HHS Public Access

Author manuscript

Comput Biol Med. Author manuscript; available in PMC 2017 July 13.

Published in final edited form as:

Comput Biol Med. 2017 February 01; 81: 111–120. doi:10.1016/j.compbiomed.2016.12.011.

A Bayesian model for estimating multi-state disease progression

Shiwen Shen[a,b,*], **Simon X. Han**[a,b], **Panayiotis Petousis**[a,b], **Robert E. Weiss**[c], **Frank Meng**[b], **Alex A.T. Bui**[b], and **William Hsu**[b]

[a]Department of Bioengineering, University of California, Los Angeles, CA, USA

[b]Medical Imaging Informatics (MII) Group, Department of Radiological Sciences, University of California, Los Angeles, CA, USA

[c]Department of Biostatistics, University of California, Los Angeles, CA, USA

Abstract

A growing number of individuals who are considered at high risk of cancer are now routinely undergoing population screening. However, noted harms such as radiation exposure, overdiagnosis, and overtreatment underscore the need for better temporal models that predict who should be screened and at what frequency. The mean sojourn time (MST), an average duration period when a tumor can be detected by imaging but with no observable clinical symptoms, is a critical variable for formulating screening policy. Estimation of MST has been long studied using continuous Markov model (CMM) with Maximum likelihood estimation (MLE). However, a lot of traditional methods assume no observation error of the imaging data, which is unlikely and can bias the estimation of the MST. In addition, the MLE may not be stably estimated when data is sparse. Addressing these shortcomings, we present a probabilistic modeling approach for periodic cancer screening data. We first model the cancer state transition using a three state CMM model, while simultaneously considering observation error. We then jointly estimate the MST and observation error within a Bayesian framework. We also consider the inclusion of covariates to estimate individualized rates of disease progression. Our approach is demonstrated on participants who underwent chest x-ray screening in the National Lung Screening Trial (NLST) and validated using posterior predictive p-values and Pearson's chi-square test. Our model demonstrates more accurate and sensible estimates of MST in comparison to MLE.

Keywords

Bayesian analysis; Markov model; Mean sojourn time; Chest x-ray; Lung cancer; Transition probability; Observation error; Posterior predictive p-value; Markov chain Monte Carlo

[*]Correspondence to: UCLA Medical Imaging Informatics, 924 Westwood Boulevard, Suite 420, Los Angeles, CA 90024, USA. shiwenshen@ucla.edu (S. Shen).

Conflict of interest statement

None Declared.

1. Introduction

Lung cancer, breast cancer, diabetes and coronary heart disease are today's leading causes of death [1]. A better understanding of these diseases' progression and dynamics, such as the expected time to reach a certain disease state, may lead to more appropriate prevention, management and treatment, as well as early detection [2]. Periodic screening using imaging is one of the most common ways to detect early stage disease, especially for cancer. Longitudinal data collected as a result of screening [3] provide an opportunity to discover better approaches for characterizing natural disease progression and generate predictions for individualized screening or diagnostic policies [4]. Traditionally, a "one size fits all" approach has been used for programs such as mammography screening. However, patients at lower risk of cancer should likely have longer screening intervals or not be screened at all.

The mean sojourn time (MST) measures how fast a disease progresses from a preclinical state (imaging detectable but without observable symptoms) to a clinical state (with observable symptoms). MST has been widely used [5] to model disease progression and in the context of population screening, calculate the optimal interval between screens and estimate the extent of overdiagnosis. The overarching objective of our work is to determine how the estimation of MST can be used to inform individualized screening strategies [6]. Four informatics-related challenges exist in leveraging retrospective screening data. First, observations for disease states made in clinical practice are often subject to interpretation error such as when radiologists incorrectly miss a cancerous nodule due to reasons such as noise and artifacts in an image. Failure to model this observation error will bias the MST estimation [7]. Second, missing or partial observations are common in clinical practice. For example, some patients may miss a scheduled screening exam or undergo care at another facility where data is not shared. Third, the interval between screening exams is frequently irregular (e.g., patients do not always come back exactly within one year). Thus, the discretization of continuous time information results in the loss of valuable information [8]. Fourth, the sample size of certain observed disease states may be very small (sparse), thus making the estimation difficult. For instance, patients will usually undergo an intervention if early state cancer is detected, thereby removing them from further observations. As a result, transitions to later states have fewer individuals with which probabilities can be estimated.

To overcome the aforementioned challenges, we use a continuoustime Markov model to represent disease transitions between states. Maintaining continuous time information permits estimation of unobserved states and maintains the interval between screens that is unique for each individual. We then utilize a Bayesian approach to jointly estimate MST, interpretation error, and disease incidence rates using the CMM model and to derive the observed transition probability between states for subsequent rounds of screening. Finally, we demonstrate how the MST can be estimated for different subgroups that are stratified by covariates such as demographics and patient history. We evaluate our model using data from the National Lung Screening Trial (NLST) [9]. In particular, we model the natural history of lung cancer in the chest x-ray (CXR) arm, whose participants underwent three rounds of lung cancer screening.

In Section 2, we introduce prior work related to estimating MST using Bayesian approaches and CMMs. In Section 3, we describe the NLST dataset and the corresponding data pre-processing. The theoretical formulations of our CMM-based Bayesian approach along with specific implementation details of the framework are presented in Sections 3.2–3.4. In Section 4, we summarize the results, comparing the performance of our framework with that of maximum likelihood estimation (MLE) for a three-state Markov model. Finally, in Section 5, we discuss the advantages and limitations of our models and future directions.

2. Background

Numerous techniques for modeling multi-state disease progression, especially for MST, have been proposed. Aalen et al. modeled HIV/AIDS progression using a discrete-time Markov model [10]; Chen et al. presented a three-state discrete progressive model for breast cancer [11]. Multi-state continuous-time Markov models can be adapted to solve the loss of continuous-time information [8] due to interval censoring. In particular, they have been used to model hepatocellular carcinoma [12], liver cirrhosis [13], periodontal disease [2] and diabetic retinopathy [14]. Duffy et al. applied a three-state continuous Markov model to data from a breast cancer randomized controlled trial to estimate the MST and the sensitivity of the screening process [8]. This method assumes perfect sensitivity in estimating the transition times between states and then subsequently estimates the sensitivity using fixed transition times. Chen et al. extended and applied the continuous-time Markov model in breast screening to jointly estimate mean sojourn time, screening sensitivity, and the positive predictive value [15]. Nevertheless, information from the control group (e.g., individuals who received usual care) was needed to properly estimate the desired parameters. Bayesian approaches have been increasingly applied [16–20] to infer MST and screening sensitivity. Our model is capable of modeling the situation where no control group information is available. This is especially relevant in clinical settings where it is unethical to deny treatment. A Bayesian framework applied to breast cancer screening data was used in [16] to obtain age-dependent sensitivity and estimates of transition probabilities. Chien et al. applied a Bayesian approach to validate the effectiveness of computed tomography (CT) for mortality reduction in lung cancer and to estimate the MST [17]. In 2010, Wu et al. used data from the Mayo Lung Project (MLP) to estimate lung cancer screening sensitivity, age-dependent transition probability between states, and the distribution of sojourn time using a Bayesian approach [18]. Bayesian methods have advantages over classical techniques such as enabling small sample inference, providing appropriate measures of uncertainties, allowing inference on non-linear functions of parameters, and constructing predictive distributions to allow for additional inferences of interest [16]. More recently, Jiang et al. [21] used the Day and Walter model [22] to estimate the MST and the false negative rate from the Ontario breast cancer screening program in Canada. Taghipour et al. [23] modeled the natural history of breast cancer with a 4-state hidden Markov model and analyzed the effects of covariates and over different subpopulations. Jia et al. [24] used a 5-state Markov model to detect the worsening of patient symptoms in order to prioritize by symptom severity. Ma et al. [25] used a Bayesian approach on a 5-state continuous time Markov model to investigate a transtheoretical model. The advent of lung cancer and in particular lung screening trials also stimulated the development of a number of risk models to predict

lung cancer incidence from epidemiological and clinical data. Bach et al. [26] developed and combined two logistic regression models that predict the 10-year cumulative probability of dying from lung cancer and dying without lung cancer. Cronin et al. [27] validated this model with the placebo Arm of the Alpha-Tocopherol Beta-Carotene Cancer Prevention (ATBC) study. The model underestimated the observed lung cancer risk and the observed non-lung cancer risk individuals that smoked less than 20 cigarettes per day. A cox proportional hazards regression was developed from the COSMOS trial from epidemiological and clinical data [28]. Model's performance was poor on early cancers but it could identify lower risk individuals and prevent overdiagnosis. Using the PLCO dataset Tammemagi et al. [29] developed a logistic regression model that predicts the six year probability of cancer from a wider range, of incrementally validated using AUC, epidemiological and clinical factors. Petousis et al. [30] developed discrete time dynamic Bayesian networks (DBNs) that predict lung cancer incidence at the different screening points of the NLST trial. The models achieved results comparable to expert's decisions.

In this paper, we extend previous probabilistic models and demonstrate how our model yields a more accurate picture of lung cancer progression. The contributions of this paper are:

1. We provide an approach that serves as the basis for generating individualized screening policies based on estimations of MST for a specific group of individuals stratified by their covariates.
2. We describe how a CMM model parameterized using a Bayesian approach can be applied to accurately model data collected from three rounds of screening.
3. We explore the effect of age and gender on MST in the lung cancer screening population.

Results are validated using Pearson's chi-squared test and posterior predictive p-value to measure the model's fit to the data.

3. Materials and methods

3.1. Overview

As with prior work, we model the natural progression of lung cancer as transitioning through three states (see Fig. 1): a disease-free state (State 1), a preclinical state detectable via screening but asymptomatic (State 2), and a symptomatic state (State 3) [7,8,15,17,31]. The model assumes that a patient in State 1 must go through State 2 to reach State 3. When a patient undergoes screening, one of two states can be observed: if the screening result is positive and confirmed by a diagnostic evaluation, such as a biopsy, the patient is in the preclinical state; otherwise, the patient is in the disease-free state. Thus, the second state (preclinical state) is identified under two conditions: 1) a positive screening test; 2) a confirmed positive pathology diagnosis. However, patients in the preclinical state include both false-negatives due to interpretation error, and true-negatives, both of which can progress to the clinical state. When cancer is first detected by emerging lung cancer symptoms (not through screening), the patient is in the clinical state. In multi-round screening settings, patients who do not progress to the clinical state and are not found to be

preclinical during screening will repeat the process in subsequent rounds. Those found to be symptomatic of lung cancer prior to another round of screening are considered to be interval cases. Fig. 2 illustrates this process. There is observation error when we observe State 1 (i.e., the underlying real state could be either State 1 or State 2), but no observation error is assumed when we observe State 2 and State 3, because both of them are confirmed clinically.

MST is difficult to estimate because the direct transition from the disease-free state to the preclinical state is clinically unobservable. Patients will undergo intervention/treatment after being observed in a preclinical state (a positive cancer screening), thus obviating the natural progression from a preclinical state to a clinical state. Therefore, interval cases become the only source of information for estimating MST if no control group (individuals who never undergo screening) is available. As the discovery of interval cancers is affected by false-negative screening results, estimation of MST is affected by detection sensitivity; a biased estimate of sensitivity can influence the estimate of MST [17]. Sensitivity is the unknown probability of screening detecting preclinical cancer.

3.2. National lung screening trial (NLST) data

The National Lung Screening Trial (NLST) was a large multi-center randomized controlled trial (RCT) of over 53,000 high-risk current or former smokers. Participants were initially between 55 and 74 years old, had smoking histories of at least 30 pack-years and were cancerfree at the start of the trial. The study followed participants between 2002 – 2007, with follow-ups through 2009. Each participant had up to three rounds of screening, with roughly one year between screenings. The study consisted of two arms, chest x-ray (CXR) and computed tomography (CT). If at any point in the study the participant was found to have cancer, he/she did not receive further screenings and was removed from the trial. In this study, we utilize data from the CXR arm. Of the 26,730 total patients originally in the CXR arm, 807 were removed from our analysis due to withdrawal from the study or loss of contact, and 100 were removed because they were discovered to have been ineligible after enrollment (e.g., patient had a CT within 18 months of enrollment). A further 579 patients who did not receive the first round of screening were removed. Only clinically confirmed positive screenings are considered preclinical (State 2). Interval cancers are cases detected after a negative screen, but before the next screen (between first and second screen or between second and third screen). Post-screening cancer cases are those detected after a negative third screen during follow-up, and the follow-up time is up to 5.09 years. Both interval and post-screening cancers are assumed to be symptomatic cancers and defined as clinical (State 3). False-positives were considered to not be cancer (State 1). Table 1 presents a detailed breakdown of events.

3.3. Continuous-time markov model

Let k and l denote one of the three disease states, where $k, l \in \{1, 2, 3\}$. Suppose the disease is in State k at time t and let $p_{kl}(\Delta t) = p_{kl}(t, t + \Delta t)$ denote the probability of transition from State k to l during time period Δt. Then, the instantaneous $\lambda_{kl}(t)$ transition intensity, which represents the instantaneous hazard rate of progression to State l [32], is

$$\lambda_{kl}(t)=\lim_{\Delta t\to 0} p_{kl}(t,t+\Delta t)/\Delta t. \quad (1)$$

Using a time-homogeneous model, both the transition intensity and transition probability is independent of t, where $\lambda_{kl}(t) = \lambda_{kl}$. In this case, the process is stationary and the transition probability $p_{kl}(\Delta t) = p_{kl}(t, t + \Delta t) = p_{kl}(0, \Delta t)$. The three-state instantaneous transition rate matrix Q is

$$Q=\begin{bmatrix} -\lambda_{12} & \lambda_{12} & 0 \\ 0 & -\lambda_{23} & \lambda_{23} \\ 0 & 0 & 0 \end{bmatrix}, \quad (2)$$

whose rows sum to 0, so that the diagonal entries are [32,33]

$$\lambda_{kk}=-\sum_{k\neq 1}\lambda_{kl}. \quad (3)$$

As shown in Fig. 1, transitions could only happen from State 1 to State 2 and from State 2 to State 3. For other undefined transitions, transition rates are 0. Transition rate λ_{12} represents the instantaneous hazard rate to the preclinical state from the disease free state and λ_{23} is the instantaneous hazard rate of transitioning to clinical state from the preclinical state. In the CMM, MST is calculated as $1/\lambda_{23}$. The transition probability matrix in time Δt is the matrix exponential [33] $P(\Delta t) = \exp(Q\Delta t)$

$$P(\Delta t)=\begin{bmatrix} \exp(-\lambda_{12}\Delta t) & \frac{\lambda_{12}(\exp(-\lambda_{23}\Delta t)-\exp(-\lambda_{12}\Delta t))}{\lambda_{12}-\lambda_{23}} & 1+\frac{\lambda_{23}\exp(-\lambda_{12}\Delta t)-\lambda_{12}\exp(\lambda_{23}\Delta t)}{\lambda_{12}-\lambda_{23}} \\ 0 & \exp(-\lambda_{23}\Delta t) & 1-\exp(-\lambda_{23}\Delta t) \\ 0 & 0 & 1 \end{bmatrix},$$

(4)

whose (k,l)th entry is $p_{kl}(\Delta t)$. We first assume no observation error, that is, sensitivity equals 1. We can easily write the likelihood function for each observation, and the parameters can be estimated using maximum likelihood. Transition probabilities from baseline to the first screening time point are conditional on no cancer at baseline. Probabilities for different transitions can be computed and are given in Table 2. The Markov assumption states that transitions only depend on the previous state. Patients are independent given the parameters and thus the log-likelihood function for all subjects is equal to the summation over all participants.

3.4. Modeling imperfect screening sensitivity

Eq. (4) and Table Table (2) assume that sensitivity of the low-dose lung cancer screening exam is 100%. However, in practice, false negatives reduce the test sensitivity. In this section, we introduce a Bayesian model for inferring both sensitivity and transition probabilities simultaneously.

We develop the model for three rounds of screening, including interval cancer and post-screening cancer cases. Time intervals between the first and second screenings and second and third screenings are Δt_{12} and Δt_{23}, respectively. We assume Δt_{12} and Δt_{23} are the same for all participants.

Let A be the average age of all participants at first screening and A is used as the time interval for the first screening (participants are assumed to be disease-free at birth and the first observation time is at first screening) [7,8,17,31]. The real state at time t is Y_t, a random variable with three possible states {1, 2, 3} modeled by the three-state homogeneous Markov process (Fig. 1) with transition hazard matrix Q and transition probability matrix $P(\Delta t)$. The observed state at time t is denoted as Z_t, also with three possible states {1, 2, 3}. At screening, the observed state Z_t is subject to error due to false-negatives, meaning a preclinical state may be incorrectly observed as disease-free. The sensitivity S is $S = Pr(Z_t = 2|Y_t = 2)$ and $1 - S = Pr(Z_t = 1|Y_t = 2)$. The observation error for the other two states are assumed to be zero because they are confirmed clinically: $Pr(Z_t = 1|Y_t = 1) = 1$ and $Pr(Z_t = 3|Y_t = 3) = 1$. Suppose there are T rounds of screening in total (T observations), we make two conditional independence assumptions about Y_t and Z_t, (a) observational independence assumption, that the t^{th} observation given the t^{th} real state is independent of all other observations and real states on that subject

$$P\left(Z_t|Y_T,\, Z_T,\, Y_{T-1},\, Z_{T-1},\, \ldots, Y_{t+1},\, Z_{t+1},\, Y_t,\, Y_{t-1},\, Z_{t-1},\, \ldots, Y_1,\, Z_1\right) = P\left(Z_t|Y_t\right),$$

and (b) Markov assumption, that the tth real state given previous real states is independent of all previous observations and real states except the most recent real state,

$$P\left(Y_t|Y_{t-1},\, Z_{t-1},\, Y_{t-2},\, Z_{t-2},\, \ldots, Y_1,\, Z_1\right) = P\left(Y_t|Y_{t-1}\right).$$

Given the total number of participants in the jth round of screening and the probability that a participant will be observed as preclinical this round, the number of observed preclinical cases at this screening follows a binomial distribution [34]. Let N_j be the total number of attendees at screening j; let π_j be the probability of being observed at the preclinical state at the jth screening, and let n_j be the number of subjects observed at the preclinical state at the jth screening, then we have

$$n_j|\pi_j \sim B(N_j, \pi_j). \quad (5)$$

The probability π_1 that a subject is observed in the preclinical state on the first screening given the subject is not in the clinical state is

$$\pi_1=Pr(Z_A=2|Y_A \neq 3)=S \cdot p_{12}(A)/(p_{12}(A)+p_{11}(A)). \quad (6)$$

Detailed derivations are given in the appendix. Next $\boldsymbol{\pi}_2$ is the probability that a subject is observed in the preclinical state on the second screening given that the patient is not observed in the clinical state after the first screen and was disease-free on the first screening. The true preclinical case here may come from two sources: 1) the patient is disease-free at the first screening and progresses to preclinical at the second screening; 2) the false-negative patient whose real state is preclinical at the first screening and stays in the preclinical state for the second screening.

$$\begin{aligned}\pi_2&=Pr(Z_{A+\Delta t_{12}}=2|Z_A=1, Y_{A+\Delta t_{12}} \neq 3)\\ &=S \cdot (\beta_1 \cdot p_{12}(\Delta t_{12})+\beta_2 \cdot p_{22}(\Delta t_{12}))/\theta_1, \quad (7)\end{aligned}$$

where

$$\beta_2=Pr(Y_A=2|Z_A=1, Y_A \neq 3)=(1-S) \cdot \pi_1/((1-\pi_1) \cdot S), \quad (8)$$

$$\beta_1=Pr(Y_A=1|Z_A=1, Y_A \neq 3)=1-\beta_2, \quad (9)$$

$$\theta_1=Pr(Y_{A+\Delta t_{12}} \neq 3|Z_A=1, Y_A \neq 3)=1-\beta_1 \cdot p_{13}(\Delta t_{12})-\beta_2 \cdot p_{23}(\Delta t_{12}). \quad (10)$$

Finally, $\boldsymbol{\pi}_3$ is the probability that one is observed as preclinical on the third screening given disease-free observations in the two previous screenings and not in the clinical state in any screening. The formula for $\boldsymbol{\pi}_3$ is similar to that of $\boldsymbol{\pi}_2$

$$\begin{aligned}\pi_3&=Pr(Z_{A+\Delta t_{12}+\Delta t_{23}}=2|Z_{A+\Delta t_{12}}=1, Z_A=1, Y_{A+\Delta t_{12}+\Delta t_{23}} \neq 3)\\ &=S \cdot (\gamma_1 \cdot p_{12}(\Delta t_{23})+\gamma_2 \cdot p_{22}(\Delta t_{23}))/\theta_2, \quad (11)\end{aligned}$$

where

$$\begin{aligned}\gamma_2&=Pr(Y_{A+\Delta t_{12}}=2|Z_{A+\Delta t_{12}}=1, Y_{A+\Delta t_{12}} \neq 3, Z_A=1, Y_A \neq 3)\\ &=(1-S) \cdot (p_{22}(\Delta t_{12})\beta_2+p_{12}(\Delta t_{12})\beta_1)/((1-S) \cdot (p_{22}(\Delta t_{12})\beta_2+p_{12}(\Delta t_{12})\beta_1)\\ &\quad +p_{11}(\Delta t_{12})\beta_1), \quad (12)\end{aligned}$$

$$\gamma_1=Pr(Y_{A+\Delta t_{12}}=1|Z_{A+\Delta t_{12}}=1, Y_{A+\Delta t_{12}} \neq 3, Z_A=1, Y_A \neq 3)=1-\gamma_2, \quad (13)$$

$$\theta_2 = Pr(Y_{A+\Delta t_{12}+\Delta t_{23}} \neq 3 | Z_{A+\Delta t_{12}} = 1, Y_{A+\Delta t_{12}} \neq 3, Z_A = 1, Y_A \neq 3) = 1 - \gamma_2 \cdot p_{23}(\Delta t_{23}) - \gamma_1 \cdot p_{13}(\Delta t_{23}). \quad (14)$$

Observed interval cancers and follow-up cancers are assumed to be clinical. Let $N_{j\Delta t}^{(3)}$ be the number of subjects observed in the clinical state between the jth and (j+1)th screening within time interval Δt. Then $N_{j\Delta t}^{(3)}$ is a random variable that can be modeled as a Poisson process [34]. Let $M_j(\Delta t)$ be the mean of the Poisson distribution. Then $M_j(\Delta t)$ is the sum of two parts: 1) the number of patients who progress to the clinical state from the disease-free state within Δt; and 2) the number of progressions from the preclinical states, which are false-negatives of the previous screening that transit into the clinical state within Δt [34].

The first part is the product of the number of disease-free patients at jth screening $N_j^{(1)}$ times $p_{13}(\Delta t)$; and the second part is the product of the number of false-negative subjects at jth screening times $p_{23}(\Delta t)$.

Section 4 discusses the priors adopted for the model. We use Markov Chain Monte Carlo (MCMC) to generate random samples from the joint posterior distribution of the parameters with the WinBUGS software program [35].

3.5. Considering covariates

Our model can be extended to evaluate the effects of covariates, such as gender and age, on parameter estimation. In this study, the effects of covariates are investigated with a stratified analysis by fitting our model separately for age groups >60 and ≤60 to yield independent estimates of parameters (λ_{12}, λ_{23}, S) for each age group [31]. The same analysis is also performed on different gender groups. The NLST CXR dataset used in this work enrolled only high-risk lung cancer subjects, who are former or current smokers and have a minimum of 30 pack-years of cigarette smoking history. To further divide the dataset into sub-cohorts and investigate the cancer progression differences, covariates are identified from demographics, smoking history and disease history, including age, gender, family history of cancer, body mass index, disease history, cancer history, current or former smoker, number of packs of cigarette per year and smoke years. Distributions of each covariates within no-cancer, non-symptomatic cancer, symptomatic cancer and post-screening cancer groups are plotted and compared to identify the significant covariates in this high-risk cohort for further stratification. Age and gender are two significant and important factors that have also been used by previous lung cancer studies [18,36], thus being used here and reported in Section 4.3 and Appendices A and B.

3.6. Evaluation

To validate the proposed model and compare with other methods, two metrics are employed. First, we use Pearson's chi-square to check the adequacy of the proposed model and validate parameter estimates by checking whether there are significant differences between the observed and expected counts [7,17,31,34,37]. A p-value larger than 0.05 suggests no significant difference indicating a good fit and accurate estimation of parameters. Then,

posterior predictive p-values (PPPV) [38] are employed to check the inconsistency between model predictions and observed counts for the Bayesian approach. A p-value away from 0 for the PPPV indicates a good fit.

In the Results section, we first present the results of a CMM model (no observation error) fit with MLE using NLST dataset for estimation of MST. We then present the results of the model including observation error using the proposed Bayesian approach. Using Pearson's chi- square and PPPV, we show that the proposed model fits the data better than the model without observation error.

4. Results

Section 4.1 gives results using the simple CMM model assuming 100% sensitivity. Pearson's chi-square reveals a poor fit using simple CMM model. Section 4.2 gives results for transition intensities and sensitivity using the Bayesian approach with and without covariates. Both Pearson's chi-square test and the PPPV suggest a good fit for the proposed model. Based on our covariate analysis in Section 4.3, the MST is longer in the older population.

4.1. Maximum likelihood without observation error

MLE is used to estimate the parameters of the three-state CMM when assuming sensitivity is 1. Two parameters $\theta(\lambda_{12}, \lambda_{23})$ need to be estimated. The likelihood calculation is implemented in R and quasiNewton function maximization is used for the optimization step. The initial value for λ_{12}, the incidence rate of preclinical disease, is set to 0.00552 based on a study done by Manser et al. [34,39]. The initial value for λ_{23}, the incidence rate of clinical disease and the inverse of MST, is set to 0.52 based on the inverse of the average reported CXR MST range (0.46 – 3.35) [17,19,20,34]. With these initial values, λ_{12} is estimated at 0.0154 (95% CI: 0.0143 – 0.0164), λ_{23} is estimated at 3.31 (95% CI: 2.90 – 3.72) and MST is estimated as 0.302 years (95% CI: 0.269 – 0.345). However, the chi-square is 610.7 with p smaller than 0.00001, indicating a poor fit for the model.

4.2. Bayesian approach

In our analysis of the CXR data using the proposed Bayesian approach, there are now three parameters to be estimated, $\theta = (\lambda_{12}, \lambda_{23}, S)$. In previous studies, sensitivity of CXR in lung cancer is reported as being in the range of 69 – 90% [18,34,40] with a mean around 80%. Thus, a Beta distribution with $\alpha = 5$ and $\beta = 2$ is adopted as the prior for sensitivity with a mode at 80%. Uniform distributions are employed as priors for λ_{12} and λ_{23}. A range from 0.0005 to 0.05 is selected as the prior for λ_{12} to allow enough flexibility given previously reported values in studies [34,39] as described in Section 4.1. Similarly, the prior of λ_{23} is chosen to be uniform of 0.2–5 based on previous studies [17,19,20,34].

We use two sub-chains and WinBUGS [35] to sample from the posterior for $\theta = (\lambda_{12}, \lambda_{23}, S)$. Each MCMC simulation is run for 45,000 iterations, with a burn-in of 5,000 iterations. Convergence was essentially immediate. After the burn-in, the posteriors are sampled and stored every 8 iterations, generating 5,000 posterior samples per chain. The 10,000 posterior samples, θ^i, $i = 1, \ldots, 10000$, are pooled for analysis. The program is running on a Windows

8.1 desktop with a Intel Xeon CPU (3.3 GHz and 3.30 GHz) and 16 GB RAM. The WinBUGS program running time is 3.2 min.

Table 3 shows summaries of the posterior for the parameters. MST and sensitivity are estimated as 1.09 years (95% CI: 0.914 – 1.34) and 0.899 (95% CI: 0.761 – 0.984), respectively. Table 4 shows results of the chi-square test. There is no significant difference between observed and expected values, indicating a good fit to the empirical data. Compared to the model with perfect sensitivity, the expanded model fits the data better. The estimated MST is much longer compared to the reduced model fit assuming a sensitivity of 1. This corresponds to the expectation that lower sensitivity will lead to higher MST. This trend is also demonstrated in Fig. 3, which plots 1000 randomly selected posterior samples of MST and sensitivity.

To further evaluate the expanded model, we also employ a posterior predictive p-value [38] to assess the model fit. Let $y = (y_1, \ldots y_9)$ denote the observed data, where y_a is the number of positive, negative and interval cases or post-screening cancer subjects for all three screenings. Let $y_{rep} = (y_{rep1}, \ldots y_{rep9})$ be the replicated data that could have been observed. A χ^2 discrepancy is the sum of squares of standardized residuals of the data with respect to their expectations under a posterior model and defined as [38]

$$\chi^2(y;\theta)=\sum_{a=1}^{9}\frac{(y_a - E(y_a|\theta))^2}{\mathrm{Var}(y_a|\theta)}. \quad (15)$$

where $Var(y_a|\theta)$ represents the variance of y_a given the parameter vector θ and $E(y_a|\theta)$ represents the expectation. For each posterior sample $\theta^{(b)}$, $b = 1, \ldots,10000$, draw a simulated replicated data set, $y_{\mathrm{rep}}^{(b)}$, from the sampling distribution $p(y_{\mathrm{rep}}^{(b)}|\theta^{(b)})$. Then, calculate $\chi^2(y;\theta^{(b)})$and $\chi^2(y_{\mathrm{rep}}^{(b)};\theta^{(b)})$. Fig. 4 plots $\{(\chi^2(y;\theta^{(b)}), \chi^2(y_{\mathrm{rep}}^{(b)};\theta^{(b)})), b=1, \ldots, 10000\}$. The estimated PPPV is calculated as the proportion of the 10,000 pairs for which $\chi^2(y_{\mathrm{rep}}^{(b)};\theta^{(b)})$ exceeds $\chi^2(y; \theta^{(b)})$ [38]. The estimated PPPV is 0.381 as shown in Fig. 4 and it does not indicate a lack fit for the model.

4.3. Covariate analysis

The explanatory variables (covariates) can be included using a stratified analysis. For gender, the data is divided into male and female and the model refit for each group. At the first screening, there are 14,936 males and 10,308 females. Table 5 shows posterior summaries for these two groups. There is little difference between females and males in terms of MST, sensitivity, and incidence rate of preclinical disease (λ_{12}). For age, the sample is divided into two groups: age≤60 (12,669 participants) and those older than 60 (12,575 participants). Table 6 shows the posterior summaries for these two groups. The incidence rate λ_{12} of preclinical disease is two times larger in the older age group, and the 95% confidence intervals do not overlap. However, the MST and sensitivity are similar. This indicates that subjects (heavy smokers) older than 60 are twice as likely to progress to the preclinical state compared to those who are younger, consistent with the observations that

there are significantly more detected preclinical cases as shown in Appendix Table A1. There is also a significantly higher percentage of observed interval/post-screening cancers in the older group. This is reasonable given that there are more subjects in the preclinical state. Tables A1 and A2 give the chi-square test results for the models fit separately by age and gender. The results suggest that the model fits well in all sub-groups. Compared to the whole population, fit is improved in each of the age groups. The estimated PPPV for the age sub-groups ≤60, >60, male and female are 0.604, 0.507, 0.610 and 0.259, respectively. All values indicate no lack of fit (Table 7).

5. Discussion

We present a CMM-based model that incorporates observation error to model estimate multi-state disease progression. Applied to lung cancer screening data from the NLST dataset, the model produces results that are plausible and consistent with published literature [17,18,34,41]. The CMM is a natural approach to take for modeling the transitions of discrete health states [31]. It can model transitions over time by incorporating longitudinal patient data and variable observation intervals. Sensitivity and MST are two important unknown parameters in the model. However, these two parameters are correlated and difficult to untangle as shown in Fig. 3, especially when no information is available for the incidence rate from a control group [7]. Without a control group, MST can only be estimated from the occurrence of interval cancer cases. On the other hand, more falsenegative cases leads to more occurrences of interval cases, resulting in a shorter MST estimate. Thus, MST and sensitivity should be modeled jointly [42,43] and estimates are sensitive to small changes in interval cancer counts [7]. In [17,31], MST is estimated assuming sensitivity is 1, which is quite optimistic in reality. Dufiy and Chen et al. [8] proposed a two step method: firstly, MST is estimated by assuming sensitivity to be 1, and secondly, sensitivity is re-estimated using the obtained MST. This method is similarly still subject to error due to not estimating both jointly. In our study, we first investigated MST assuming sensitivity to be 1 similar to step 1 in [8]. As shown in Section 4.1, the model did not fit well. In [7,15,32,37], sensitivity is modeled as part of the likelihood function, and MLE is adopted to estimate the parameters.

To compare with other methods, we implemented a three-state model from [7] using the R programming language [44] to model the same NLST CXR data. A general multi-state Markov model software package developed in [37] was also used to try to fit the data. However, no stable estimates were obtained in either case. Overall, our method uses the probabilistic Bayesian approach to model the observed occurrences for each state and jointly estimated MST and sensitivity and provides improved fit. An additional benefit is that the likelihood is able to use all the data including data from the third screening and post-screening cancer cases.

There are some limitations to this work. First, the preclinical state is defined as the state in which the disease is detectable by screening. Therefore, depending on the screening modality, the probability of transition into the preclinical state will also vary. For instance, CT has better resolution for detecting lung cancer, and a CT-screened patient might enter the preclinical state earlier relative to a CXR-screened patient. Therefore, MST and sensitivity

are specific to each screening modality. Second, our current model models population level information by using average transition times between states as in [7,17,31,34]. By using individual data, we can make inferences on the MST distribution across the patient population, possibly improving accuracy of the estimates. This also opens the path towards using patients' electronic health record (EHR) data for individualized screening schedules. Future work will focus on an individualized Bayesian framework that models each patient's information separately. In the reduced model where sensitivity is 1, both average and individualized patient transition times were investigated and the estimates for MST were very similar. Thus, the fit of our proposed model does not lose much generalizability using average transition time.

This work builds the basis for providing individualized screening recommendations for a specific group of individuals stratified by their covariates. In traditional clinical settings, screening programs, such as mammography screening, employ a "one size fits all" paradigm. Nevertheless, subjects with lower risk (cancer progression rates) should receive less frequent screenings to reduce cost and radiation exposure. MST provide a way to quantify this screening period for difierent cohorts. This work provide a way to more robustly calculate MST for specific cohorts with sparse observations, while considering the observation error. In clinical practice, with collected screening data (may be noisy and sparse), this work make it possible to more accurately determine suitable screening periods. It also serves as the foundation to move towards individualized screening, where a personalized screening paradigm will be provided for each subject.

6. Conclusion

In this paper, we propose a Bayesian approach to model disease progression based on a continuous-time Markov model using periodic screening data. Observation error (sensitivity) is incorporated into the model as one of the parameters. The model is able to capture continuous time information, while compensating for missing observations in disease transitions. Priors are included to narrow the parameter space. The method is applied to a large chest x-ray screening dataset collected on heavy smokers as part of the National Lung Screening Trial. Gender and age are evaluated as two explanatory covariates in the proposed model. The participants older than 60 are found to have twice the incidence rate of preclinical lung cancer compared to subjects between ages 55 and 60. Pearson's chi-square and posterior predictive p-values are adopted to validate the model. Compared to previous Markov models, estimation using our method is stable and yields a better fit of the empirical data.

Acknowledgments

The authors acknowledge the National Cancer Institute and National Lung Screening Trial for sharing the data. This work has been partially supported by NCI R01 Grant CA1575533 and NIBIB Grant EB000362.

Appendix A. Chi-square test table

Table A1

Goodness of fit by age group.

	Observed	Expected	Residual
Age≤60			
First screen negative	12631	12633.5	−2.5
First screen positive	38	35.5	2.5
Interval cancers after first screen	14	19.1	−5.1
Second screen negative	11828	11830.2	2.2
Second screen positive	18	15.8	−2.2
Interval cancers after second screen	16	15.2	0.8
Third screen negative	11258	11257.2	0.8
Third screen positive	20	20.8	−0.8
Post-screening cancers after third screen	154	146.9	7.1
$\chi^2 = 2.261$, $P=0.894$			
age>60			
First screen negative	12481	12481.3	−0.3
First screen positive	94	93.7	0.3
Interval cancers after first screen	34	38	−4
Second screen negative	11614	11605.3	8.7
Second screen positive	46	54.7	−8.7
Interval cancers after second screen	26	31.9	−5.9
Third screen negative	11079	11082.5	−3.5
Third screen positive	54	50.5	3.5
Post-screening cancers after third screen	330	316.8	13.2
$\chi^2 = 3.697$, P=0.718			

Table A2

Goodness of fit by gender group.

	Observed	Expected	Residual
Male			
First screen negative	14858	14859.7	−1.7
First screen positive	78	76.3	1.7
Interval cancers after first screen	30	34	−4
Second screen negative	13932	13927.8	4.2
Second screen positive	42	46.2	−4.2
Interval cancers after second screen	23	29	−6
Third screen negative	13307	13303.3	3.7
Third screen positive	39	42.7	−3.7
Post-screening cancers after third screen	293	277.9	15.1
$\chi^2 = 3.275$, $P=0.774$			

	Observed	Expected	Residual
Female			
First screen negative	10254	10255.4	−1.4
First screen positive	54	52.6	1.4
Interval cancers after first screen	18	22.8	−4.8
Second screen negative	9510	9500.8	9.2
Second screen positive	22	31.2	−9.2
Interval cancers after second screen	19	18.6	0.4
Third screen negative	9030	9036.3	−6.3
Third screen positive	35	28.7	6.3
Post-screening cancers after third screen	191	185.6	5.4
χ^2 = 5.323, P=0.503			

Appendix B. Formulas for π1, π2 and π3

$$\begin{aligned}\pi_1 &= Pr(Z_A=2|Y_A\neq 3)=Pr(Z_A=2,Y_A\neq 3)/Pr(Y_A\neq 3)\\ &=(Pr(Z_A=2|Y_A=2)Pr(Y_A=2)+Pr(Z_A=1|Y_A=2)Pr(Y_A=2))/\\ &\quad (Pr(Y_A=2)+Pr(Y_A=1))\\ &=(S\cdot p_{12}(A))/(p_{12}(A)+p_{11}(A)).\end{aligned} \tag{B.1}$$

$$\begin{aligned}\pi_2 &= Pr(Z_{A+\Delta t_{12}}=2|Z_A=1,Y_A\neq 3,Y_{A+\Delta t_{12}}\neq 3)\\ &=Pr(Z_{A+\Delta t_{12}}=2,Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3)/Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3)\\ &=(Pr(Z_{A+\Delta t_{12}}=2,Y_{A+\Delta t_{12}}=2|Z_A=1,Y_A\neq 3)+Pr(Z_{A+\Delta t_{12}}=2,(Y_{A+\Delta t_{12}}=1|\\ &\quad Z_A=1,Y_A\neq 3))/(Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3))\\ &=(Pr(Z_{A+\Delta t_{12}}=2|Y_{A+\Delta t_{12}}=2,Z_A=1,Y_A\neq 3)Pr(Y_{A+\Delta t_{12}}=2|Z_A=1,Y_A\neq 3)\\ &\quad +Pr(Z_{A+\Delta t_{12}}=2|Y_{A+\Delta t_{12}}=1,Z_A=Y_A\neq 3)Pr(Y_{A+\Delta t_{12}}=1|Z_A=1,Y_A\neq 3))/\\ &\quad Pr(Z_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3).\\ &=S\cdot Pr(Z_{A+\Delta t_{12}}=2|Z_A=1,Y_A\neq 3)/Pr(Y_{A+\Delta t_{12}}\neq 3|Z_A=1,Y_A\neq 3)\\ &=S\cdot(\beta_1\cdot p_{12}(\Delta t_{12})+\beta_2\cdot p_{22})(\Delta t_{12}))/\theta_1\end{aligned} \tag{B.2}$$

where (B.2) follows from assumptions (a) and (b) and

$$\begin{aligned}\beta_2 &= Pr(Y_A=2|Z_A=1,Y_A\neq 3)=Pr(Z_A=1|Y_A=2,Y_A\neq 3)Pr(Y_A=2|Y_A\neq 3)/\\ &\quad Pr(Z_A=1|Y_A\neq 3)\\ &=(1-S)\cdot\pi_1/((1-\pi_1)\cdot S),\end{aligned} \tag{B.3}$$

$$\beta_1=Pr(Y_A=1|Z_A=1,Y_A\neq 3)=1-\beta_2, \tag{B.4}$$

$$\begin{aligned}\theta_1 &= Pr(Y_{A+\Delta t_{12}} \neq 3 | Z_A = 1, Y_A \neq 3) \\ &= 1 - Pr(Y_{A+\Delta t_{12}} \neq 3 | Z_A = 1, Y_A \neq 3) \\ &= 1 - Pr(Y_A = 1 | Z_A = 1, Y_A \neq 3) Pr(Y_{A+\Delta t_{12}} = 3 | Y_A = 1, Z_A = 1, Y_A \neq 3) - Pr(\\ & Y_A = 2 | Z_A = 1, Y_A \neq 3) Pr(Y_{A+\Delta t_{12}} = 3 | Y_A = 2, Z_A = 1, Y_A \neq 3) \\ &= 1 - \beta_1 \cdot p_{13}(\Delta t_{12}) - \beta_2 \cdot p_{23}(\Delta t_{12}).\end{aligned} \tag{B.5}$$

$$\begin{aligned}\pi_3 &= Pr(Z_{A+\Delta t_{12}+\Delta t_{23}} = 2 | Z_{A+\Delta t_{12}} = 1, Z_A = 1, Y_A \neq 3, Y_{A+\Delta t_{12}} \neq 3, Y_{A+\Delta t_{12}+\Delta t_{23}} \neq 3) \\ &= S \cdot (\gamma_1 \cdot p_{12}(\Delta t_{23}) + (\gamma_2 \cdot p_{22}(\Delta t_{23}))/\theta_2\end{aligned} \tag{B.6}$$

where

$$\begin{aligned}\gamma_2 &= Pr(Y_{A+\Delta t_{12}} = 2 | Z_{A+\Delta t_{12}} = 1, Y_{A+\Delta t_{12}} \neq 3, Z_A = 1, Y_A \neq 3) \\ &= Pr(Y_{A+\Delta t_{12}} = 2, Z_{A+\Delta t_{12}} = 1 | Z_A = 1, Y_A \neq 3)/(Pr(Y_{A+\Delta t_{12}} = 2, Z_{A+\Delta t_{12}} = 1 | Z_A = 1, \\ & Y_A \neq 3) + Pr(Y_{A+\Delta t_{12}} = 1, Z_{A+\Delta t_{12}} = 1 | Z_A = 1, Y_A \neq 3)) \\ &= (Pr(Y_{A+\Delta t_{12}} = 1 | Y_{A+\Delta t_{12}} = 2, Z_A = 1, Y_A \neq 3, Y_A = 2) Pr(Y_{A+\Delta t_{12}} = 2 | Z_A = 1, Y_A \\ & \neq 3, Y_A = 2) Pr(Y_A = 2 | Z_A = 1, Y_A \neq 3) + Pr(Z_{A+\Delta t_{12}} = 1 | Y_{A+\Delta t_{12}} = 2, Z_A = 1, Y_A \\ & \neq 3, Y_A = 1) Pr(Y_{A+\Delta t_{12}} = 2 | Z_A = 1, Y_A \neq 3, Y_A = 1) Pr(Y_A = 1 | Z_A = 1, Y_A \neq 3))/ \\ & (Pr(Y_{A+\Delta t_{12}} = 2, Z_{A+\Delta t_{12}} = 1 | Z_A = 1, Y_A \neq 3) + Pr(Y_{A+\Delta t_{12}} = 1, Z_{A+\Delta t_{12}} = 1 | Z_A \\ & = 1, Y_A \neq 3)) = (1 - S) \cdot (p_{22}(\Delta t_{12})\beta_2 + p_{12}(\Delta t_{12})\beta_1)/((1 - S) \cdot (p_{22}(\Delta t_{12})\beta_2 + p_{12}(\Delta t_{12})\beta_1) + \\ & p_{1,1}(\Delta t_{12})\beta_1),\end{aligned}$$

(B.7)

$$\gamma_1 = Pr(Y_{A+\Delta_{12}} = 1 | Z_{A+\Delta t_{12}} = 1, Y_{A+\Delta_{12}} \neq 3, Z_A = 1, Y_A \neq 3) = 1 - \gamma_2, \tag{B.8}$$

$$\begin{aligned}\theta_2 &= Pr(Y_{A+\Delta t_{12}+\Delta t_{23}} \neq 3 | Z_{A+\Delta t_{12}} = 1, Y_{A+\Delta t_{12}} \neq 3, Z_A = 1, Y_A \neq 3) \\ &= 1 - \gamma_2 \cdot p_{23}(\Delta t_{23}) - \gamma_1 \cdot p_{13}(\Delta t_{23}).\end{aligned} \tag{B.9}$$

References

1. Russell LB. Preventing chronic disease: an important investment, but don't count on cost savings. Health Aff. 2009; 28:42–45.
2. Mdala I, Olsen I, Haffajee AD, Socransky SS, Thoresen M, Blasio BF. Comparing clinical attachment level and pocket depth for predicting periodontal disease progression in healthy sites of patients with chronic periodontitis using multi-state Markov models. J Clin Periodontol. 2014; 41:837–845. [PubMed: 24888705]
3. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986:121–130. [PubMed: 3719049]

4. Zhou J, Wang F, Hu J, Ye J. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records, in. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. :135–144.
5. Duffy SW. Screening, Sojourn Time, Wiley Online Library. 2005
6. Shen S, Han S, Petousis P, Meng F, Bui AA, Hsu W. Continuous markov model approach using individual patient data to estimate mean sojourn time of lung cancer, AMIA. Annu Symp Proceedings. 2015
7. Uhry Z, Hedelin G, Colonna M, Asselain B, Arveux P, Rogel A, Exbrayat C, Guldenfels C, Courtial I, Soler-Michel P, et al. Multi-state Markov models in cancer screening evaluation: a brief review and case study. Stat Methods Med Res. 2010; 19:463–486. [PubMed: 20231370]
8. Duffy SW, Chen HH, Tabar L, Day NE. 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. 1995; 14:1531–1543. [PubMed: 7481190]
9. Aberle DR, DeMello S, Berg CD, Black WC, Brewer B, Church TR, Clingan KL, Duan F, Fagerstrom RM, Gareen IF, et al. Results of the two incidence screenings in the national lung screening trial. New Engl J Med. 2013; 369:920–931. [PubMed: 24004119]
10. Aalen OO, Farewell VT, de Angelis D, Day NE, Noel Gill O. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. Stat Med. 1997; 16:2191–2210. [PubMed: 9330428]
11. Chen JS, Prorok PC. Lead time estimation in a controlled screening program. Am J Epidemiol. 1983; 118:740–751. [PubMed: 6638001]
12. Kay R. A Markov model for analysing cancer markers and disease states in survival studies. Biometrics. 1986; 44:855–865.
13. Andersen PK, Hansen LS, Keiding N. Assessing the influence of reversible disease indicators on survival. Stat Med. 1991; 10:1061–1067. [PubMed: 1876794]
14. Marshall G, Jones RH. Multi-state models and diabetic retinopathy. Stat Med. 1995; 14:1975–1983. [PubMed: 8677398]
15. Chen H, Duffy S, Tabar L. A. Markov, chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. Statistician. 1996:307–317.
16. Wu D, Rosner GL, Broemeling L. MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. Biometrics. 2005; 61:1056–1063. [PubMed: 16401279]
17. Chien CR, Lai MS, Chen THH. Estimation of mean sojourn time for lung cancer by chest X-ray screening with a Bayesian approach. Lung Cancer. 2008; 62:215–220. [PubMed: 18400331]
18. Wu D, Erwin D, Rosner GL. Sojourn time and lead time projection in lung cancer screening. Lung Cancer. 2011; 72:322–326. [PubMed: 21075475]
19. Kim S, Erwin D, Wu D. Efficacy of dual lung cancer screening by chest X-ray and sputum cytology using Johns Hopkins lung project data. J Biomet Biostat. 2012; 3
20. Chen Y, Erwin D, Wu D. Over-diagnosis in lung cancer screening using the MSKC-LCSP data. J Biomet Biostat. 2014; 5:2.
21. Jiang H, Walter S, Brown P, Chiarelli A. Estimation of screening sensitivity and sojourn time from an organized screening program. Cancer Epidemiol. 2016; 44:178–185. [PubMed: 27619724]
22. Day NE, Walter SD. Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. Biometrics. 1984:1–13. [PubMed: 6733223]
23. Taghipour S, Caudrelier LN, Miller AB, Harvey B. Using simulation to model and validate invasive breast cancer progression in women in the study and control groups of the Canadian national breast screening studies i and ii. Med Decis Mak. 2016:1–12.
24. Jia J, Barbera L, Sutradhar R. Using markov multistate models to examine the progression of symptom severity among an ambulatory population of cancer patients: are certain symptoms better managed than others? J Pain Sympt Manag. 2016; 51:232–239.
25. Ma J, Chan W, Tilley BC. Continuous time markov chain approaches for analyzing transtheoretical models of health behavioral change: a case study and comparison of model estimations. Stat Methods Med Res. 2016:1–14.

26. Bach PB, Kattan MW, Thornquist MD, Kris MG, Tate RC, Barnett MJ, Hsieh LJ, Begg CB. Variations in lung cancer risk among smokers. J Natl Cancer Inst. 2003; 95:470–478. [PubMed: 12644540]
27. Cronin KA, Gail MH, Zou Z, Bach PB, Virtamo J, Albanes D. Validation of a model of lung cancer risk prediction among smokers. J Natl Cancer Inst. 2006; 98:637–640. [PubMed: 16670389]
28. Maisonneuve P, Bagnardi V, Bellomi M, Spaggiari L, Pelosi G, Rampinelli C, Bertolotti R, Rotmensz N, Field JK, DeCensi A, et al. Lung cancer risk prediction to select smokers for screening CT? A model based on the italian COSMOS trial. Cancer Prev Res. 2011; 4:1778–1789.
29. Tammemagi CM, Pinsky PF, Caporaso NE, Kvale PA, Hocking WG, Church TR, Riley TL, Commins J, Oken MM, Berg CD, et al. Lung cancer risk prediction: prostate, lung, colorectal and ovarian cancer screening trial models and validation. J Natl Cancer Inst. 2011; 103:1058–1068. [PubMed: 21606442]
30. Petousis P, Han SX, Aberle D, Bui AA. Prediction of lung cancer incidence on the low-dose computed tomography arm of the national lung screening trial: a dynamic bayesian network. Artif Intell Med. 2016; 72:42–55. [PubMed: 27664507]
31. Shih HC, Chou P, Liu CM, Tung TH. Estimation of progression of multi-state chronic disease using the Markov model and prevalence pool concept. BMC Med Inform Decis Mak. 2007; 7:34. [PubMed: 17996074]
32. Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multistate Markov models for disease progression with classification error. J R Stat Soc: Ser D (Stat). 2003; 52:193–209.
33. Kalbfleisch J, Lawless JF. The analysis of panel data under a Markov assumption. J Am Stat Assoc. 1985; 80:863–871.
34. Chien CR, Chen THH. Mean sojourn time and effectiveness of mortality reduction for lung cancer screening with computed tomography. Int J Cancer. 2008; 122:2594–2599. [PubMed: 18302157]
35. Lunn DJ, Thomas A, Best N, Spiegelhalter D. Winbugs – a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput. 2000; 10:325–337.
36. Zang EA, Wynder EL. Differences in lung cancer risk between men and women: examination of the evidence. J Natl Cancer Inst. 1996; 88:183–192. [PubMed: 8632492]
37. Jackson CH. Multi-state models for panel data: the MSM package for R. J Stat Softw. 2011; 38:1–29.
38. Gelman A, Meng XL, Stern H. Posterior predictive assessment of model fitness via realized discrepancies. Stat Sin. 1996; 6:733–760.
39. Manser R, Dalton A, Carter R, Byrnes G, Elwood M, Campbell DA. Cost- effectiveness analysis of screening for lung cancer with low dose spiral CT (computed tomography) in the Australian setting. Lung Cancer. 2005; 48:171–185. [PubMed: 15829317]
40. Toyoda Y, Nakayama T, Kusunoki Y, Iso H, Suzuki T. Sensitivity and specificity of lung cancer screening using chest low-dose computed tomography. Br J Cancer. 2008; 98:1602–1607. [PubMed: 18475292]
41. ten Haaf K, van Rosmalen J, de Koning HJ. Lung cancer detectability by test, histology, stage, and gender: estimates from the NLST and the PLCO trials. Cancer Epidemiol Biomark Prev. 2015; 24:154–161.
42. Weedon-Fekjær H, Vatten LJ, Aalen OO, Lindqvist B, Tretli S. Estimating mean sojourn time and screening test sensitivity in breast cancer mammography screening: new results. J Med Screen. 2005; 12:172–178. [PubMed: 16417693]
43. Weedon-Fekjær H, Lindqvist BH, Vatten LJ, Aalen OO, Tretli S. Estimating mean sojourn time and screening sensitivity using questionnaire data on time since previous screening. J Med Screen. 2008; 15:83–90. [PubMed: 18573776]
44. R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria: 2013.

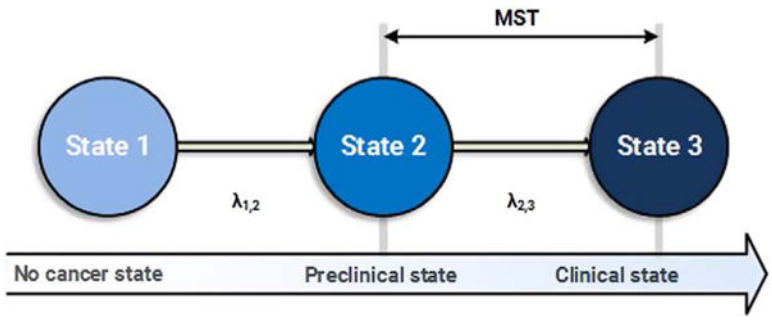

Fig. 1.

Model state transition diagram. State 1 is the disease-free state, State 2 is the preclinical state and State 3 is the clinical state. Parameters λ_{12} and λ_{23} are the transition intensities for transitioning from State 1 to State 2 and State 2 to State 3, respectively.

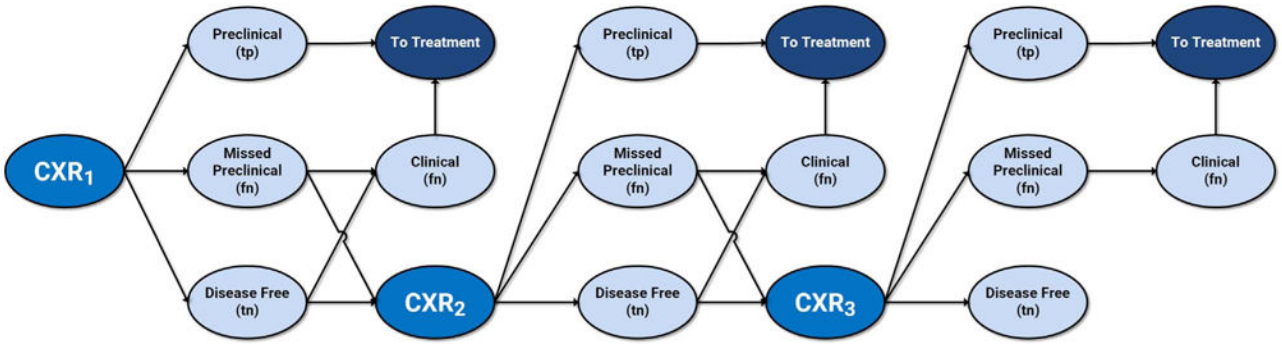

Fig. 2.

An illustration of possible outcomes from periodic CXR screening, where CXR_j represents the jth screening. CXR_2 and following screening will have similar possible outcomes and procedure as with CXR_1. If the subjects are observed in the preclinical state in the first screening, they will enter treatment (and stop periodic screening CXR). Otherwise, subjects are observed to be in the disease free state. However, these observed disease-free subjects include both false-negatives (missed preclinical cases) and true-negatives. Some subjects, who are found at the clinical state (lung cancer symptoms emerge) prior to another round of screening, are called interval cases and also will not undergo additional screening. These interval cases may come from missed preclinical subjects or true disease-free subjects. Subjects who do not progress to the clinical state repeat the process in subsequent rounds.

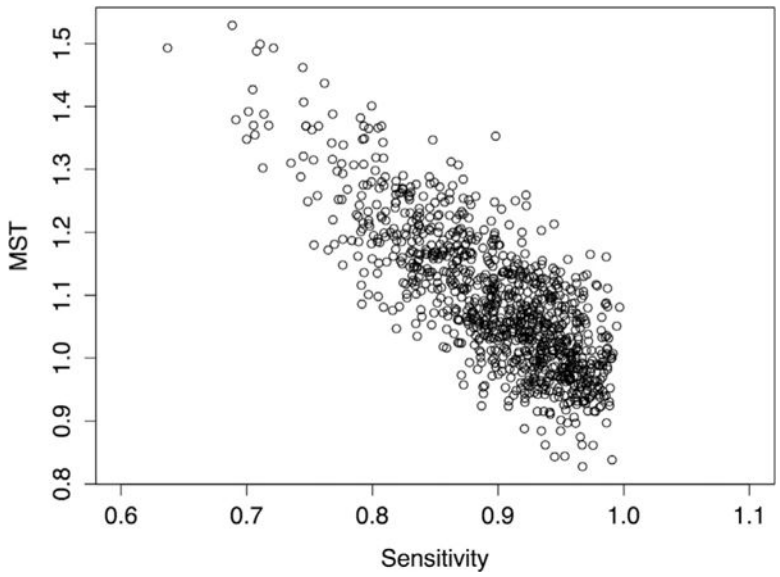

Fig. 3.

Scatter plot of 1000 randomly selected posterior samples of sensitivity and corresponding MST.

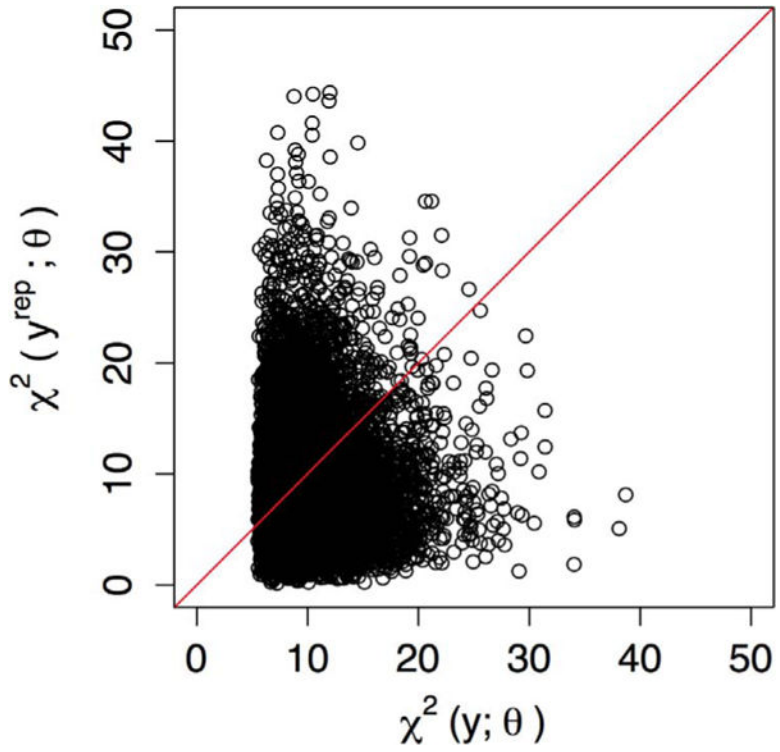

Fig. 4.

Scatter plot of predictive and realized log likelihood ratio discrepancies for the proposed Bayesian model using the whole CXR data set; the proportion of points above the red 45° line represents the proportion of $\chi^2(y_{\text{rep}}^{(b)};\theta^{(b)})$ exceeding $\chi^2(y;\theta^{(b)})$ and is the posterior predictive p-value (PPPV). A PPPV away from 0 indicates a good model fit. The PPPV is 0.381.

Table 1

Detailed chest x-ray participant breakdown.

Screening	Detection Mode	Number	Transition Types
First screening	Participants	25244	
	Screening-Detected Cases	132	No disease→Preclinical
	Negative Screening Cases	25112	No disease→No disease
	Interval Cancers (Between 1st/2nd screenings)	48	No disease→Clinical
Second screening	Participants	23506	
	Screening-Detected Cases	64	No disease→Preclinical
	Negative Screening Cases	23442	No disease→No disease
	Interval Cancers (Between 2nd/3rd screenings)	42	No disease→Clinical
Third Screening	Participants	22411	
	Screening-Detected Cases	74	No disease→Preclinical
	Negative Screening Cases	22337	No disease→No disease
	Post-screening Cancers (after 3rd screening)	484	No disease→Clinical

Table 2

Likelihood function for the Markov model.

Observation Type	Probability
Disease free at 1st screening	$P_1=\frac{\exp(-\lambda_{12}\Delta t)}{\exp(-\lambda_{12}\Delta t)+\frac{\lambda_{12}(\exp(-\lambda_{23}\Delta t)-\exp(-\lambda_{12}\Delta t))}{\lambda_{12}-\lambda_{23}}}$
Disease free at 2nd or 3rd screening	$P_2=\exp(-\lambda_{12}\Delta t)$
Preclinical disease at 1st screening	$P_3=1-P_1$
Preclinical disease at 2nd or 3rd screening	$P_4=\frac{\lambda_{12}(\exp(-\lambda_{23}\Delta t)-\exp(-\lambda_{12}\Delta t))}{\lambda_{12}-\lambda_{23}}$
Clinical disease (interval and post-screening cases)	$P_5=1-\frac{\lambda_{23}\exp(-\lambda_{12}\Delta t)-\lambda_{12}\exp(\lambda_{23}\Delta t)}{\lambda_{23}-\lambda_{12}}$

Table 3

Summaries of the posterior.

Parameter	Mean	SD	2.5%	97.5%
λ_{12}	0.00525	0.000185	0.00489	0.00562
λ_{23}	0.927	0.0889	0.748	1.09
MST (year)	1.09	0.108	0.914	1.34
Sensitivity	0.899	0.0589	0.761	0.984

Table 4

Goodness of fit with sensitivity<1.

	Observed	Expected	Residual
First screening negative	25112	25115.1	−3.1
First screening positive	132	128.9	3.1
Interval cancers after first screening	48	55	−7
Second screening negative	23442	23428.9	13.1
Second screening positive	64	77.1	−13.1
Interval cancers after second screening	42	47	−5
Third screening negative	22337	22339.2	2.2
Third screening positive	74	71.8	−2.2
Post-screening cancers after third screening	484	464.2	19.8
$\chi^2 = 4.643$, $P=0.590$			

Table 5

Summaries of the posterior for the two gender groups.

Parameter	Gender	Mean	SD	2.5%	97.5%
λ_{12}	Male	0.00528	0.000241	0.00483	0.00577
	Female	0.00519	0.000294	0.00463	0.00578
λ_{23}	Male	0.908	0.112	0.694	1.13
	Female	0.889	0.125	0.652	1.14
MST	Male	1.12	0.143	0.884	1.44
	Female	1.15	0.171	0.878	1.54
Sensitivity	Male	0.871	0.0728	0.705	0.980
	Female	0.866	0.0801	0.677	0.980

Table 6

Summaries of the posterior for the two age groups.

Parameter	Age Group	Mean	SD	2.5%	97.5%
λ_{12}	≤60	0.00322	0.000205	0.00284	0.00364
	>60	0.00732	0.000314	0.00670	0.00795
λ 23	≤60	0.986	0.164	0.687	1.33
	>60	0.872	0.0988	0.676	1.07
MST	≤60	1.04	0.181	0.753	1.46
	>60	1.16	0.137	0.937	1.48
Sensitivity	≤60	0.848	0.0883	0.646	0.976
	>60	0.881	0.0684	0.723	0.982

Table 7

Comparison between modeling approaches.

Model	Comment
Shih et al. [31]	Uses a Markov model in conjunction with the prevalence pool concept, but has the limitations of assuming a steady state disease rate. Parameters are estimated using an Expectation-Maximum likelihood algorithm.
Chien et al. [17]	Uses a Bayesian approach to estimate parameters of a 3-state lung cancer Markov model. The authors assumes imaging exams have 100% sensitivity.
Petousis et al. [30]	Uses a discrete time dynamic Bayesian network to predict lung cancer incidence across time points. But the discrete time nature make it impossible to estimate MST.
Proposed method	Uses continuous-time Markov model and developed a Bayesian framework for parameter estimation. Provides analytical solutions to model observed occurrences for each state and jointly estimated MST and sensitivity. It is able to use all observed data including data from the third screening and postscreening cancer cases.