Computers in Biology and Medicine 81 (2017) 111–120

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

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], 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

ARTICLE INFO

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

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.

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

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

http://dx.doi.org/10.1016/j.compbiomed.2016.12.011
Received 30 June 2016; Received in revised form 15 November 2016; Accepted 18 December 2016

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 continuous-time 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

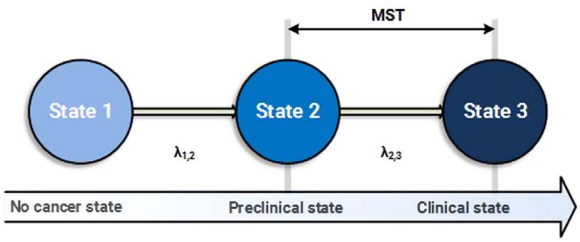

Fig. 1. Model state transition diagram. State 1 is the disease-free state, State 2 is the preclinical state and State 3 is the clinical state. Parameters λ_{12} and λ_{23} are the transition intensities for transitioning from State 1 to State 2 and State 2 to State 3, respectively.