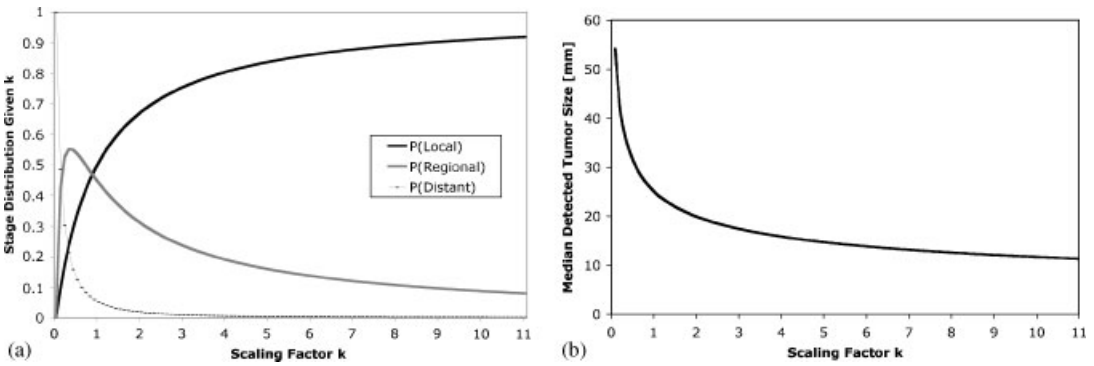

Figure 5. (a) The stage distribution of clinically detected breast tumours versus the symptomatic detection function scaling factor k; and (b) The median size of clinically detected breast tumours versus the symptomatic detection function scaling factor k.

Consider a new hazard function for the time to symptomatic detection that is $k_\gamma V(t)$ where $k>0$ is scaling factor. The impact of k on the overall stage distribution and the median tumour size for symptomatically detected breast cancer is given in Figures 5(a) and (b), respectively. Compared to $k=1$, for $k=2$ the median tumour size decreases from 2.5 to 2.0 cm (20 per cent reduction) and the proportion of distant disease decreases from 5.1 to 1.8 per cent (65 per cent reduction). Assuming a mean tumour volume doubling time of 8 months, $k=2$ would advance the median time of symptomatic detection by 5.3 months.

7. DISCUSSION

We proposed a natural history model of breast cancer that relies on three simple, but biologically and clinically reasonable, assumptions: (1) the primary tumour grows exponentially from 2 mm, (2) the disease progresses from local, regional to distant stages, and (3) the hazard function of the times to stage progression and symptomatic detection are proportional to the volume. We produced a trackable, closed form expression for the likelihood function using data on the tumour size and stage of invasive breast cancer for patients who were symptomatically detected during the period preceding the wide dissemination of screening mammography.

The model gives a reasonable, but not exact fit, to the data. The model fit is some-what better for symptom-detected tumours detected above 1 cm, than those below 1cm. For tumours below 1 cm, which represent about 5 per cent of the population, the model overestimates the proportion of local disease and underestimates the proportion of advanced stage disease. The weaker fit for the smaller tumours (<1 cm) suggests that there may be a small subset of tumours that progress to advanced disease stages faster than the majority of the tumours. These tumours may progress to advanced stages disease under a different mechanism than the one proposed here. Factors not considered in our model include the spatial location of the primary tumour, which may have a role in determining its propensity to successfully metastasize to the lymphnodes and beyond.

Since the model comes close to explaining the detected tumour size and stage for invasive breast cancer, we use it to predict characteristics of disease that cannot be observed for ethical reasons. The model indicates that the rate of symptom-prompted detection and the rate of transition from local to regional stage are similar and an order of magnitude larger than the rate of transition from regional to distant stage. This finding may be of particular interest to policy makers who are concerned with the value of educating women on how to identify early clinical symptoms of breast cancer. Our findings imply that clinical attention to symptoms is preventing a significant fraction of patients from presenting with advanced stage disease at initial diagnosis. We showed how hypothetically scaled versions of the symptomatic detection function would further impact the tumour size and stage distribution, but this exercise does not identify how to achieve such an improvement in the symptomatic detection.

Our findings for the median tumour volume at stage transitions are fairly consistent with published results. The work by Kimmel *et al.*, in Reference [3], reported 'single nodal metastases have a probability of 0.5 of developing from primary tumours by the time they reach 2 cm in diameter.' Our work reports on the median tumour size at the transition from local to regional at 2.5 cm. Even though we do not report on the first nodal involvement, by definition, the onset of the regional stage can occur with at least one involved node that is clinically detectable. The close but not exact agreement is likely to be due to differences in modelling assumptions and data sets. Kimmel *et al.* made assumptions that the hazard for stage transitions is a function of volume, as opposed to a function of time as in our model; also, Kimmel *et al.* fit their model to a data set other than the SEER data.

Work by Koscielny *et al.*, in Reference [2], report on the biological onset of remote metastases, stating that 'the volume at which 50 per cent of the tumours have remote metastases, some that may be occult when the primary is detected is 23.6 ml (diameter = 3.56 cm).' Our approach does not allow us to estimate the same quantity, but it is reassuring that this estimated tumour volume at the biological onset of remote metastases is smaller than 5.5 cm which we estimate as the median tumour size for the onset of clinically detectable metastatic disease. The model assumptions and data underlying the work by Koscienly *et al.* differ from ours. Koscienly *et al.* relies on long-term survival data for patients who did not receive adjuvant treatment and exploits a linear relationship observed between the logarithm of the tumour volume and the probability of metastases.

Dr Shwartz, in References [7, 8], used a formalism for the assumption that stage transitions are proportional to the volume that is more general than ours. He assumes that the hazard rate for nodal involvement has three components: a constant, a term proportional to volume and a term proportional to the first derivative of the volume. This generalization increases the number of model parameters and does not allow a closed-form analytical expression for the likelihood function. When we applied Dr Shwartz's generalized model to our data, we found that the constant term slightly improved the data fit at the smaller tumour sizes (< 1 cm) but the impact of the derivative term was negligible.

Our model does not rely on data that is collected in the presence of screening. While a model that includes a screened population could be more informative, it would be more complex because it would require characterizing the detection function of the screening test and screening compliance. It would also need to account for complex effects of leadtime, lengthtime and overdiagnosis introduced by screening.

Stochastic modelling of the type that we present here is less explored than regression-based analyses of breast cancer data, yet can provide insights into unobservable properties of the

disease. Alternative stochastic models of the natural history of breast cancer may produce a similar or better fit to the data, but yield different predictions of unobservable events. Future modelling efforts should not only compare the data fit but also predictions of unobservable events.

8. CONCLUSION

We present a stochastic model of the natural history of cancer to quantify the relative rates of stage transitions, the tumour volume at stage transitions and the impact of the symptomatic detection function on the observed tumour size and stage distribution in an unscreened population. We computed maximum likelihood estimates of model parameters using data on tumour size and stage from breast cancer patients who were not undergoing screening mammography. Our results indicate that the rate of symptomatic detection is similar to the rate of transition from the local to regional stage and an order of magnitude larger than the rate of transition from the regional to distant stage. We demonstrate that, in the even absence of screening mammography, symptom-prompted detection alone has a large effect on reducing the occurrence of distant stage disease at initial diagnosis. Even though our analysis was limited to breast cancer, the formalism presented here may be applicable to other solid tumours where it is reasonable to assume that the disease progresses from local to regional to distant stages and that the hazard function for the time to symptomatic detection and stage transition is proportional to the tumour volume expressed as a function of time.

APPENDIX

The joint density of $\{D, I\}$ for the likelihood function in Section 5 was expressed with a continuous variable for tumour size. Here is it expressed with discrete tumour size bins. In particular, the probability of a clinically detected tumour falling in the kth tumour size bin, i.e. belonging to tumour sizes in the range $[d_k, d_{k+1}]$

$$\Pr(D \in (d_k, d_{k+1}), I = i) = \int_{d_k}^{d_{k+1}} \Pr(D = d, I = i)\, dd$$

For symptomatic detection in the local stage

$$\Pr(D \in (d_k, d_{k+1}), I = 0) = \frac{-\gamma}{\gamma + \eta} \times \left[\left(\frac{\beta}{\beta + (\gamma + \eta)(d - c_0)} \right)^{\alpha} \right] \Bigg|_{d=d_k}^{d_{k+1}}$$

For symptomatic detection in the regional stage

$$\Pr(D \in (d_k, d_{k+1}), I = 1)$$

$$= \frac{\eta}{\eta - \omega} \times \left[\frac{\gamma}{\gamma + \eta} \times \left(\frac{\beta}{\beta + (\gamma + \eta)(d - c_0)} \right)^{\alpha} - \frac{\gamma}{\gamma + \omega} \times \left(\frac{\beta}{\beta + (\gamma + \omega)(d - c_0)} \right)^{\alpha} \right] \Bigg|_{d=d_k}^{d_{k+1}}$$

For symptomatic detection in the distant stage

$$\begin{aligned}
&\Pr(D \in (d_k, d_{k+1}), I = 2) \\
&= \left[\left(\frac{\beta}{\beta + \gamma(d - c_0)} \right)^{\alpha} \right. \\
&\quad + \frac{\gamma}{\gamma + \eta} \times \left(\frac{\beta}{\beta + (\gamma + \eta)(d - c_0)} \right)^{\alpha} \\
&\quad - \frac{\eta}{\eta - \omega} \times \left(\frac{\gamma}{\gamma + \eta} \times \left(\frac{\beta}{\beta + (\gamma + \eta)(d - c_0)} \right)^{\alpha} \right. \\
&\quad \left. \left. - \frac{\gamma}{\gamma + \omega} \times \left(\frac{\beta}{\beta + (\gamma + \omega)(d - c_0)} \right)^{\alpha} \right) \right] \Bigg|_{d=d_k}^{d_{k+1}}
\end{aligned}$$

ACKNOWLEDGEMENTS

We gratefully acknowledge funding from the National Cancer Institute (R01 CA82904 and U01 CA097420). We are also thankful to Drs Frank Stockdale, Robert Tibshirani and Andrei Yakovlev for their valuable insights.

REFERENCES

1. Young JLJ *et al. SEER Summary Staging Manual—2000*: *Codes and Coding Instructions*. National Cancer Institute: Bethesda, MD, 2001.
2. Koscielny S, Tubiana M, Le MG, Valleron AJ, Mouriesse H, Contesso G, Sarrazin D. Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. *British Journal of Cancer* 1984; **49**(6):709–715.
3. Kimmel M, Flehinger BJ. Nonparametric estimation of the size-metastasis relationship in solid cancers. *Biometrics* 1991; **47**(3):987–1004.
4. Atkinson EN, Brown BW, Thompson JR. On estimating the growth function of tumours. *Mathematical Biosciences* 1983; **67**:145–166.
5. Bartoszynski R. A modeling approach to metastatic progression of cancer. In *Cancer Modeling*. Thompson JR, Brown PW (eds). Marcel Dekker: New York, 1987.
6. Bartoszynski R *et al.* Modeling cancer detection: tumour size as a source of information on unobservable stages of carcinogenesis. *Mathematical Biosciences* 2001; **171**(2):113–142.
7. Shwartz M. Validation and use of a mathematical model to estimate the benefits of screening younger women for breast cancer. *Cancer Detection and Prevention* 1981; **4**(1–4):595–601.
8. Shwartz M. An analysis of the benefits of serial screening for breast cancer based upon a mathematical model of the disease. *Cancer* 1978; **41**(4):1550–1564.
9. Duffy SW *et al.* Markov models of breast tumour progression: some age-specific results. *Journal of the National Cancer Institute Monographs* 1997; (22):93–97.
10. Duffy SW *et al.* Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase. *Statistics in Medicine* 1995; **14**(14):1531–1543.
11. Chen HH *et al.* Evaluation by Markov chain models of a non-randomised breast cancer screening programme in women aged under 50 years in Sweden. *Journal of Epidemiology and Community Health* 1998; **52**(5):329–335.
12. Brown BW *et al.* Lack of concordance of growth rates of primary and recurrent breast cancer. *Journal of National Cancer Institute* 1987; **78**(3):425–435.

13. Brown BW *et al.* Estimation of human tumour growth rate from distribution of tumour size at detection. *Journal of National Cancer Institute* 1984; **72**(1):31–38.
14. *Surveillance*, *Epidemiology and End Results* (*SEER*) *Program* (*www.seer.cancer.gov*) *Public Use Data* (*1973–2000*). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2003.
15. Cronin KA *et al.* Modeling the dissemination of mammography in the United States. *Cancer Causes Control* 2005; **16**(6):701–712.
16. *MATLAB ver. 5.3*. the Mathworks (www.mathworks.com).

