Cancer Causes Control (2009) 20:1559–1569
DOI 10.1007/s10552-009-9401-4

ORIGINAL PAPER

Modeling the transition of lung cancer from early to advanced stage

Maksim A. Pashkevich · Bronislava M. Sigal · Sylvia K. Plevritis

Received: 7 May 2008 / Accepted: 7 July 2009 / Published online: 23 July 2009

Abstract We present a stochastic parametric model of the natural history of lung cancer that predicts the primary tumor volume at the moment the disease transits from early to advanced stage. Our model also produces estimates for the probability of symptomatic detection as a function of tumor volume and clinical stage. We estimate model parameters by likelihood maximization using data from the Mayo Lung Project (MLP), which was a clinical trial that evaluated screening for lung cancer in the 1970s. Mayo Lung Project cancer cases reported in Stage III or greater, according to the 1979 AJCC staging for lung cancer, were considered advanced stage. Our estimator distinguishes between the cases detected because of clinical symptoms and cases detected by screening. For nonsmall cell lung cancer cases detected in MLP, we estimate that the median primary tumor diameter at the onset of advanced stage disease was 4.1 cm. In addition, we estimate that the rate of patients symptomatically detected with their disease increases as their primary tumor increases in size, and for patients with a primary tumor of a given size, the rate of symptomatic detection is 12.8 times greater among patients with advanced stage disease compared to patients with early stage disease.

Keywords Lung cancer · Mayo Lung Project · Metastatic progression · Natural history model · Screening

Electronic supplementary material The online version of this article (doi:10.1007/s10552-009-9401-4) contains supplementary material, which is available to authorized users.

M. A. Pashkevich · B. M. Sigal · S. K. Plevritis (✉)
Department of Radiology, Stanford University
School of Medicine, Stanford, CA 94305, USA
e-mail: sylvia.plevritis@stanford.edu

Introduction

Screening for cancer promises to reduce cancer mortality by detecting the disease before the onset of clinical symptoms, and more importantly before on the onset of lethal metastases. Screening has proven to be an effective strategy for reducing mortality from breast, cervical, and colorectal cancer, but it has not proven effective for lung cancer, which is the leading cancer-related death in the US and worldwide [1]. Numerous clinical trials have demonstrated that chest radiography is not effective in reducing lung cancer mortality [2], yet these findings have been debated for decades [3, 4]. Recently, interest in early detection of lung cancer has grown considerably as Computerized Tomography (CT) has demonstrated the ability to detect lung tumors in the millimeter-range that are missed on chest radiography [5]. Because the clinical significance of these small lesions is not clear [6], screening for lung cancer with CT is currently being evaluated in a randomized controlled trial with a mortality endpoint [7, 8]. One nonrandomized study demonstrated a large survival benefit due to CT [9] that would advocate for CT screening, but concerns about potential biases confounding the interpretation of this study have been raised [10].

At the center of the controversies regarding the potential efficacy of lung cancer screening are questions about the natural history of lung cancer; in particular: can the primary tumor be screen-detected before the onset of lethal metastases? Mathematical models of the natural history of cancer, in general, have been developed to identify the volume of the primary tumor at the onset of clinically detectable metastases. These natural history models are typically incorporated into decision analytic models that evaluate the health benefit and cost

of alternative screening strategies [11–16]. The natural history models often estimate the transition from early to advanced stage as function of tumor size. Because stage transition is not observable, a variety of methods are proposed to infer it from observable data. From data on patients who were symptomatically detected in early stage, one can infer that the size at transition must be greater than the size at detection. From data on patients who were detected in advanced stage, one can infer that the size at transition must be smaller than the size at detection. The majority of models that aim to infer primary tumor size at stage transition have been applied to breast cancer [11, 12, 17–24], but we have found that they can not be directly applied to lung cancer. When modeling the natural history of breast cancer, a common assumption made is that disease stage does not impact the detection of the disease due to clinical symptoms. In the case of lung cancer, this assumption could not hold because a large percentage of patients experience their initial symptoms when they have advanced disease stage.

We propose a parametric stochastic model of the natural history of lung cancer that assumes that symptomatic detection depends on both the primary tumor volume and the disease stage. We demonstrate how well different models of the symptomatic detection function fit clinical trial data from the Mayo Lung Project (MLP), which was a lung cancer screening clinical trial of chest radiography and sputum cytology. The model is fit to data on both MLP screen detected and nonscreen-detected cases.

In our earlier work related to modeling the natural history of breast cancer [23], we described a related model but it assumed that symptomatic detection was only a function of size (not size and stage as assumed here). Also, we fit our model of the natural history of breast cancer to data from symptomatically detected breast cancer patients who were not screened. In this analysis of lung cancer, we include cancer patients whose tumor is asymptomatic and was detected either by screening or incidentally. We find that including the asymptomatic cases is necessary because the symptomatic detected cases, which are often large and advance-staged, are not sufficient to inform the natural history at the smaller tumor sizes.

The work presented here is related to that presented by Kimmel and Flehinger [25], and special cases of their model were investigated by Xu and Prorok [26, 27]. Kimmel et al. inferred the primary tumor volume when the disease transits from early to advanced clinical stage and explored the dependence of lung cancer detection on the volume and disease stage. However, Kimmel et al. presented a nonparametric approach, using a model expressed in terms of primary tumor volume, whereas we propose a parametric approach which is based on assumptions of the growth function of the primary tumor and is expressed in the temporal domain. Also in the work by Kimmel et al., a single detection function is used for both symptomatic and screen detection, whereas we distinguish between screen and symptomatic detection and we produce estimates of the symptomatic detection function. In the discussion, we will further compare these approaches.

The remainder of the paper is organized as follows. The next section presents our model formulation and assumptions. We then identify the types of observations used for model identification. We next derive the expressions required for maximum likelihood estimation. We describe the application to a dataset of nonsmall cell lung cancer cases detected in the MLP. Finally, we present our results, discussion, and conclusions.

Model formulation

Our natural history model of cancer is based on assumptions about the growth of the primary tumor, the distribution of the tumor growth rate, the transition from early to advanced stages, and the symptomatic detection as a function of the primary tumor volume and disease stage. The model assumptions are summarized below.

Assumption 1 (On the tumor growth function). The volume of the primary tumor with the inverse growth rate r at the time t is given by $V(t \mid r) = v(t/r)$, which increases monotonically with t for fixed r. $V(0 \mid r) = c_0$ is the tumor volume at time $t_0 = 0$; the tumor prior to c_0 (presumably the avascular phase) is not modeled [28]. We assume that the primary tumor grows exponentially: $V(t|r) = v(t/r) = c_0 \cdot e^{t/r}$ and $c_0 = 8$ mm^3 (which is a cube of a tumor whose diameter is 2 mm). While the assumption of exponential growth may be overly simplistic, it is one of the oldest assumptions made about solid tumors [29] and continues to be used [11, 23, 28]. Alternative growth models have been proposed since exponential growth suggests that a tumor could experience a long preclinical period (>20 years) and grow to unrealistic sizes [30, 31]. We favor the simplicity of the exponential model and assume it holds from the moment the tumor diameter is 2 mm to the moment of symptomatic detection, thereby avoiding the undesirable properties of exponential growth at the smaller and larger sizes. Also, with exponential growth, the inverse growth rate is proportional to the tumor doubling time: $T_D = R \cdot \ln(2)$.

Assumption 2 (On the tumor growth rate). The inverse growth rate R is a random variable that has gamma probability distribution with the parameters α, β:

$$f_R(r;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}. \quad (1)$$

The gamma distribution is frequently used to describe the distribution of the inverse growth rates because it simplifies the likelihood function used for parameter estimation.

Assumption 3 (On tumor size at stage transition). The random variable for the primary tumor volume $M = V(T_M)$ at which the patient transitions from early to advanced stage is characterized by the cumulative distribution function $F_M(m;\theta)$ and the probability density function $f_M(m;\theta)$, where T_M is the random variable for the time of stage transition, and θ is a parameter vector. Because the time of transition is not observed, its distribution is nonidentifiable without making additional assumptions, as previously noted [25–27]. Implicit in our assumption is that the disease stage transition to the advanced stage is irreversible. Also, we implicitly assume that progression from early to advanced stage disease is size-driven. This assumption is also made in the nonparametric approach of [25], but in our parametric approach, which incorporates the tumor growth function, the implications of this assumption become more apparent. In particular, two patients with primary tumors of the same size have the same probability of being in advanced stage disease even if their primary tumors have different tumor volume doubling times. We assume the Weibull probability distribution [32] for the tumor size at stage transition M: $F_M(m;\omega_0,\omega_1) = 1 - e^{-\omega_0\cdot(x-c_0)^{\omega_1}}$, when formalizing special cases for the likelihood function.

Assumption 4 (On symptomatic detection). The probability distribution of the random variable for the time of symptomatic detection T_{SD} depends on whether the disease has progressed to the advanced stage and is defined as follows:

$$\mathbf{P}\{T_{SD} \in (t, t+dt] | T_{SD} > t,\ R = r,\ T_M = t_M\} = \begin{cases} h_0(V(t\,|\,r))\,dt + o(dt), & t < t_M \\ h_1(V(t\,|\,r))\,dt + o(dt), & t \geq t_M \end{cases}, \quad (2)$$

where $h_0(.)$, $h_1(.)$ are the hazard functions in early stage and advanced stage, respectively, and $o(.)$ is a Landau symbol; $o(dt)$ represents a term that is negligibly small when compared to dt. We assume that at the volume c_0 the patient is early staged, or equivalently $T_M > 0$.

We assume that the hazard function for symptomatic detection takes the following form:

$$h_0(v(t/r)) = \gamma_0 \cdot v(t/r),\ h_1(v(t/r)) = \gamma_1 \cdot v(t/r) + \delta. \quad (3)$$

and consider four special cases:

a. the $\{\gamma\}$-detection model: $\gamma_0 = \gamma_1 = \gamma, \delta = 0, h_0(v(t/r)) = h_1(v(t/r)) = \gamma \cdot v(t/r)$;
b. the $\{\gamma_0,\gamma_1\}$-detection model: $\delta = 0,\ h_0(v(t/r)) = \gamma_0 \cdot v(t/r),\ h_1(v(t/r)) = \gamma_1 \cdot v(t/r)$;
c. the $\{\gamma,\delta\}$-detection model: $\gamma_0 = \gamma_1 = \gamma,\ h_0(v(t/r)) = \gamma \cdot v(t/r),\ h_1(v(t/r)) = \gamma \cdot v(t/r) + \delta$;
d. the $\{\gamma_0,\gamma_1,\delta\}$-detection model: $h_0(v(t/r)) = \gamma_0 \cdot v(t/r),\ h_1(v(t/r)) = \gamma_1 \cdot v(t/r) + \delta$.

Data structure

We assume that the following two types of observations are available for parameter estimation:

O_1 Cases detected at the onset of symptoms: $\{x_{1i}, s_{1i}\}_{i=1}^n$.
O_2 Cases detected before the onset of symptoms: $\{x_{2i}, s_{2i}\}_{i=1}^m$.

Here x_{ki} is the volume of the i-th tumor, $s_{ki} \in \{0, 1\}$ is an indicator function equal to one when the disease is in the advance stage at detection and zero otherwise, and k specifies the observation type ($k = 1, 2$).

The first type of observation (O_1) arises when a patient seeks medical attention due to symptoms. The detection mechanism for these symptomatically detected cases is described by Assumption A_4.

The second type of observation (O_2) arises when the primary tumor is detected either as a result of screening or on a clinical examination that was indicated for reasons other than lung cancer risk. In the case of lung cancer, O_2-type observations can arise from screening procedures, such as chest X-ray, CT scan, or sputum cytology, with each of these procedures having significantly different detection functions. Because the observations of the second type were detected before the onset of symptoms, it follows from the model assumptions that on average they are smaller than tumors at symptomatic detection.

In the following section, we consider O_2-type observations as censored O_1-type observations, assuming independent censoring when formulating the likelihood function with data from the MLP. We found, by simulation analysis, that this assumption generates a negligible bias in parameter estimates given the screen detection characteristics in the MLP. However, the bias does increase as the sensitivity of the screening test increases. (See Supplement for details.)

Parameter estimation

To obtain the likelihood function for model parameter estimation, we first derive the expressions for the probabilities of observing the primary tumor of volume x before ($s = 0$) and after ($s = 1$) the metastatic spread for O_1- and O_2-type observations, as defined in the previous section. We denote $D = V(T_{SD})$ as the random variable for the

primary tumor volume at symptomatic detection. Initially, we express the likelihood function without specific assumptions about the tumor growth function $V(.)$, the hazards of the symptomatic detection $h_0(.)$, $h_1(.)$, and the probability distribution of primary tumor volume at transition from early to advanced stage $f_M(m;\theta)$.

The cumulative distribution function (CDF) of the size at detection D, given inverse growth rate $R = r$ and size at stage transition $M = m$, when the tumor is detected before stage transition ($x < m$) is:

$$\begin{aligned} F_{D|R,M}(x|r,m) &= \mathbf{P}\{D<x|R=r,\ M\ =\ m\} \\ &= 1-\exp\left(-r\int_0^{v^{-1}(x)} h_0(v(\tau))\,\mathrm{d}\tau\right) \\ &= 1-\mathrm{e}^{-r\cdot\gamma_0(x)}, \end{aligned} \tag{4}$$

where

$$\gamma_0(x) = \int_0^{v^{-1}(x)} h_0(v(\tau))\,\mathrm{d}\tau. \tag{5}$$

When the tumor is detected after stage transition ($x \geq m$), then this conditional CDF of D is:

$$\begin{aligned} &F_{D|R,M}(x|r,m) \\ &= 1-\exp\left(-r\int_0^{v^{-1}(m)} h_0(v(\tau))\,\mathrm{d}\tau - r\int_{v^{-1}(m)}^{v^{-1}(x)} h_1(v(\tau))\,\mathrm{d}t\right), \\ &= 1-\mathrm{e}^{-r\cdot\gamma_1(x,m)} \end{aligned} \tag{6}$$

where

$$\gamma_1(x,m) = \gamma_0(m) + \int_{v^{-1}(m)}^{v^{-1}(x)} h_1(v(\tau))\,\mathrm{d}\tau. \tag{7}$$

The probability density function (PDF) of D given $R = r$ and $M = m$, obtained by differentiating Eqs. 4 and 6, is:

$$f_{D|R,M}(x|r,\ m) = \begin{cases} r\,\gamma_0'(x)\cdot\mathrm{e}^{-r\cdot\gamma_0(x)}, & x<m, \\ r\,\gamma_1'(x,m)\cdot\mathrm{e}^{-r\cdot\gamma_1(x,m)}, & x\geq m. \end{cases} \tag{8}$$

where the derivatives $\gamma_0'(x)$, $\gamma_1'(x)$ are with respect to x. Integrating over r, based on Assumption A_2, yields the PDF of D given $M = m$, for $x < m$:

$$\begin{aligned} f_{D|M}(x|m) &= \int_0^{+\infty} f_{D|R,M}(x|r,\ m)\cdot f_R(r;\alpha,\beta)\,\mathrm{d}r \\ &= \gamma_0'(x)\cdot\frac{\alpha\cdot\beta^\alpha}{(\beta+\gamma_0(x))^{\alpha+1}}. \end{aligned} \tag{9}$$

Similarly, the PDF of D, given $M = m$, for $x \geq m$ is:

$$f_{D|M}(x|m) = \gamma_1'(x,m)\cdot\frac{\alpha\cdot\beta^\alpha}{(\beta+\gamma_1(x,m))^{\alpha+1}}. \tag{10}$$

The CDF of D, given $M = m$, for $x < m$ is:

$$\begin{aligned} F_{D|M}(x|m) &= \int_{c_0}^{x} \gamma_0'(y)\cdot\frac{\alpha\cdot\beta^\alpha}{(\beta+\gamma_0(y))^{\alpha+1}}\,\mathrm{d}y \\ &= 1-\beta^\alpha(\beta+\gamma_0(x))^{-\alpha}. \end{aligned} \tag{11}$$

and since $\gamma_0(m) = \gamma_1(m, m)$, the CDF of D, given $M = m$, for $x \geq m$ is:

$$\begin{aligned} F_{D|M}(x|m) &= \int_{c_0}^{m} \gamma_0'(y)\cdot\frac{\alpha\cdot\beta^\alpha}{(\beta+\gamma_0(y))^{\alpha+1}}\,\mathrm{d}y \\ &\quad+\int_{m}^{x} \gamma_1'(y,m)\cdot\frac{\alpha\cdot\beta^\alpha}{(\beta+\gamma_1(y,m))^{\alpha+1}}\,\mathrm{d}y \\ &=1-\beta^\alpha(\beta+\gamma_1(x,m))^{-\alpha}. \end{aligned} \tag{12}$$

We will use the following simplifying notation for the PDF and CDF of D, given M throughout the rest of the paper:

$$\begin{aligned} f_{D|M}(x|m) &= \begin{cases} f_D^0(x), & x<m, \\ f_{D|M}^1(x|m), & x\geq m, \end{cases} \\ F_{D|M}(x|m) &= \begin{cases} F_D^0(x), & x<m, \\ F_{D|M}^1(x|m), & x\geq m. \end{cases} \end{aligned} \tag{13}$$

Then the components of the likelihood function $p_{ks}(x)$ have the form:

$$p_{10}(x) = \mathbf{P}\{D\ =\ x,\ M\ >\ x\} = f_D^0(x)\cdot(1-F_M(x;\theta)), \tag{14}$$

$$\begin{aligned} p_{11}(x) &= \mathbf{P}\{D\ =\ x,\ M\leq x\} \\ &= \int_{c_0}^{x} f_{D|M}^1(x|m)\cdot f_M(m;\theta)\,\mathrm{d}m, \end{aligned} \tag{15}$$

$$\begin{aligned} p_{20}(x) &= \mathbf{P}\{D\ >\ x,\ M\ >\ x\} \\ &= (1-F_D^0(x))\cdot(1-F_M(x;\theta)), \end{aligned} \tag{16}$$

$$\begin{aligned} p_{21}(x) &= \mathbf{P}\{D\ >\ x,\ M\leq x\} \\ &= \int_{c_0}^{x} (1-F_{D|M}^1(x|m))\cdot f_M(m;\theta)\,\mathrm{d}m, \end{aligned} \tag{17}$$

and the likelihood function is:

$$\begin{aligned} L(\vartheta) &= \prod_{i=1}^{n}(p_{10}(x_{1i}))^{1-s_{1i}}(p_{11}(x_{1i}))^{s_{1i}} \\ &\quad\cdot\prod_{i=1}^{m}(p_{20}(x_{2i}))^{1-s_{2i}}(p_{21}(x_{2i}))^{s_{2i}}, \end{aligned} \tag{18}$$

where ϑ is a vector of all model parameters, which includes the inverse growth rate distribution parameters α, β, the parameter vector θ for the distribution of tumor size at stage transition, and parameters used to specify the hazard functions $h_0(.)$, $h_1(.)$ for the symptomatic detection.

For the four detection models presented in earlier section, expressions (5) and (7) can be integrated and functions $\gamma_0(.)$ and $\gamma_1(.)$ have closed form expressions. Likelihood functions for these models are given below, using exponential tumor growth (based on Assumption A_1) and the Weibull distribution for M (based on Assumption A_3).

Likelihood function for the $\{\gamma\}$*-detection model*: Here the hazard of symptomatic detection depends only on primary tumor volume and not on disease stage. The likelihood function (18) becomes

$$\begin{aligned} L(\gamma,\omega_0,\omega_1) &= \prod_{i=1}^{n}\left[f_D^0(x_{1i};\gamma)\cdot(1-F_M(x_{1i};\omega_0,\omega_1))\right]^{1-s_{1i}} \\ &\times \left[f_D^0(x_{1i};\gamma)\cdot F_M(x_{1i};\omega_0,\omega_1)\right]^{s_{1i}} \\ &\times \prod_{i=1}^{m}\left[(1-F_D^0(x_{2i};\gamma))\cdot(1-F_M(x_{2i};\omega_0,\omega_1))\right]^{1-s_{2i}} \\ &\times \left[(1-F_D^0(x_{2i};\gamma))\cdot F_M(x_{2i};\omega_0,\omega_1)\right]^{s_{2i}}, \end{aligned} \tag{19}$$

$$\begin{aligned} f_D^0(x;\gamma) &= \frac{\alpha\beta^\alpha\gamma}{(\beta+\gamma\cdot(x-c_0))^{\alpha+1}}, \\ F_D^0(x;\gamma) &= 1-\beta^\alpha(\beta+\gamma\cdot(x-c_0))^{-\alpha}. \end{aligned} \tag{20}$$

Likelihood for the $\{\gamma,\delta\}$*-,*$\{\gamma_0,\gamma_1\}$*- and* $\{\gamma_0,\gamma_1,\delta\}$*-detection models*: Terms involved in calculating the likelihood function $L(\gamma_0, \gamma_1, \delta, \omega_0, \omega_1)$ defined by (18) are

$$f_{D|M}^1(x|m) = \frac{\alpha\cdot\beta^\alpha\cdot(\gamma_1+\delta/x)}{[\beta+\gamma_0\cdot(m-c_0)+\gamma_1\cdot(x-m)+\delta\cdot\log(x/m)]^{\alpha+1}}, \tag{21}$$

$$F_{D|M}^1(x|m) = 1-\beta^\alpha\,[\beta+\gamma_0\cdot(m-c_0)+\gamma_1\cdot(x-m)+\delta\cdot\log(x/m)]^{-\alpha}. \tag{22}$$

and $f_D^0(x;\gamma_0)$, $F_D^0(x;\gamma_0)$ are given by the expression (20).

Maximization of the likelihood function was performed using the Nelder–Mead technique for each symptomatic detection model. Model selection was based on the likelihood ratio test.

Note that because our likelihood function (18) does not contain time-dependent components, parameters of the inverse growth rate distribution (α, β) can not be estimated. The likelihood function (18) does not change when the subset of model parameters $\{h_0(.),\ h_1(.),\ \beta\}$ is scaled by a nonzero constant. Hence, to reduce complexity, we proceed by setting $\alpha = \beta = 1$, that is, assuming the growth rate distribution is exponential. Under the more general constraint $\alpha = \beta$, we estimated the parameter α jointly with the other model parameters and found that when compared to $\alpha = \beta = 1$, there was no statistically significant improvement in model fit.

Lung cancer dataset

The natural history model, with each symptomatic detection function described earlier, was fit to data from the MLP. The MLP was a randomized controlled trial designed to evaluate the reduction in lung cancer mortality due to screening with chest X-ray and sputum cytology for a high risk population [33]. A total of 10,933 asymptomatic males, who were current or former heavy-smokers aged from 44 to 76 years old, were recruited into the study between the calendar years 1971 and 1977; 9,211 were qualified for the MLP randomized control trial after the prevalence screen (4,618 in the intervention arm and 4,593 in the control arm). The intervention group was offered chest X-ray and sputum cytology every 4 months for 6 years, and the control group received standard 1970 Mayo guidelines for annual chest X-ray film and sputum cytology tests with no further reminders during the study. The study was completed in July 1983, with 206 cases of lung cancer in the screening arm and 160 cases in the control arm; a considerable number of cases were detected on chest X-ray outside of the study. No statistically significant difference in lung cancer mortality was found between the study and control groups, but alternative interpretations of the MLP results are still being discussed [4, 34].

Table 1 summarizes the MLP data on nonsmall cell lung cancer detected with and without the clinically detectable metastases during the study due to symptoms or by chest X-ray; cases with missing sizes, in situ disease and small cell carcinomas were excluded. We define patients with advanced stage disease as patients reported in the stage III according to the AJCC staging for lung cancer in 1979, which was applied to the MLP medical records. These cases include tumors more extensive than T2, mediastinal lymph nodes involvement, and distant metastasis [35]. The tumor diameter was the pathologic size when available, otherwise the radiographic size. There were 65 cases detected due to symptoms (13 without metastases and 52 with metastases), with the size of the tumor ranging from 8 to 130 mm (mean 44.2 mm). There were 120 cases detected on chest X-ray (93 without the metastases and 27 with the metastases), with the size of the tumor ranging from 5 to 80 mm (mean 28 mm).

Table 1 Nonsmall cell lung cancer detected as early versus advanced stage disease in the Mayo Lung Project among patients with symptoms ("symptomatic") or without symptoms but who were tested positive by medical examination ("nonsymptomatic")

Tumor diameter at *symptomatic* detection (mm)	Number of tumors		Tumor diameter at *nonsymptomatic* detection (mm)	Number of tumors	
	Early stage	Advanced stage		Early stage	Advanced stage
8	1	0	5	0	1
10	1	0	8	5	0
15	1	3	9	1	0
18	1	0	10	4	1
20	0	1	11	1	0
22	0	1	12	2	1
23	0	4	14	2	0
25	2	3	15	9	1
26	0	1	16	1	1
27	0	1	17	3	0
30	1	1	18	6	1
32	0	1	20	9	2
33	0	1	23	1	0
35	1	7	24	2	1
40	1	3	25	14	3
42	0	1	26	1	0
45	0	4	27	1	0
46	1	0	28	2	0
50	1	4	30	6	3
53	0	1	32	1	0
55	0	1	34	2	0
60	0	3	35	1	0
65	0	1	40	10	2
67	0	1	45	2	2
70	1	3	50	2	0
76	0	1	55	1	2
80	1	1	58	1	0
90	0	1	60	3	3
110	0	1	65	0	1
115	0	1	75	0	1
130	0	1	80	0	1

Results

Based on the data in Table 1, the log-likelihood function values for the natural history model using the $\{\gamma\}$-, $\{\gamma,\delta\}$-, $\{\gamma_0,\gamma_1\}$- and $\{\gamma_0,\gamma_1,\delta\}$-detection models are given in Table 2; parameter estimates with their asymptotic 95% confidence intervals are also presented. The $\{\gamma_0,\gamma_1,\delta\}$-detection model produces the highest log-likelihood value; however, when compared to the $\{\gamma_0,\gamma_1\}$-detection model, the improvement due to the parameter δ is not statistically significant (χ^2-value is 0.02, likelihood ratio test p-value is 0.89), and the confidence interval for the parameter δ includes zero. Therefore, we exclude the $\{\gamma_0,\gamma_1,\delta\}$-model from further consideration and select the $\{\gamma_0,\gamma_1\}$-detection model as the model of choice based on its likelihood. Under the $\{\gamma_0,\gamma_1\}$-detection model, the symptomatic detection rate of lung cancer in patients with primary tumor of a given size is 12.8 times greater among patients with advanced stage disease compared to those with early stage disease.

Figure 1 compares the cumulative distribution function (CDF) for the tumor size at symptomatic detection under the $\{\gamma\}$-,$\{\gamma,\delta\}$-,$\{\gamma_0,\gamma_1\}$-detection models to each other and to the Kaplan–Meier (KM) estimate. To generate the KM estimate, we regard observations of the second kind (O_2) as independently censored observations with respect to the symptomatically detected cases (O_1). In other words, screen detection is regarded as a censoring event for the size at clinical detection, which is the endpoint for our KM estimate. Interestingly, all three detection models generate

Table 2 Maximum likelihood estimates of the model parameters, their asymptotic confidence intervals and the corresponding log-likelihood values

Model	Model parameter	Parameter estimate	95% Confidence interval	
			Lower bound	Upper bound
$\{\gamma\}$-model, log-likelihood = −983.62				
	γ	7.02×10^{-6}	4.62×10^{-6}	9.42×10^{-6}
	ω_0	1.11×10^{-2}	7.11×10^{-3}	1.50×10^{-2}
	ω_1	3.82×10^{-1}	3.56×10^{-1}	4.08×10^{-1}
$\{\gamma_0,\gamma_1\}$-model, log-likelihood = −965.48				
	γ_0	1.90×10^{-6}	6.87×10^{-7}	3.12×10^{-6}
	γ_1	2.43×10^{-5}	1.03×10^{-5}	3.84×10^{-5}
	ω_0	2.42×10^{-3}	1.64×10^{-3}	3.21×10^{-3}
	ω_1	5.07×10^{-1}	4.88×10^{-1}	5.25×10^{-1}
$\{\gamma,\delta\}$-model, log-likelihood = −975.39				
	γ	2.34×10^{-6}	8.45×10^{-7}	3.83×10^{-6}
	δ	6.95×10^{-1}	6.29×10^{-1}	7.61×10^{-1}
	ω_0	3.20×10^{-4}	8.37×10^{-5}	5.56×10^{-4}
	ω_1	6.92×10^{-1}	6.24×10^{-1}	7.60×10^{-1}
$\{\gamma_0,\gamma_1,\delta\}$-model, log-likelihood = −965.47				
	γ_0	1.90×10^{-6}	7.04×10^{-7}	3.10×10^{-6}
	γ_1	2.41×10^{-5}	1.08×10^{-5}	3.74×10^{-5}
	δ	2.55×10^{-6}	0	1.41×10^{-1}
	ω_0	2.56×10^{-3}	1.96×10^{-3}	3.17×10^{-3}
	ω_1	5.02×10^{-1}	4.81×10^{-1}	5.22×10^{-1}

The estimates presented are for the cube of tumor diameter; $\alpha = \beta = 1$

results that lie within the 95% confidence interval of the KM estimate (Fig. 1), yet the models produce different inferences on the natural history of the disease (Figs. 2, 3). Figure 2 illustrates the cumulative distribution function (CDF) of primary tumor size at which the disease transits from early to advanced stage. As shown in Fig. 2, the CDF of the primary tumor size at stage transition under the three different detection models differ significantly at the smaller diameters (<40 mm), which would correspond to tumor sizes targeted by screening. For example, when a tumor reaches 2 cm, the probability of clinically detectable metastases is approximately 30, 20, and 15% for the $\{\gamma\}$-, $\{\gamma_0,\gamma_1\}$-, and $\{\gamma,\delta\}$-detection models, respectively. The $\{\gamma\}$-detection model, which does not depend on stage, generates the highest probability of advanced stage disease for tumors less than 4 cm, and the $\{\gamma,\delta\}$-detection model generates the highest probability of advanced stage disease for tumors greater than 5 cm. The three models are in greater agreement for tumor sizes that are more typical for symptomatic detection (4.0–5.0 cm). For the detection model selected based on its likelihood, namely the $\{\gamma_0,\gamma_1\}$-detection model, the probability that nonsmall cell lung cancer has clinically detectable metastases is 25% at the primary tumor diameter of 2.3 cm, 50% at 4.1 cm, 75% at 6.5 cm, and 90% at 9.1 cm.

Figure 3 illustrates the probability that a patient with lung cancer has advanced stage disease if the patient is symptomatically detected with a primary tumor at a specified size, in the absence of screening. The curves in Figs. 2

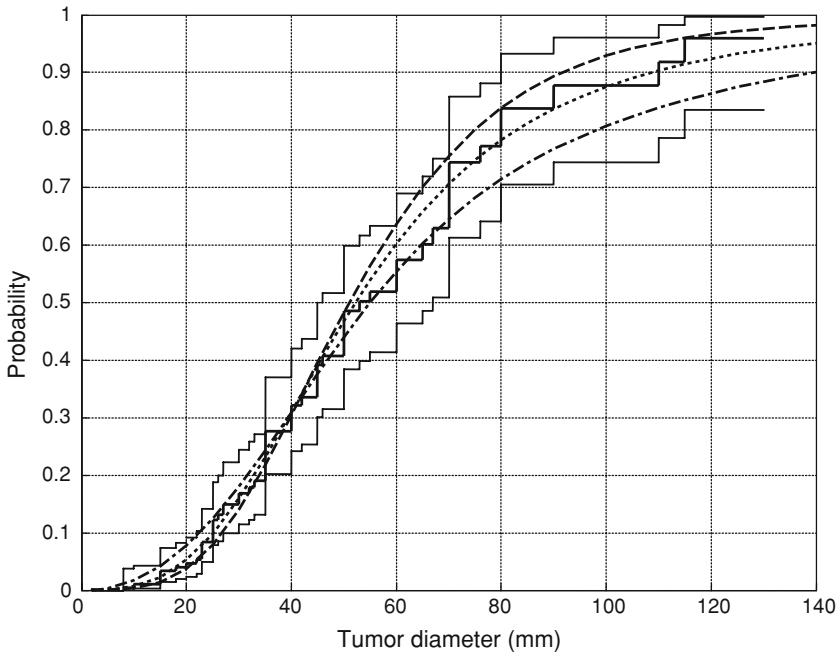

Fig. 1 Cumulative distribution functions for primary tumor size of nonsmall cell lung cancer at symptomatic detection: Kaplan–Meier estimator ("—", grey) and the corresponding 95% confidence intervals ("—") versus $\{\gamma\}$-model ("····"), $\{\gamma_0,\gamma_1\}$-model ("- - -") and $\{\gamma,\delta\}$-model ("-·-·-")

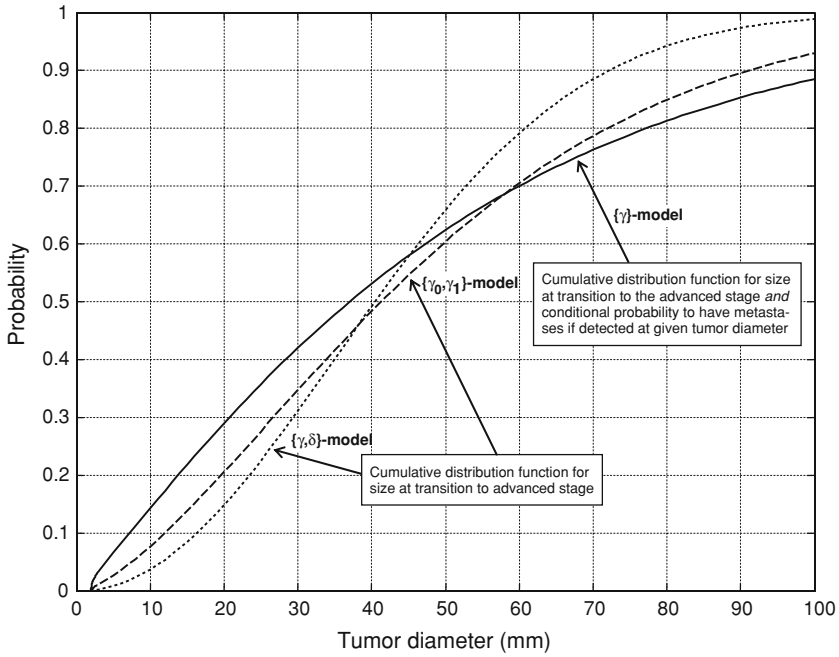

Fig. 2 Cumulative distribution function of the primary tumor size at transition from early to advanced stage; legend: "—" $\{\gamma\}$-model; "- - -" $\{\gamma_0,\gamma_1\}$-model; "····" $\{\gamma,\delta\}$-model

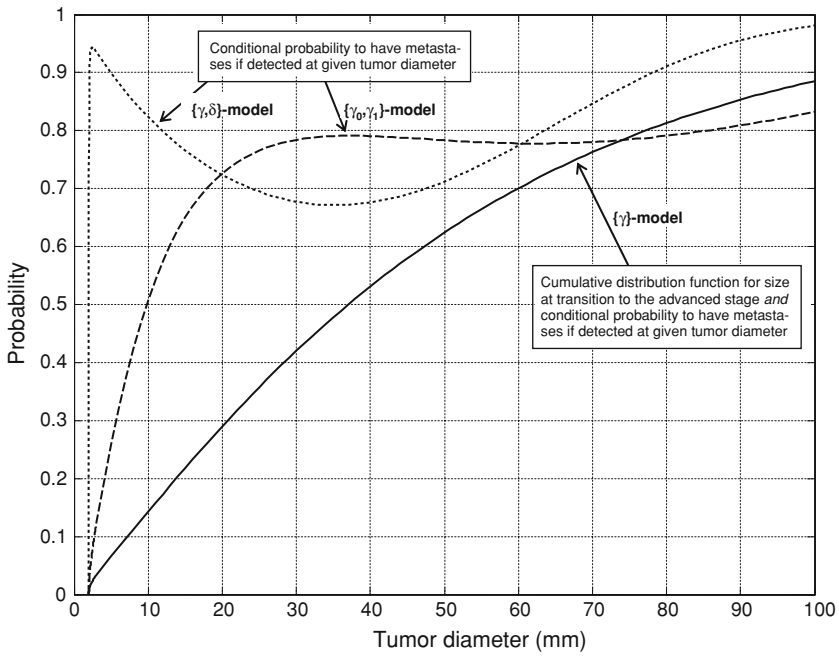

Fig. 3 Probability of advanced stage disease given symptomatic detection at a specified primary tumor size; legend: "—" $\{\gamma\}$-model; "- - -" $\{\gamma_0,\gamma_1\}$-model; "····" $\{\gamma,\delta\}$-model

and 3 are equivalent for the $\{\gamma\}$-detection model but differ for the $\{\gamma_0,\gamma_1\}$- and $\{\gamma,\delta\}$-detection models. Compared to the $\{\gamma\}$-detection model, the $\{\gamma_0,\gamma_1\}$- and $\{\gamma,\delta\}$-detection models produce a higher probability of clinically detectable metastases when patients are symptomatically detected with a small primary tumor (Fig. 3). Keep in mind that the shape of this distribution at the small sizes is not readily observable, since the probability of symptomatic detection at the small sizes is low (e.g., for a tumor diameter of 10 mm or less, the probability of symptomatic detection is less than 2% for all models). The $\{\gamma,\delta\}$-detection model, which has the highest probability of clinically detectable metastases at small sizes, demonstrates an abrupt jump at the small sizes because it does not allow the presence of advanced stage disease at the initial tumor volume c_0. It converges to the cumulative density function of size at metastatic spread because the effect of δ decreases as the tumor volume increases. For large tumor volumes, the dependence of the detection function on stage is negligible for the $\{\gamma,\delta\}$-model, but for $\{\gamma_0,\gamma_1\}$-model, the hazard of

the symptomatic detection is higher for the tumors with the clinically detectable metastases at all sizes.

Discussion

Our work demonstrates that inferences about the natural history of lung cancer are influenced by assumptions about symptomatic detection. The $\{\gamma\}$-detection model, which assumes that stage does not contribute to symptomatic detection, produces the higher probability of advanced stage disease at the smaller sizes and predicts that the transition from early to advanced stage disease occurs at smaller tumor sizes. However, this model produces the worst data fit to MLP. The data fit to MLP improves by allowing stage to contribute to the symptomatic detection function, as in the $\{\gamma_0,\gamma_1\}$- and $\{\gamma,\delta\}$-detection models, which predict that the disease transits to advanced stages at larger sizes. For example, these models predict that lung cancer patients who are symptomatically detected with a large primary tumor (typically >4 cm) and advanced stage would most likely be diagnosed with early stage disease if their primary tumors were detected at smaller sizes (<2 cm).

Our motivation for developing a natural history model of lung cancer is to simulate the impact of lung cancer screening on stage shift. We embedded the $\{\gamma\}$-, $\{\gamma_0,\gamma_1\}$-, $\{\gamma,\delta\}$-, and $\{\gamma_0,\gamma_1,\delta\}$-detection models into a simulation model of MLP and compared the predicted and observed cancer stage and tumor size distribution for X-ray detected cases. The $\{\gamma_0,\gamma_1\}$-detection model, which was selected earlier based on the likelihood ratio test, also closely estimates the probability of advanced staged disease in MLP when screening is explicitly simulated (results not shown).

Compared to the results reported by Kimmel and Flehinger [25], we estimate a higher probability of clinically detectable metastases when nonsmall cell lung cancer reaches a specified size. In their work, the probability that clinically detectable metastases are present when a tumor reaches the diameter of 4 and 6 cm is approximately 30 and 50% for adenocarcinoma and 10 and 40% for squamous cell carcinoma, which are not consistent with Fig. 2. One reason for the discrepancies is the differences in datasets used. When we applied the nonparametric approach proposed by Kimmel and Flehinger [25, as stated on page 997] to the MLP dataset, we estimate that transition probabilities for nonsmall cell lung cancer with diameters of 4 and 6 cm are 48 and 76%, respectively, which are close to the estimates in Fig. 2. In contrast to the work by Kimmel and Flehinger [25], we also estimate symptomatic detection, in the absence of screening, as shown in Fig. 3. Moreover, we systematically explore the impact of the symptomatic detection function on estimates of the tumor size at stage transition. Knowledge of the symptomatic detection function is critical when simulating alternative lung cancer screening programs, since detection by symptoms and screening must be modeled as separate and competing events.

In addition, we note differences between our work and that by Xu and Prorok [26, 27]. Xu and Prorok [26, 27] further developed [25] by proposing additional assumptions to make the model identifiable in the general case. In [26], Xu and Prorok assume that the odds ratio for detecting a tumor at given size before and after stage transition is constant for all tumor sizes, and known. In [27], they assume that the hazard for detecting a tumor after the metastatic spread is constant and known. In both papers, the authors compare their results with the ones presented in [25], and report similar behavior. In contrast, our approach does not rely on expert estimates of the model parameters (such as the defined odds ratio or the constant hazard for detection in advanced stage). That said, the limitations of our approach are related the underlying assumptions 1–4, which include exponential tumor growth, gamma distributed growth rate, Weibull distributed tumor size at stage transition, and symptomatic detection hazard as a linear function of tumor size.

Our findings can be valuable when contemplating the effectiveness of new screening tests in reducing lung cancer mortality, however, the potential effects of "stage migration" must be considered. Suppose a MLP-like cohort were screened and the majority of patients were screen-detected with a primary tumor around 10 mm, then roughly 90% of those patients would have early stage disease, as per our analysis. However, if the proportion of early stage disease were substantially lower than predicted by our analysis, it is possible that changes in the practice of staging the disease could have impacted the outcomes. With more sensitive staging technologies, the stage transition will likely occur at a smaller size than we estimated. To evaluate this possibility, a dataset similar to MLP would be used to make new inferences on the properties of stage transition that would need to reflect the new staging practice (and possibly new staging definition).

Suppose again that a MLP-like cohort were evaluated under a screening protocol and the detected tumors were around 10 mm in size. If the proportion of early staged disease were even higher than predicted by our analysis, some might argue that the screening test may be detecting a new class of lesion that would not have been symptomatically detected in MLP and classify these lesions as over-diagnosed cases [3]. A model-based analysis of the potential for over-diagnosis from sensitive screening tests would require analyzing nonrandomized screening trial data with a natural history model of lung cancer and an accurate characterization of the screening and staging tests. Such an analysis may show that even in the presence of

over-diagnosis, screening may have a net positive health benefit if lethal tumors are screen-detected at earlier, nonlethal, stages.

Conclusions

This work describes the natural history of lung cancer in terms of the relationship for size of the primary tumor, the occurrences of clinically detectable metastases, and stage of the lung cancer at symptomatic detection. We present a parametric stochastic model of the natural history of lung cancer that predicts the distribution of tumor size at stage transition when the symptomatic detection function depends on the tumor size and stage of the disease. The model is applied to lung cancer data from the MLP lung cancer screening trial to evaluate the structure and parameters of the hazard function for symptomatic detection, and to estimate the distribution of the primary tumor size as the disease transits from early to advanced stage. For nonsmall cell lung cancer in the MLP population, we find that the median diameter of the primary tumor at the onset of clinically detectable metastases is 4.1 cm. Also, the rate of symptomatic detection is proportional to the tumor volume, and for patients with primary tumors of a given size, it is 12.8 times greater among patients with advanced stage disease compared to those with early stage disease.

Acknowledgments We gratefully acknowledge funding by the National Cancer Institute under the research grant R01 CA105366. We also acknowledge valuable discussions about the clinical progression of lung cancer with Dr. Michael Gould.

References

1. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ (2005) Cancer statistics, 2005. CA Cancer J Clin 55:10–30. doi:10.3322/canjclin.55.1.10
2. Eddy DM (1989) Screening for lung cancer. Ann Intern Med 111: 232–237
3. Marcus PM, Bergstralh EJ, Zweig MH, Harris A, Offord KP, Fontana RS (2006) Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. J Natl Cancer Inst 98:748–756
4. Marcus PM, Bergstralh EJ, Fagerstrom RM, Williams DE, Fontana R, Taylor WF, Prorok PC (2000) Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up. J Natl Cancer Inst 92:1308–1316. doi:10.1093/jnci/92.16.1308
5. Henschke CI, McCauley DI, Yankelevitz DF, Naidich DP, McGuinness G, Miettinen OS, Libby DM, Pasmantier MW, Koizumi J, Altorki NK, Smith JP (1999) Early Lung Cancer Action Project: overall design and findings from baseline screening. Lancet 354:99–105. doi:10.1016/S0140-6736(99)06093-6
6. Black WC (2000) Overdiagnosis: an underrecognized cause of confusion and harm in cancer screening. J Natl Cancer Inst 92:1280–1282. doi:10.1093/jnci/92.16.1280
7. Black WC, Baron JA (2007) CT screening for lung cancer: spiraling into confusion? J Am Med Assoc 297:995–997. doi:10.1001/jama.297.9.995
8. National Cancer Institute Web site (2008) National lung screening trial. http://www.cancer.gov/nlst. Accessed 4 September 4 2008
9. Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, Miettinen OS (2006) Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 355:1763–1771. doi:10.1056/NEJMoa060476
10. Bach PB, Jett JR, Pastorino U, Tockman MS, Swensen SJ, Begg CB (2007) Computed tomography screening and lung cancer outcomes. J Am Med Assoc 297:953–961. doi:10.1001/jama.297.9.953
11. Shwartz M (1978) A mathematical model used to analyze breast cancer screening strategies. Oper Res 26:955–973. doi:10.1287/opre.26.6.937
12. Shwartz M (1981) Validation and use of a mathematical model to estimate the benefits of screening younger women for breast cancer. Cancer Detect Prev 4:595–601
13. Plevritis SK, Sigal BM, Pashkevich MA (2007) Mortality reduction due to x-ray screening: stochastic simulation of Mayo Lung Project. Manuscript in preparation
14. Plevritis SK, Kurian A, Sigal BM, Daniel B, Ikeda D, Stockdale F, Garber AM (2006) Cost-effectiveness of screening for breast cancer with magnetic resonance imaging in BRCA1/2 mutation carriers. JAMA 295:2374–2384. doi:10.1001/jama.295.20.2374
15. Plevritis SK, Sigal BM, Salzman P, Glynn P, Rosenberg J (2006) A stochastic simulation model of US breast cancer mortality trends from 1975 to 2000. J Natl Cancer Inst Monogr 36:86–95
16. Berry DA, Cronin KA, Plevritis SK, Fryback DG, Clark LC, Zelen M, Mandelblatt JS, Yakovlev AY, Habbema JDF, Feuer EJ (2005) Contributions of screening and adjuvant treatment to reduction in breast cancer mortality in the US from 1975 to 2000. N Engl J Med 353:1784–1792. doi:10.1056/NEJMoa050518
17. Koscielny S, Tubiana M, Le MG, Valleron AJ, Mouriesse H, Contesso G, Sarrazin D (1984) Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. Br J Cancer 49:709–715
18. Atkinson EN, Brown BW, Thompson JR (1983) On estimating the growth function of tumours. Math Biosci 67:145–166. doi:10.1016/0025-5564(83)90097-4
19. Bartoszynski R, Edler L, Hanin L, Kopp-Schneider A, Pavlova L, Tsodikov A, Zorin A, Yakovlev AY (2001) Modeling cancer detection: tumour size as a source of information on unobservable stages of carcinogenesis. Math Biosci 171:113–142. doi:10.1016/S0025-5564(01)00058-X
20. Duffy SW, Chen HH, Tabar L, Day NE (1995) 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. Stat Med 14:1531–1543. doi:10.1002/sim.4780141404
21. Duffy SW, Day NE, Tabar L, Chen HH, Smith TC (1997) Markov models of breast tumour progression: some age-specific results. J Natl Cancer Inst Monogr 22:93–97
22. Ghosh D (2006) Modelling tumour biology—progression relationships in screening trials. Stat Med 25:1872–1884. doi:10.1002/sim.2363
23. Plevritis SK, Salzman P, Sigal BM, Glynn PW (2007) A natural history model of stage progression applied to breast cancer. Stat Med 26:581–595. doi:10.1002/sim.2550
24. Hanin L, Rose J, Zaider M (2006) A stochastic model for the sizes of detectable metastases. J Theor Biol 243:407–417. doi:10.1016/j.jtbi.2006.07.005
25. Kimmel M, Flehinger BJ (1991) Nonparametric estimation of the size-metastasis relationship in solid cancers. Biometrics 47:987–1004. doi:10.2307/2532654

26. Xu JL, Prorok PC (1997) Nonparametric estimation of solid cancer size at metastasis and probability of presenting with metastasis at detection. Biometrics 53:579–591. doi:10.2307/2533959
27. Xu JL, Prorok PC (1998) Estimating a distribution function of the tumor size at metastasis. Biometrics 54:859–864. doi:10.2307/2533840
28. Bartoszynski R (1987) A modeling approach to metastatic progression of cancer. In: Thompson JR, Brown PW (eds) Cancer modeling. Marcel Dekker, New York
29. Collins VP, Loeffler RK, Tivey H (1956) Observations on growth rates of human tumors. Am J Roentgenol Radium Ther Nucl Med 76:988–1000
30. Geddes DM (1979) The natural history of lung cancer: a review based on rates of tumour growth. Br J Dis Chest 73:1–17
31. Sherratt JA, Chaplain MAJ (2001) A new mathematical model for avascular tumour growth. J Math Biol 43:291–312. doi:10.1007/s002850100088
32. Johnson NL, Kotz S, Balkrishan N (1994) Continuous univariate distributions. Wiley, New York
33. Fontana RS, Sanderson DR, Woolner LB, Taylor WF, Miller WE, Muhm JR, Bernatz PE, Payne WS, Pairolero PC, Bergstralh EJ (1991) Screening for lung cancer. A critique of the Mayo Lung Project. Cancer 67:1155–1164. doi:10.1002/1097-0142(19910215)67:4+<1155::AID-CNCR2820671509>3.0.CO;2-0
34. Henschke CI, Yankelevitz DF, Kostis WJ (2003) CT screening for lung cancer. Semin Ultrasound CT MR 24:23–32. doi:10.1016/S0887-2171(03)90022-9
35. American Joint Committee for Cancer Staging and End Results Reporting (1979) Task force on lung cancer: staging of lung cancer. American Joint Committee for Cancer Staging and End Results Reporting, Chicago