OPEN ACCESS Freely available online

PLOS ONE

Modeling the Natural History and Detection of Lung Cancer Based on Smoking Behavior

Xing Chen[1,2]*, Millennia Foy[3], Marek Kimmel[4], Olga Y. Gorlova[2,5]

1 Department of Biomedical Engineering, Key Laboratory of Biomedical Engineering of Ministry of Education of China, Zhejiang University, Hangzhou, Zhejiang, China, **2** Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **3** Brown Foundation Institute of Molecular Medicine, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America, **4** Departments of Statistics and Bioengineering, Rice University, Houston, Texas, United States of America, **5** Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth College, Lebanon, New Hampshire, United States of America

Abstract

In this study, we developed a method for modeling the progression and detection of lung cancer based on the smoking behavior at an individual level. The model allows obtaining the characteristics of lung cancer in a population at the time of diagnosis. Lung cancer data from Surveillance, Epidemiology and End Results (SEER) database collected between 2004 and 2008 were used to fit the lung cancer progression and detection model. The fitted model combined with a smoking based carcinogenesis model was used to predict the distribution of age, gender, tumor size, disease stage and smoking status at diagnosis and the results were validated against independent data from the SEER database collected from 1988 to 1999. The model accurately predicted the gender distribution and median age of LC patients of diagnosis, and reasonably predicted the joint tumor size and disease stage distribution.

Citation: Chen X, Foy M, Kimmel M, Gorlova OY (2014) Modeling the Natural History and Detection of Lung Cancer Based on Smoking Behavior. PLoS ONE 9(4): e93430. doi:10.1371/journal.pone.0093430

Editor: Dominik Wodarz, University of California Irvine, United States of America

Received December 12, 2013; **Accepted** March 4, 2014; **Published** April 4, 2014

Funding: This study was funded by the following grants: CISNET 5 U01 CA097431 to M.K. and O.G., FAMRI Young Clinical Scientist Award and Prevent Cancer Foundation grants to O.G., National Institutes of Health (NIH) grants R01 CA149462 to OYG, R03CA1338885 and R03128025 to I.G., CA55769 to M.R.S., National Natural Science Foundation of China No. 81201166 to X.C., Research Fund for the Doctoral Program of Ministry of Education of China No. 20120101120165 to X.C., Zhejiang Key Science and Technology Innovation Team 2011R50018 to X.C., and CA016672 to MD Anderson. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: cnxingchen@gmail.com

Introduction

Lung cancer is one of the most deadly diseases worldwide, largely because most patients present with advanced-stage disease at the time of diagnosis [1,2]. Most patients have clinical stage III or IV disease when they first notice symptoms and seek medical attention, which results in a poor prognosis. Age, gender, smoking status, tumor size, and disease stage at the time of diagnosis are highly related to the prognosis of patients with lung cancer [3,4,5]. In this article, we use mathematical methods to disentangle the tumorigenesis and detection processes. The goal of this model is to trace the timeline of an individual from his/her birth to the time of lung cancer initiation, progression, detection, and death. Thus, we combined models of carcinogenesis (cancer development until the first malignant cell), tumor progression (growth and metastasis), and detection, to construct a framework for modeling lung cancer at an individual level. Using this framework, we could infer characteristics that cannot be observed in clinical practice, including age of the patient when the primary tumor and nodal and distant metastases are formed. We were also able to evaluate characteristics that can only be partially observed, such as the tumor growth rate, and closely reconstruct characteristics that can be observed clinically, notably, tumor size and disease stage at the time of diagnosis. This procedure may be useful for better understanding the formation of the current lung cancer patient population and characterization of the future lung cancer trends with changes in smoking behavior and detection methods.

Materials and Methods

1. Lung cancer patients identified in the Surveillance, Epidemiology and End Results (SEER) database

Age, sex, disease stage, and tumor size for lung cancer patients who were diagnosed between 1973 and 2008 are available in the SEER database [6]. For patients diagnosed from 1988 to 1999, we used information on tumor size, with staging determined according to SEER extent of disease codes, which categorize tumors as localized, regional, and distant. For patients diagnosed from 2004 to 2008, the staging system developed by the American Joint Committee on Cancer was used to obtain tumor size and tumor node metastasis (TNM) disease stage information. Tumor size was measured as the maximum diameter, and we calculated the volume according to an assumption that a tumor grows as a sphere. We re-categorized the tumors into groups from 0 to 20 cm (with 1-cm increments) according to their maximum diameter.

2. Carcinogenesis modeling

For the carcinogenesis model, we used the two-stage clonal expansion (TSCE) model developed by Moolgavkar and Venzon [7] to calculate the age of the patient at tumor initiation. This model leads to an explicit formula for the distribution of the total

duration, T, of the first two stages in the carcinogenesis process (the transitions from normal to initiated cell and initiated to malignant cell), which encompasses the time from the birth of an individual to the onset of malignancy [8]. A smoking-based modification of the TSCE model relating smoking intensity measured in packs per day (ppd) to the parameters of the TSCE model through response functions (with the parameters v_0, α_0, γ_0, a_1, and a_2, as listed in File S1) was chosen. Smoking duration was incorporated to produce a more specific Survival function of the age at tumor initiation for individual never, current and former smokers.[9,10,11].

The smoking history generator (SHG, version 5.2.1) [12]from the Cancer Intervention and Surveillance Modeling Network (CISNET), which produced smoking duration and intensity data for individuals, was incorporated into the smoking-based TSCE model by using the parameters listed in Table S1. Parameters for males and females were estimated using the data from Cancer Prevention Study I (CPS-I) and from Nurses' Health Study (NHS), respectively. The SHG was also used to generate an age of death due to causes other than lung cancer.

3. Tumor growth and metastasis modeling

For the tumor growth and metastasis models, we assumed that the hazard of tumor progression is based on the activity of the tumor cells, and tumor cells detach from the primary tumor and transfer to another part of the body, leading to metastases [13]

3.1. Assumptions. The following assumptions were made in modeling tumor growth and metastasis:

1) The primary tumor grows from a single cell, with an assumed volume of 1×10^{-9} cm^3 [14]. The growth rate λ, which is related to the tumor doubling time by the expression $DT=Ln2/\lambda$, is determined at the time of tumor initiation and is assumed to remain the same over time.
2) The growth rate follows a gamma distribution, with shape and scale parameters θ and K.
3) All metastases are derived from the primary tumor, which means cells detach from the primary tumor at the rate ξ and are transferred and deposited at the rate μ at a new metastasis site [13]. We don't consider secondary metastasis from existing metastasis.
4) The activity of the tumor cell is related to how fast the tumor grows and how easily the cells detach. Specifically, the faster the primary tumor grows, the easier it is for the cells to detach. We define the tumor cell's activity α, to which the growth rate λ is proportional, $\lambda=\varepsilon1\times\alpha$ (where $\varepsilon1$ is a constant). The cell-detachment rate, β, is also proportional to α, $\beta=\varepsilon2\times\alpha$ (where $\varepsilon2$ is another constant). Thus, $\beta=\frac{\varepsilon2}{\varepsilon1}\lambda$, where $\frac{\varepsilon2}{\varepsilon1}=\xi$ is a parameter representing the relationship between β and λ. If the tumor with volume S grows exponentially, $S=e^{\lambda t}$, the total number of detached cells before time τ_0 is $e^{\xi\lambda\tau_0}=S_0^{\xi}$; we assume $0<\xi<1$, the interpretation of which is that cells always detach from the primary tumor but not all tumor cells will detach.
5) The detached cells will be transferred and deposited at new locations. The aggregate rate of transfer and deposition is μ. μ could be a constant parameter or a functional parameter determined by a biological process, such as the rate of synthesis of proteins that help transfer the tumor cells across the blood vessel wall [13].
6) Metastases are defined as either nodal or distant. We assume a different rate μ (μ_n and μ_m) for each type of metastasis, μ_n for nodal and μ_m for distant metastasis. We also assume that the detached cells can move to nodal sites at least as easily as to distant sites ($\mu_n\geq\mu_m$).
7) The hazard function for metastasis (nodal or distant) is related to the number of tumor cells that have detached from the primary tumor and have been successfully transferred and deposited at nodal or distant locations. Assuming exponential growth, the hazard functions for nodal and distant metastases are $h_n=\mu_n\times S^{\xi}$ and $h_m=\mu_m\times S^{\xi}$, respectively. The cumulative distribution functions (c.d.f.) are defined below:

$$F_n(s)=1-e^{-\int_0^S h_n(s)ds}=1-e^{-\frac{\mu_n}{\xi+1}*S^{(\xi+1)}}$$

$$F_m(s)=1-e^{-\int_0^S h_m(s)ds}=1-e^{-\frac{\mu_m}{\xi+1}*S^{(\xi+1)}}$$

with the tail functions (or survival functions) $\bar{F}_n$, $\bar{F}_m$. If $\xi=0$, we assume no cells detach from the primary tumor, $\mu_n=0$ and $\mu_m=0$.

1) Assumption $\mu_n\geq\mu_m$ implies that $F_n(s)\geq F_m(s)$, where $F_n(s)$ and $F_m(s)$ are corresponding cumulative distribution functions defined above. Primary tumor sizes at the time of initiation for nodal and distant metastases, respectively, are denoted s_{ni} and s_{mi}.
2) We assume that the cell's activity changes after detachment from the tumor, transfer and deposition at the metastatic site. Cell at the nodal and distant metastatic sites grow three (3λ) or four (4λ) times faster than the primary tumor [13,15,16], correspondingly.

The primary tumor size is calculated using the tumor growth model by giving the growing time t, with a constant growth rate λ. Thus, we rewrite $F_n(s),F_m(s)$ to $F_n(t,\lambda)=\int_0^t f_n(\tau,\lambda)d\tau$ and $F_m(t,\lambda)=\int_0^t f_m(\tau,\lambda)d\tau$, where $f_n(\tau,\lambda)$ and $f_m(\tau,\lambda)$ are the probability density functions (p.d.f.) of time that nodal and distant metastases happened in a group of patients with the same tumor growth rate λ. Then,

$$f_n(\tau)=\int f_n(\tau,\lambda)*\gamma(\lambda|k,\theta)d\lambda\ ;\ f_m(\tau)$$
$$=\int f_m(\tau,\lambda)*\gamma(\lambda|k,\theta)d\lambda$$

where f_n and f_m are the p.d.f. of time that nodal and distant metastases occurred in patients with tumor growth rate having a Gamma distribution, and γ is the Gamma distribution function with parameters k and θ. Then,

$$f_{n_1m_0}(\tau)=\int(f_n(\tau,\lambda)-f_m(\tau,\lambda))*\gamma(\lambda|k,\theta)d\lambda$$

where $f_{n_1m_0}(\tau)$ is the p.d.f. of time for patients who had only nodal metastases (without distant metastases) in the whole time period.

4. Cancer detection modeling

To model cancer detection, we introduced a competing process of detecting the disease through the primary tumor or nodal or distant metastases, adapting the framework developed by Kimmel and Flehinger. [17]

4.1. Assumptions. The following assumptions were made in modeling cancer detection:

1) The detection of cancer is based on the detection method used. The hazard of detection has a linear relationship with tumor size. It also depends on the reasons (e.g., symptoms or results of a screening test) that prompted the patient to seek medical attention. For details, see equations (*) for h_{Dp}, h_{Dn}, and h_{Dm} further on.
2) The detection of cancer is considered a competing process of detecting the primary tumor or nodal or distant metastases. The specific, mode-dependent hazard functions for the detection through the primary tumor and nodal and distant metastases are denoted as h_{Dp}, h_{Dn}, and h_{Dm} respectively (*). Then the c.d.f. of the detection of primary tumors and nodal and distant metastases are denoted as $D_p(s)$, $D_n(s)$, and $D_m(s)$, (**) with tail functions $\bar{D}_p(s)$, $\bar{D}_n(s)$,$\bar{D}_m(s)$, respectively (for details see equations (**)).
3) We also define primary tumor-size dependent hazard functions for detection (irrespective of the mode of detection, whether through the primary tumor, nodal metastases, or distant metastases), $z_{00}(s)$, $z_{10}(s)$, $z_{01}(s)$, and $z_{11}(s)$, where $z_{00}(s)$ is the hazard function for detecting, at size s, a cancer with no detectable metastases; $z_{10}(s)$ is the hazard of detecting, through whatever means, a cancer with the primary tumor of size s and with detectable nodal but not distant metastases; likewise, $z_{01}(s)$ is the hazard function for detecting, at size s, a cancer with detectable distant but not nodal metastases; and $z_{11}(s)$ is the hazard function for detecting, at size s, a cancer with detectable nodal and distant metastases. Associated with these hazard functions are the c.d.f. $Z_{nm}(s)$ before they reach size S, with tails $\bar{Z}_{nm}(s)$, where n, m = 0,1.
4) The observable variables in the study of size-dependent metastases are sizes S of the primary tumor at detection and the indicators N and M, where N, M = 1/0 if nodal or distant metastases are present/absent.

The relationship between assumptions (2) and (3) for the detection model is shown in the following functions.

$$Z_{00} = D_p$$

$$Z_{10} = (D_p \cup D_n) \backslash (D_p \cap D_n) = D_p \div D_n$$

$$Z_{01} = (D_p \cup D_m) \backslash (D_p \cap D_m) = D_p \div D_m$$

$$Z_{11} = (D_p \cup D_n \cup D_m \backslash D_p \cap D_n \backslash D_p \cap D_m \backslash D_m \cap D_n) \cup (D_p \cap D_n \cap D_m)$$

To give the explicit expressions for Z_{nm}, the detailed expressions of D_p,D_n and $D_m$are required. According to the assumptions of metastases model, we know that:

$$P_r\{s_0 \in S | N=0, M=0\} = 1 - F_n(s) - F_m(s) + F_n(s) \cap F_m(s) \triangleq 1 - F_n(s)$$

$$P_r\{s_0 \in S | N=1, M=0\} = F_n(s) - F_n(s) \cap F_m(s) \triangleq F_n(s) - F_m(s)$$

$$P_r\{s_0 \in S | N=0, M=1\} = F_m(s) - F_n(s) \cap F_m(s) \triangleq 0$$

$$P_r\{s_0 \in S | N=1, M=1\} = F_n(s) \cap F_m(s) \triangleq F_m(s)$$

Where $0 < S_0 \leq S$, is the size of the primary tumor that was not observed and Δ is considered the assumption that $F_n(s) \geq F_m(s)$. $P_r\{s_0 \in S | N, M\}$ represented the probability of a primary tumor with (1) or without (0) nodal (N) or distant (M) metastasis.

The joint density/ probability functions $p(s,n,m)$ of random variable S, N and M are presented below.

$$p(s,0,0) = Z'_{00}(1 - F_n)$$

$$p(s,1,0) = Z'_{10}(F_n - F_m)$$

$$p(s,0,1) = Z'_{01} \times 0 = 0$$

$$p(s,1,1) = Z'_{11} F_m$$

Where Z'_{nm} are probability density functions for detection.

The tumor-size dependent probability that nodal and distant metastases are present at diagnosis, $\Phi_n(s) = P_r\{N=1 | s=S\}$ and $\Phi_m(s) = P_r\{M=1 | s=S\}$, respectively, where

$$\Phi_n(s)=\frac{p(s,1,0)+p(s,1,1)}{p(s,0,0)+p(s,1,0)+p(s,0,1)+p(s,1,1)}$$

$$\Phi_m(s)=\frac{p(s,0,1)+p(s,1,1)}{p(s,0,0)+p(s,1,0)+p(s,0,1)+p(s,1,1)}$$

Substitute $p(s,n,m)$ with Z'_{nm}, F_n, and F_m we obtain:

$$F_n(s)=\frac{\Phi_m Z'_{00}Z'_{10}+(\Phi_n-\Phi_m)Z'_{00}Z'_{11}}{Z'_{11}Z'_{10}+\Phi_n(Z'_{00}-Z'_{10})Z'_{11}+\Phi_m(Z'_{10}-Z'_{11})Z'_{00}}$$

$$F_m(s)=\frac{\Phi_m Z'_{00}Z'_{10}}{Z'_{11}Z'_{10}+\Phi_n(Z'_{00}-Z'_{10})Z'_{11}+\Phi_m(Z'_{10}-Z'_{11})Z'_{00}}$$

5. Estimation

Methods provided by Kimmel and Flehinger [17] could be used to estimate Φ_n, Φ_m, and Z_{nm} non-parametrically. We can also estimate D_p,D_n,D_m,F_n, and F_m parametrically once the parametric tumor-growth and detection models are determined. Below we provide an example.

Assuming that a tumor grows exponentially with a growth rate λ, and the metastases model described above, we have

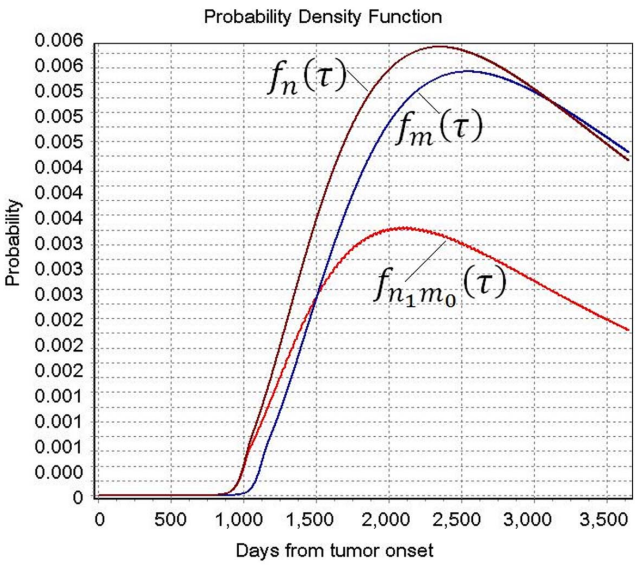


Figure 1. Probability density functions of nodal and distant metastases from the time of tumor onset, using the estimated parameters $\xi=0.01$, $\mu_n=8.05\times10^{-9}$, $\mu_m=2.78\times10^{-9}$, $K=3.80$ and $\theta=1.15$.

doi:10.1371/journal.pone.0093430.g001

$$F_n(s)=1-e^{-\int_0^s h_n(u)du}=1-e^{-\frac{\mu_n}{\xi+1}*S^{(\xi+1)}}$$

$$F_m(s)=1-e^{-\int_0^s h_m(u)du}=1-e^{-\frac{\mu_m}{\xi+1}*S^{(\xi+1)}}$$

Assuming that the hazard of tumor detection depends linearly on the size of the tumor, denoting the efficiency of the detection by tumor size as α and stage-dependent offset parameters as w_0, w_1 and w_2, we obtain

$$h_{D_p}=\alpha s_p+w_0;\ (*)\ h_{D_n}=\alpha s_n+w_1;\ (*)\ \ h_{D_m}=\alpha s_m+w_2;(*)$$

Correspondingly

$$D_p(s_p)=1-e^{-\int_0^{sp} h_{Dp}(s)ds}=1-e^{-\frac{\alpha}{2}s_p^2-w_0 s_p}\qquad(**)$$

$$D_n(s_n)=1-e^{-\int_0^{sn} h_{Dn}(s)ds}=1-e^{-\frac{\alpha}{2}s_n^2-w_1 s_n}\qquad(**)$$

$$D_m(s_m)=1-e^{-\int_0^{sm} h_{Dm}(s)ds}=1-e^{-\frac{\alpha}{2}s_m^2-w_2 s_m}\qquad(**)$$

are c.d.f. of detection by size s_p or s_n or $s_m$of the primary tumor and nodal and distant metastases. The primary tumor size when the nodal and distant metastases arise is denoted as s_{ni},s_{mi}. We can then rewrite $D_n(s)$, $D_m(s)$ as

$$D_n(s)=1-e^{-\frac{\alpha}{2}(s-s_{ni})^2-w_1(s-s_{ni})},\ s\geq s_{ni};\ D_n(s)=0,s<s_{ni}$$

$$D_m(s)=1-e^{-\frac{\alpha}{2}(s-s_{mi})^2-w_2(s-s_{mi})},\ s\geq s_{mi}\,;D_m(s)=0,s<s_{mi}$$

where s_{ni},s_{mi} are distributed with c.d.f. $F_n(\bullet)$,$F_m(\bullet)$, respectively, and $s_{ni}\leq s_{mi}$.

5.1. Simulation-based estimation. The tumor growth and metastasis model includes nine parameters (ξ, μ_n, μ_m, $\lambda(K,\theta)$, α, W_0, W_1, W_2). The joint likelihood function is difficult to maximize directly. However, the tumor-growth, metastasis, and detection models can be estimated separately, once multiple data points for tumor size and disease stage are available. Another method is to derive the least-squares function: $F(G(x|\xi,\mu_n,\mu_m,k,\theta,\alpha,W_0,W_1,W_2)-\hat{Y}(x))$, where g is the simulated joint distribution of tumor size and stage and $\hat{Y}$ is the observed joint distribution, based on these parameters, and apply

Table 1. The estimates of model parameters, with asymptotic confidence intervals.

Parameter	Description	Estimate	95% CI
ξ	Detachment rate	0.01	[0.008, 0.011]
μ_n	Transfer and deposition rate of cells to nodal metastases	8.05×10^{-9}	[7.80×10^{-9}, 8.21×10^{-9}]
μ_m	Transfer and deposition rate of cells to distant metastases	2.78×10^{-9}	[2.15×10^{-9}, 3.34×10^{-9}]
$*K$	Shape parameter of gamma distribution of tumor growth rate	3.80	[3.77, 3.82]
$*\theta$	Scale parameter of gamma distribution of tumor growth rate	1.15	[1.12, 1.19]
η	Efficiency of the detection by tumor size	1.0×10^{-4}	[1.0×10^{-5}, 1.0×10^{-3}]
W_0	Offset parameter for detection by N0M0 stage symptoms	0.065	[0.056, 0.075]
W_1	Offset parameter for detection by N1M0 stage symptoms	1.50×10^{3}	[1.30×10^{3}, 1.80×10^{3}]
W_2	Offset parameter for detection by M1 Stage symptoms	7.00×10^{4}	[6.50×10^{3}, 8.00×10^{5}]

*Assuming exponential tumor growth and the estimates of K and θ, the average tumor growth rate E(λ) corresponds to a doubling time of 55 to 60 days.
doi:10.1371/journal.pone.0093430.t001

the Nelder-Mead method [18,19,20] to achieve the best fitted parameters in the model. We used the second approach because we did not have the multiple tumor-size measurements for individuals to estimate the models separately.

$$F=\sum_{i=0}^{2}\sum_{j=0}^{20}\left(g_{ij}(\xi,\mu_n,\mu_m,k,\theta,\alpha,W_0,W_1,W_2)-\hat{y}_{ij}\right)^2$$

is the least square function where i is the stage status defined as local ($i=0$, no nodal or distant metastasis N0M0), nodal ($i=1$, nodal metastases but no distant metastases, N1M0), or distant ($i=2$, M1), j is the number of tumor size group; $g_{ij}()$ is the simulated percentage of lung cancer with tumor in the size range of j group and i stage among all detected lung cancer; $\hat{y}_{ij}$ is the observed percentage of lung cancer with tumor in the size range of j group and i stage among all detected lung cancer. Detailed simulation procedure was in File S1.

We estimated the nine parameters (ξ, μ_n, μ_m, $\lambda(K,\theta)$, η, W_0, W_1, W_2) using the TNM staging data in the SEER database from 2004 to 2008 for model fitting. Since SHG is using year 2000 as a cut-off point for the vital status observation, the joint distributions of tumor size and disease stage from 1995 to 1999 were chosen as the output of the simulation. The results were validated against independent data from the SEER database collected from 1988 to 1999. These years were chosen as closest possible to 2004–2008 periods.

To simulate the LC population, we firstly used the smoking history generator (SHG) to generate the underlying population. We assumed that the number of persons before year 1890 was zero and at the year of 1890 there were 2877000 new born babies (the number of live births in each year was shown in the Figure S1). We provided the year of birth (say 1890) and gender (half and half) to SHG as inputs and repeated the SHG for 287700 times. Then we got these persons' basic information, including the year of death (converted from the age of death, A_d, generated by SHG) and their smoking history information. We then applied our simulation strategy (described in the File S1 at section 2.1 simulation process) to get LC candidates and the information of their tumor progression. For the next year (say 1891), we added new born babies to the underlying population and removed the persons that were dead in the previous year (say 1890) from the population, whenever she or he was LCs or "normal" persons. Thus, we had underlying population, which would be approaching the real U.S. population (figure S2), and the LC candidate population, which were considered as an unperturbed (existing before detection) LC population. Assuming that no LC-related death occurred before detection, the yearly LC population would be achieved by applying the detection model to the unperturbed LC population.

Results

Figure 1 shows the probabilities of nodal metastases and distant metastases by the time from the tumor onset. The estimated parameters ξ, μ_n, μ_m, k and θ in Table 1, which gives the estimates of the model parameters, were used to draw $f_n(\tau)$, $f_m(\tau)$ and $f_{n_1m_0}(\tau)$. These probability density functions showed that the probability of nodal and distant metastasis began to fast increase at 2.5 years (about 900 days) and 3 years (about 1100 days) from the time of tumor onset, respectively. It reached the highest at 6.4 years (about 2350 days) and 6.8 years (about 2500 days) from the time of tumor onset.

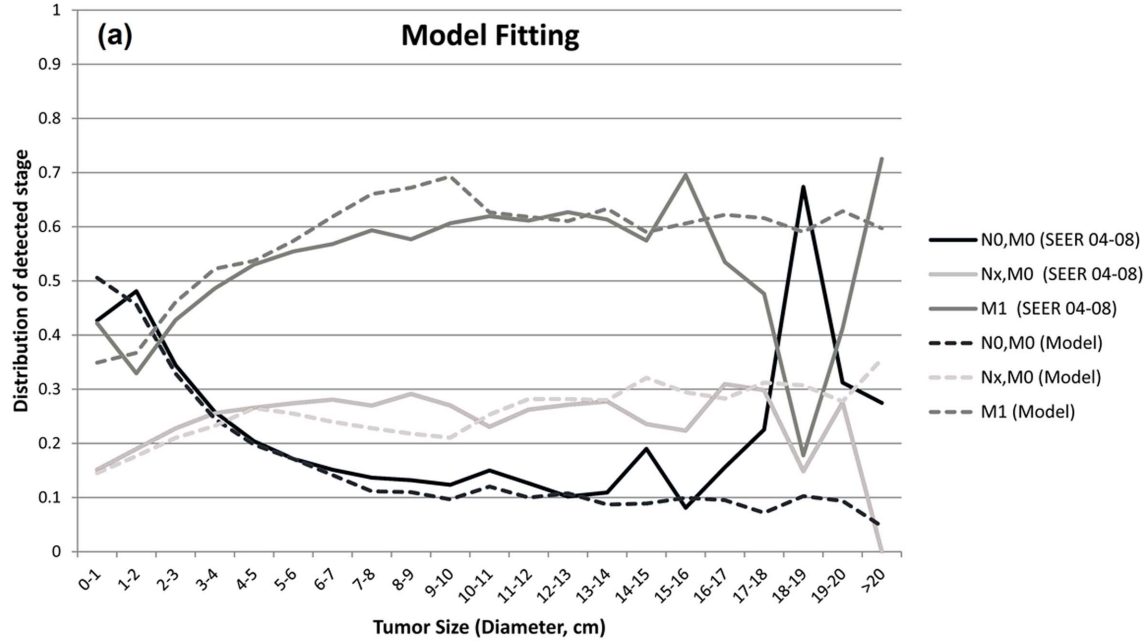
1. Model Fitting

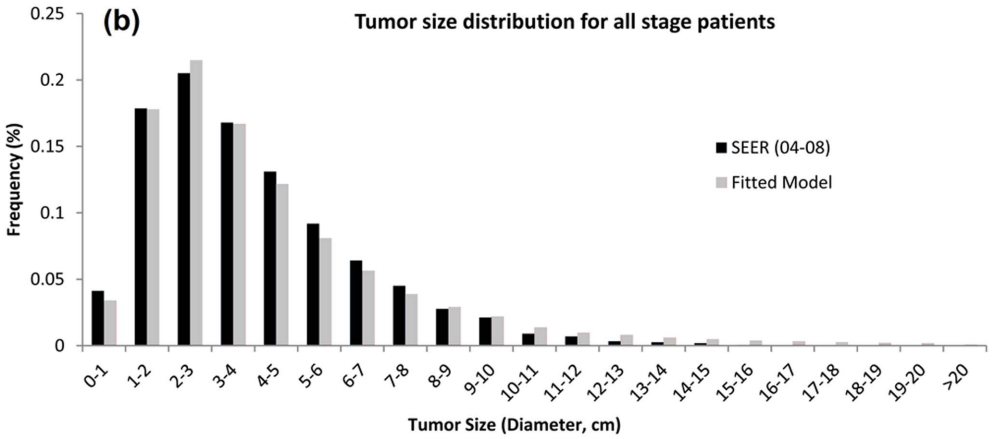
Figure 2 compares the characteristics of the population for the years 1995–1999 generated by the fitted model to the SEER data (2004–2008). For tumors smaller than 10 cm in diameter, the proportion of N0M0-stage disease (no nodal or distant metastases) more closely reproduces the SEER data (2004–2008). The proportions of NxM0- and M1-stage disease are not reproduced as accurately as the proportions of N0M0-stage disease, especially when the tumors are larger than 5 cm in diameter. For tumors smaller than 1 cm, the model predicted that about 50% and 35% would be staged as N0M0 and M1, respectively, whereas the actual percentages were 42% and 42% respectively.

2. Predicting Clinically Observable Characteristics

The fitted model was also validated by predicting the characteristics of lung cancer patient population in United States between 1988 and 1999. The model predicts both the gender distribution among LC patients and median age that are quite close to the 1988–1999 SEER data (Table 2).

Comparing the model prediction and the data, a smaller proportion of patients was diagnosed with localized disease than it was predicted (Figure 2c). One of the reasons may be the different staging definitions used. The predicted tumor size distributions were closer to the 2004–2008 SEER than to the 1988–1999 SEER data (Figure 3 (a–c)).


(a)
Model Fitting
Distribution of detected stage
Tumor Size (Diameter, cm)
N0,M0 (SEER 04-08)
Nx,M0 (SEER 04-08)
M1 (SEER 04-08)
N0,M0 (Model)
Nx,M0 (Model)
M1 (Model)


(b)
Tumor size distribution for all stage patients
Frequency (%)
Tumor Size (Diameter, cm)
SEER (04-08)
Fitted Model

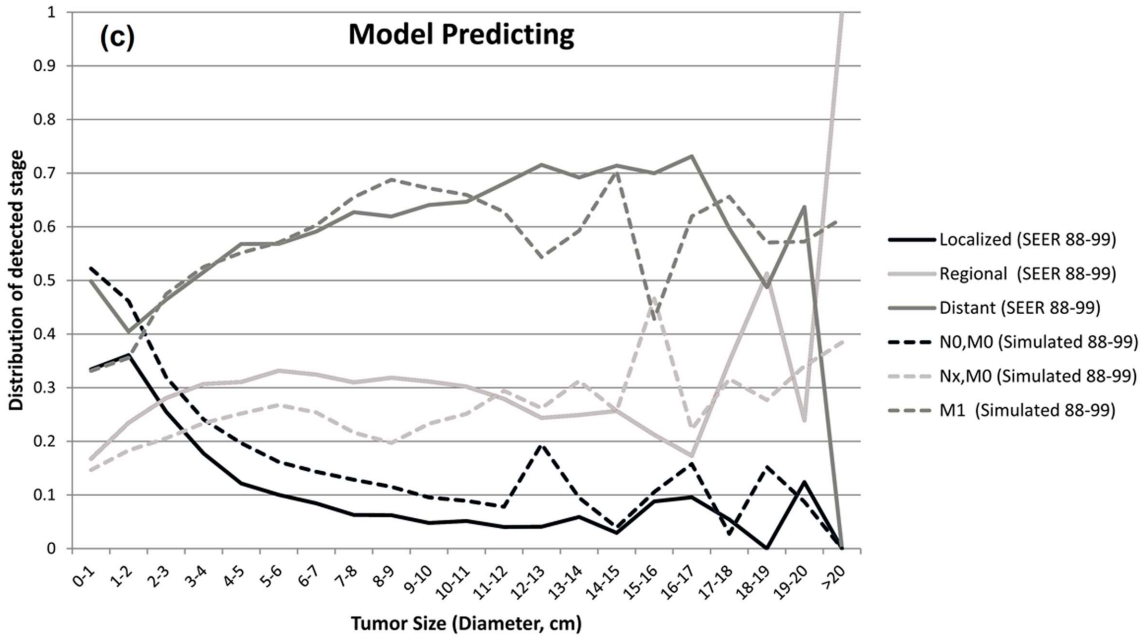

(c)
Model Predicting
Distribution of detected stage
Tumor Size (Diameter, cm)
Localized (SEER 88-99)
Regional (SEER 88-99)
Distant (SEER 88-99)
N0,M0 (Simulated 88-99)
Nx,M0 (Simulated 88-99)
M1 (Simulated 88-99)

Figure 2. Comparison of the model fit in the period of 1995–1999 to the data of SEER 2004–2008. (a) the stage distribution conditional on tumor size and (b) the tumor size distribution; (c) Comparison of the predictive model 1988–1999 and SEER 1988–1999, where the data is summarized as stage distribution conditional on tumor size.
doi:10.1371/journal.pone.0093430.g002

3. Predicting Clinically Unobservable Characteristics

Table 3 summarizes the predicted but not directly clinically observable characteristics of the detected tumors. The mean time from the tumor onset (when the first malignant cell appears) to nodal and distant metastases (when they are just formed and not yet observed) and diagnosis is about 4.77, 5.05, and 6.27 years, respectively. The average size of the primary tumors when nodal and distant metastases form was $0.09\ cm^3$ and $0.24\ cm^3$, respectively. The median age at the time of tumor onset is 63 and the average growth rate corresponds to the tumor-volume doubling time of about 60 days.

Table S2 shows the distribution of doubling time by tumor size and stage. In clinical practice, a primary tumor with distant metastases is more likely to be found with a smaller size than a primary with no or only nodal metastases. This leads to an observation that a primary tumor with distant metastases grows slower and remains smaller. This Table also demonstrates that faster growing tumors tend to be detected at larger sizes.

Discussion

The parameters estimated from the joint distribution of tumor size and stage in the SEER database from 2004 to 2008 (Table 1) were applied to generate a lung cancer patient population from 1988 to 1999 and validated by comparison of the results to the data from the SEER database from 1988 to 1999. The model accurately predicts the gender distribution and the median age of lung cancer patients, and approximates the joint tumor size and disease stage distribution. The accurate prediction of gender distribution and age at diagnosis for 1988–1999 is largely owing to the accuracy of the smoking-based TSCE model and SHG. Because smoking behavior has changed significantly over the recent decades, the output is sensitive to the year at detection, which is why we are not able to reconstruct gender and age at diagnosis in the SEER data from 2004 to 2008 as accurately. This model overestimates the proportion of patients with tumors larger than 10 cm in diameter and underestimates the proportion of patients with tumors between 4 and 9 cm. These discrepancies are more obvious for the distributions of the primary tumor size at stages NxM0 and M1 than at N0M0. The reason might be that the detection interval is fixed to 1 year in our model, whereas patients may visit a doctor more frequently when symptoms appear. For tumors smaller than 1 cm, the model underestimated the proportion of patients with distant metastases.

We also used the fitted model to predict disease characteristics that are difficult or impossible to observe in clinical practice. According to the estimates of k and θ in Table 1, the average tumor growth rate, λ, is about 4.4, which corresponds to a tumor-volume doubling time of approximately 55 to 65 days given the

Table 2. Comparison of the lung cancer patient population predicted by our model (1988 to 1999) with data from the SEER database (from 1988 to 1999 and from 2004 to 2008).

Characteristics	Prediction 1988 to 1999 (N = 1,434,024)	SEER 1988 to 1999 (N = 184,952)	SEER 2004 to 2008 (N = 84,422)
Sex, n (%)			
Male	833,754 (58.1)	108,205 (58.5)	44,228 (52.4)
Female	600,270 (41.9)	76,747 (41.5)	40,194 (47.6)
Age			
Mean(SD)	67.31(14.26)	68.06(10.98)	69.75 (11.57)
Median	69	69	71
*Stage, n (%)			
N0,M0	380,017 (26.5)	31,432 (17.0)(19.3)[£]	20,863 (24.7)(28.1)[£]
Nx,M0	321,221 (22.4)	46,934 (25.4)(28.8)[£]	17,889 (21.2)(24.1)[£]
M1	732,786 (51.1)	84,519 (45.7)(51.9)[£]	35,357 (41.9)(47.7)[£]
Missing stage	0 (0)	22,067 (11.9)	10,313 (12.2)
Tumor Size, cm (Diameter)			
Mean	4.57	4.30	4.21
Median	3.69	3.80	3.50
Std. deviation	3.21	3.03	3.11
Variance	10.33	9.20	9.65
Smoking Status, n (%)			
Never	185,885 (13.0)	**	-
Former	461,321 (32.2)	-	-
Current	786,818 (54.9)	-	-

*TNM staging being unavailable in SEER before 2004, we categorized tumors as localized, regional, and distant, for patients between 1988 and 1999.
**Smoking status is not reported in SEER.
[£]Excluding missing stage.
doi:10.1371/journal.pone.0093430.t002

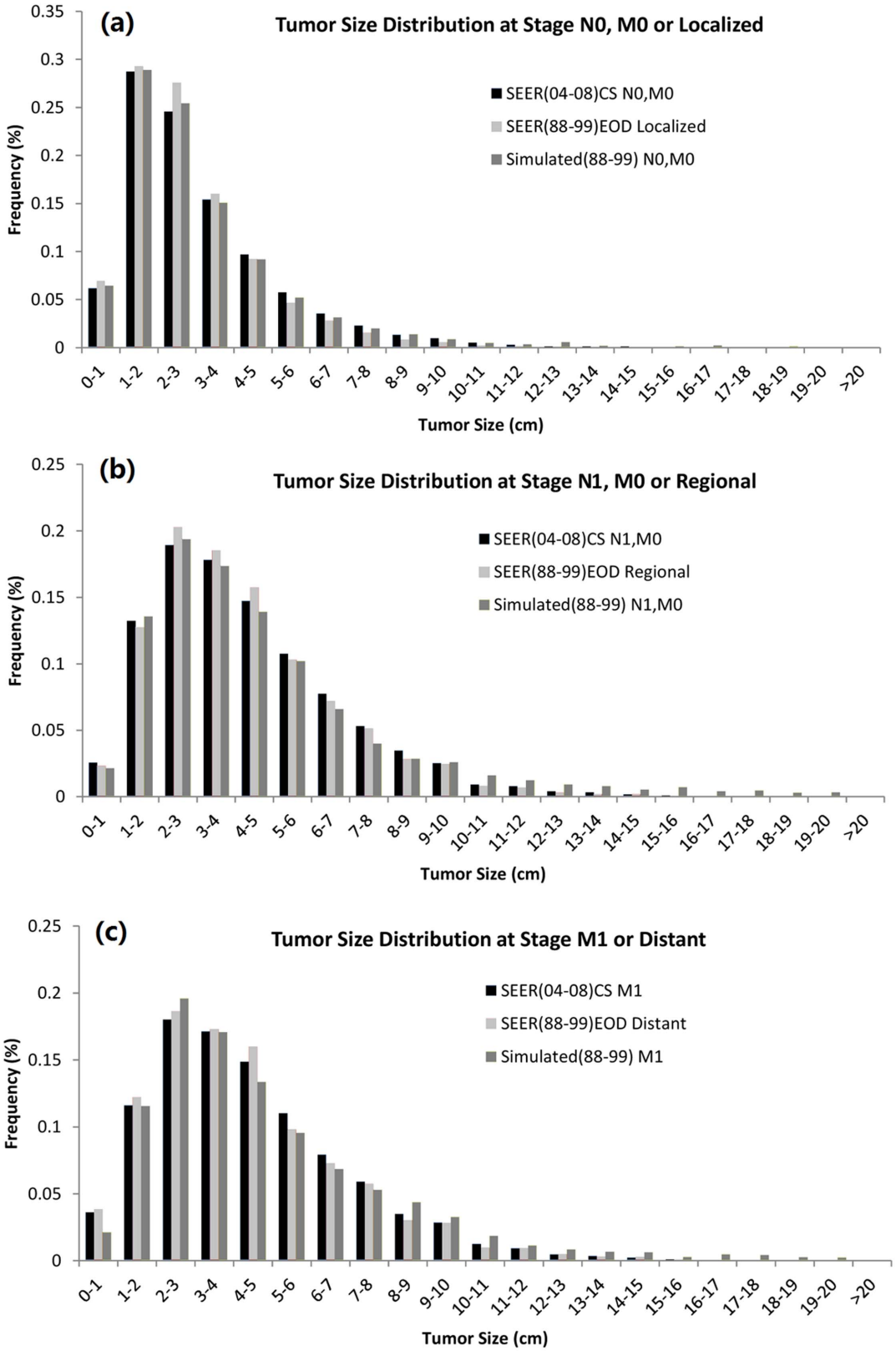


Figure 3. Tumor size distribution in predictive models, (a) Stage N0,M0 in SEER (2004–2008) and model (1988–1999), stage Localized by SEER standard in SEER (1988–1999), (b) Stage Nx,M0 (x≥1) in SEER (2004–2008) and model (1988–1999), stage Regional by SEER standard in SEER (1988–1999), (c) Stage M1 in SEER (2004–2008) and model (1988–1999), stage Distant by SEER standard in SEER (1988–1999).

doi:10.1371/journal.pone.0093430.g003

Table 3. Variables not directly observable for the detected tumors in the predicted lung cancer population (1988–1999).

Variables not directly observable	Mean (SD)	Median (IQR)
Time from the tumor onset, (years)		
To nodal metastasis	4.77 (2.70)	4.00 (3.00–6.00)
To distant metastasis	5.05 (2.86)	4.00 (3.00–6.00)
To diagnosis	6.27 (3.22)	5.00 (4.00–7.00)
Tumor volume (cm^3) at metastasis		
S_n	0.09 (8.4E-5)	0. 06 (0.03–0.13)
S_m	0.24(2.8E-4)	0.16 (0.07–0.33)
The linear tumor dimension (cm) at metastasis		
D_n	0.50 (1.7E-4)	0.49 (0.36–0.62)
D_m	0.68 (2.8E-4)	0.67 (0.50–0.85)
*Yearly Growth Rate λ by stage		
N0M0	6.17 (2.91)	6.30 (3.70–8.15)
N1M0	5.61 (1.68)	5.43 (4.45–6.71)
M1	3.79 (1.55)	3.54 (2.64–4.81)
*Doubling time by stage, (days)		
N0M0	57.15 (43.83)	40.15 (31.02–68.41)
N1M0	49.40 (16.77)	46.52 (37.73–56.88)
M1	80.01 (38.66)	71.48 (52.65–95.69)

IQR, interquartile range; SD, standard deviation;* Here is the yearly growth rate and doubling time of primary tumor.

doi:10.1371/journal.pone.0093430.t003

exponential tumor growth model. This growth rate is higher than what has been reported from screening studies [21,22,23], and thus the difference is not entirely unexpected [24]. Besides, introducing a time-dependent or size-dependent growth rate to the tumor growth model may improve the fit of the model to the data.

The hazards for detection once nodal or distant metastases are present are much larger than the hazards for detection when only the primary tumor is present (W_1, $W_2 >> W_0$), which reflects the reality of disease detection in clinical practice. The mean duration from tumor onset to detection was about 6 years in our model, which is consistent with other disease progression models [25]. Among the detected tumors, the average primary tumor size at the time of metastasis (nodal or distant) was less than 1 cm in diameter, which is considerably smaller than that implied by other predictions originating from screening data [17,26,27,28]. Assumptions regarding detection used in the models led to this difference. In our model, the chances of detecting lung cancer increase after nodal and distant metastases occur, and the competitive detection model allows for the detection of the metastasized tumor. This detection model does not require either of the two extreme assumptions used in the previous studies [8,9,17]: (1) that the probability to detect cancer is unchanged when metastases are present, or (2) cancers are detected immediately when metastatic spread occurs. To reduce the complexity of the disease stage progression model, we did not consider the possibility of a secondary spread of the disease from nodal metastases. This may be the reason that the model did not as accurately reproduce the proportion of nodal and distant metastases as it did the proportion of localized tumors for tumor sizes larger than 5 cm.

Our framework combines a carcinogenesis model with a model of the natural history of tumor growth and progression, and a detection model, to predict features of a lung cancer patient population. This modular structure allows testing of different detection strategies. One limitation of this model is that we were not able to construct the overall likelihood function for the model in the analytical form, and the Nelder-Mead estimation procedure used to optimize the least square fit is time consuming. Another limitation is that this framework largely depends on the smoking information generated by the SHG, which has to be updated before it can be used to accurately predict properties of future lung cancer patient populations. We did not perform simulation by histology, which is another limitation. Moreover, the model did not consider the difference between the lung cancer risks in CPSI/NHS and SEER, while the previously estimation of parameters in carcinogenesis model was directly used. We cannot recalibrate carcinogenesis model since no smoking information was recorded in SEER.

Conclusion

We proposed a model for predicting the natural disease progression and detection of lung cancer that relies on the following biologically and clinically reasonable assumptions: the hazard function of tumor progression is based on the activity of the tumor cells, which detach from the primary tumor and transfer to another part of the body, leading to metastases [13]. Thus, the metastasis process is related to the size of the primary tumor and the tumor growth rate (which is also related to the activity of the tumor cells). The detection of lung cancer in patients occurs as a result of competing detection of the primary tumor or nodal or distant metastasis. We used a TSCE model combined with the smoking history generator to reproduce the population with incipient tumors according to the yearly live birth number in the United States (Figure S1). We then applied our models of the tumor natural progression and its detection to re-create the lung cancer patient population at the time of diagnosis. Lung cancer data from SEER database collected between 2004 and 2008 were used to fit the lung cancer progression and detection model. The fitted model combined with a carcinogenesis model was used to reconstruct the distribution of age, gender, tumor size, and disease stage at diagnosis, and the results showed that the model accurately predicted gender and median age, and reasonably predicted the tumor size and disease stage distribution against independent data from the SEER database collected from 1988 to 1999. This model framework provides a platform for estimating the outcome of a strategy for the secondary prevention of lung cancer before it is applied in clinic.

Supporting Information

Figure S1 Yearly live birth number in US used in the simulation. For years in which the number of live births was missing (between 1890 and 1908) we used the average number of live births between 1909 and 1928 (2,877,000).
(TIF)

Figure S2 Comparison of U.S. population between simulated data and Census Bureau data; the simulated population deviates from the reality population after year 1984, since SHG could not generate new babies after 1984. However, we expect only minor if any effect of that on the LC population, as lung cancer is very rare in young individuals.
(TIF)

Table S1 Parameters of the response functions used in the TSCE model [10].
(DOCX)

Table S2 Doubling time by stages and tumor size for the simulated LC population.
(DOCX)

File S1 Supporting text.
(DOCX)

Acknowledgments

We thank the Department of Scientific Publications at M. D. Anderson for editorial assistance.

Author Contributions

Conceived and designed the experiments: XC OG MK. Performed the experiments: XC MF. Analyzed the data: XC. Contributed reagents/materials/analysis tools: XC OG MK. Wrote the paper: XC OG. Designed the software used in analysis: XC. Estimation of carcinogenesis model: MF.

References

1. Campbell M, Freeman JV (2007) Survival statistics. Br J Gen Pract 57: 410; author reply 410–411.
2. Jemal A, Thun MJ, Ries LA, Howe HL, Weir HK, et al. (2008) Annual report to the nation on the status of cancer, 1975–2005, featuring trends in lung cancer, tobacco use, and tobacco control. J Natl Cancer Inst 100: 1672–1694.
3. Gasperino J (2011) Gender is a risk factor for lung cancer. Med Hypotheses 76: 328–331.
4. Wang X, Zheng L, Zhang SY, Xie ZM, Yu H, et al. (2009) Risk factor analysis of mediastinal lymph node metastasis in non-small cell lung cancer patients and the strategy of mediastinoscopy prior to surgery. Zhonghua Zhong Liu Za Zhi 31: 456–459.
5. Boffetta P, Jayaprakash V, Yang P, Asomaning K, Muscat JE, et al. (2011) Tobacco smoking as a risk factor of bronchioloalveolar carcinoma of the lung: pooled analysis of seven case-control studies in the International Lung Cancer Consortium (ILCCO). Cancer Causes Control 22: 73–79.
6. Surveillance, Epidemiology and End Results (SEER) Program Public Use Data (1973-2008). 1973-2008. (2011). National Cancer Institue, DCCPS, Surveillance Research Program, Cancer Statistics Branch.
7. Moolgavkar SH, Venzon DJ (1979) 2-Event Models for Carcinogenesis - Incidence Curves for Childhood and Adult Tumors. Mathematical Biosciences 47: 55–77.
8. Bartoszynski R, Edler L, Hanin L, Kopp-Schneider A, Pavlova L, et al. (2001) Modeling cancer detection: tumor size as a source of information on unobservable stages of carcinogenesis. Mathematical Biosciences 171: 113–142.
9. Foy M, Yip R, Chen X, Kimmel M, Gorlova OY, et al. (2011) Modeling the mortality reduction due to computed tomography screening for lung cancer. Cancer 117: 2703–2708.
10. Foy M, Spitz MR, Kimmel M, Gorlova OY (2011) A smoking-based carcinogenesis model for lung cancer risk prediction. International Journal of Cancer 129: 1907–1913.
11. Foy M, Chen X, Kimmel M, Gorlova OY (2011) Adjusting a cancer mortality-prediction model for disease status-related eligibility criteria. Bmc Medical Research Methodology 11.
12. Moolgavkar SH, Holford TR, Levy DT, Kong CY, Foy M, et al. (2012) Impact of reduced tobacco smoking on lung cancer mortality in the United States during 1975–2000. J Natl Cancer Inst 104: 541–548.
13. Bacac M, Stamenkovic I (2008) Metastatic cancer cell. Annu Rev Pathol 3: 221–247.
14. Del Monte U (2009) Does the cell number 10(9) still really fit one gram of tumor tissue? Cell Cycle 8: 505–506.
15. Egeblad M, Werb Z (2002) New functions for the matrix metalloproteinases in cancer progression. Nat Rev Cancer 2: 161–174.
16. Ramaswamy S, Ross KN, Lander ES, Golub TR (2003) A molecular signature of metastasis in primary solid tumors. Nature Genetics 33: 49–54.
17. Kimmel M, Flehinger BJ (1991) Nonparametric estimation of the size-metastasis relationship in solid cancers. Biometrics 47: 987–1004.
18. Magdowski M, Vick R (2010) Estimation of the Mathematical Parameters of Double-Exponential Pulses Using the Nelder-Mead Algorithm. Ieee Transactions on Electromagnetic Compatibility 52: 1060–1062.
19. Barton RR, Ivey JS (1996) Nelder-Mead simplex modifications for simulation optimization. Management Science 42: 954–973.
20. Olsson DM, Nelson LS (1975) Nelder-Mead Simplex Procedure for Function Minimization. Technometrics 17: 45–51.
21. Usuda K, Saito Y, Sagawa M, Sato M, Kanma K, et al. (1994) Tumor doubling time and prognostic assessment of patients with primary lung cancer. Cancer 74: 2239–2244.
22. Arai T, Kuroishi T, Saito Y, Kurita Y, Naruke T, et al. (1994) Tumor doubling time and prognosis in lung cancer patients: evaluation from chest films and clinical follow-up study. Japanese Lung Cancer Screening Research Group. Jpn J Clin Oncol 24: 199–204.
23. Hasegawa M, Sone S, Takashima S, Li F, Yang ZG, et al. (2000) Growth rate of small lung cancers detected on mass CT screening. Br J Radiol 73: 1252–1259.
24. Gorlova O, Peng B, Yankelevitz D, Henschke C, Kimmel M (2005) Estimating the growth rates of primary lung tumours from samples with missing measurements. Stat Med 24: 1117–1134.
25. Flehinger BJ, Kimmel M (1987) The natural history of lung cancer in a periodically screened population. Biometrics 43: 127–144.
26. Koscielny S, Tubiana M, Le MG, Valleron AJ, Mouriesse H, et al. (1984) Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination. Br J Cancer 49: 709–715.
27. 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.
28. Xu JL, Prorok PC (1998) Estimating a distribution function of the tumor size at metastasis. Biometrics 54: 859–864.