Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 17

11-1-2005

Simulation Procedure In Periodic Cancer Screening Trials

Ioana Barnicescu
Mississippi State University

Ricolindo L. Cariño
Mississippi State University

DOI: 10.22237/jmasm/1130804160

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 522-527

1538 – 9472/05/$95.00

Simulation Procedure In Periodic Cancer Screening Trials

Dongfeng Wu Xiaoqin Wu
Department of Mathematics and Statistics
Mississippi State University

Ioana Banicescu
Department of Computer Science and Engineering
Mississippi State University

Ricolindo L. Cariño
Center for Computational Sciences ERC
Mississippi State University

A general simulation procedure is described to validate model fitting algorithms for complex likelihood functions that are utilized in periodic cancer screening trials. Although screening programs have existed for a few decades, there are still many unsolved problems, such as how age or hormone affects the screening sensitivity, the sojourn time in the preclinical state, and the transition probability from disease-free state to the preclinical state. Simulations are needed to check reliability or validity of the likelihood function combined with the associated effect functions. One bottleneck in the simulation procedure is the very time consuming calculations of the maximum likelihood estimates (MLE) from generated data. A practical procedure is presented, along with results for when both sensitivity and transition probability into the preclinical state are age-dependent. The procedure is also suitable for other applications.

Key words: periodic screening, breast cancer, early detection, sensitivity, sojourn time, transition probability, mammogram, clinical breast examination, incidence

Introduction

According to a recent report of the National Institute of Health (NIH 2000), breast cancer is the most common form of cancer among women in the United States and the second leading cause of cancer deaths among women. One of the procedures to manage the disease is periodic cancer screening, which has been utilized for a few decades. The motivation for screening is to detect the disease early even before clinical symptoms come up. The benefit for early detection is obvious. People in whom cancer is detected earlier usually have a better prognosis. Early treatments hopefully will lead to more cure and prolonged survival of cancer patients.

In a screening program, a large group of asymptomatic individuals are enrolled in the program to detect the presence of a specific disease. The natural history of the disease for an individual is assumed to follow a progressive stochastic model, which consists of three states, denoted by $S_0 \to S_p \to S_c$, corresponding, respectively, to the disease-free state; the preclinical disease state, in which an asymptomatic individual unknowingly has disease that the screening exam can detect; and the clinical state when the disease manifests itself in clinical symptoms. The screening sensitivity is the probability that the screening exam is positive, given that the individual is in the preclinical stage. The sojourn time refers to the time beginning when the disease first

Dongfeng Wu is an Assistant Professor, with research interests in cancer screening probability modeling and inference. She is on the Editorial Board of *JMASM*. Xiaoqin Wu is a PhD candidate. His research interest is in PDE modoling and statistical modeling. Ioana Banicescu is an Associate Professor, with research interests in parallel algorithms, scientific computing, scheduling theory, optimization and prediction. Ricolindo L. Carino received his Ph.D. from La Trobe University, and is a member of the research faculty. His main interest is parallel computing for scientific applications.

develops until the manifestation of clinical symptoms, that is $(S_c - S_p)$. The transition probability into the preclinical stage is the probability density function of making transition from the disease-free to the preclinical state. Knowledge of the sensitivity of the screening modality is necessary for evaluating the predictive performance of a screening exam. The screening sensitivity may depend on a variety of factors, including age, position, location and size of the tumor, and the experience of the radiologist, etc. For example, recent studies indicate that the sensitivity of mammography increases with age at diagnosis (Shapiro, et. al., 1988; Miller, et. al., 1992a, 1992b), attributable to the fact that breast tissue tends to be more dense and fibrous in younger women, and more soft and fatty in older women (Kerlikowske, et. al., 1996).

There is great interest in determining the properties of the sensitivity, the sojourn time distribution and the transition probability density function into the preclinical state. Much work has been done in this area (Shen & Zelen, 1999; Shen, et. al., 2001; Wu, et. al., 2005). The research is still ongoing because many researchers are trying to explore how age or hormone changes may affect the sensitivity, the sojourn time, and the transition probability. One of the common features in the research is to derive the correct likelihood function and to propose correct age effect (or hormone effect) functions based on the stochastic model and the screening data. However, it is imperative to validate the reliability of the likelihood function and the associated effect functions before these can be applied to real data. This validation may be accomplished through simulation, which has become an acceptable procedure to check that the model fitting and the complex algorithms work well with this complicated likelihood.

The remainder of the article is organized as follows. A generalized stochastic model and its likelihood function in a periodic cancer screening program is introduced, as well as the age-dependent sensitivity and transition probability density. The simulation procedure, the corresponding algorithm and results of applying it to a sample scenario are then presented. It will conclude with a discussion of the results of the research.

The Model

Consider a cohort of initially asymptomatic individuals who enroll in a screening program. The sensitivity is denoted by $\beta(t)$, where t is the individual's age at the screening exam. Define $w(t)dt$ as the probability of a transition from S_0 to S_p during $(t, t+dt)$. Let $q(t)$ be the probability density function of the sojourn time in S_p. Finally, let $Q(z) = \int_z^\infty q(x)dx$, that is, $Q(z)$ is the survivor function of the sojourn time in the preclinical state S_p. Throughout this article, the time variable t represents the participating individual's age. If random variables T and S are the duration times in S_0 and S_p respectively, then an individual will enter the clinical state S_c at age $T+S$, the probability density function of $T+S$ is

$$I(t) = \int_0^t w(x)q(t-x)dx,$$

which is the observable incidence of clinical cases.

Consider a cohort of women in the study group who are all aged t_0 at study entry, and a protocol calls for K ordered screening examinations occur at ages $t_0 < t_1 < \cdots < t_{K-1}$, where $t_i = t_0 + i$ for annual screening exams. Define the i-th screening interval as the time interval between the i-th and the $(i+1)$-th screening exams (t_{i-1}, t_i), $i=1,2,\ldots, K-1$. The i-th generation of individuals consists of those who enter S_p during this interval. The 0-th generation includes all who enter S_p before the initial screening exam; let $t_{-1} \equiv 0$.

For each screening exam, let n_{i,t_0} be the total number of individuals in this cohort examined at the i-th screening; s_{i,t_0} is the number of cases detected at the i-th screening exam; and r_{i,t_0} is the number of cases diagnosed in the clinical state S_c within the interval (t_{i-1}, t_i). The latter cases are called interval cases.

Let D_{k,t_0} be the probability that an individual will be diagnosed at the *k*-th scheduled exam (at which her age is $t_{k-1} = t_0 + k - 1$) given that she is already in the preclinical state. Let I_{k,t_0} be the probability of being incident in the *k*-th screening interval. In Wu, et. al., 2005, these two probabilities were derived as:

$$D_{k,t_0} = \beta(t_{k-1})\{\sum_{i=0}^{k-2}[1-\beta(t_i)]\cdots[1-\beta(t_{k-2})]$$
$$\int_{t_{i-1}}^{t_i} w(x)Q(t_{k-1}-x)dx + \int_{t_{k-2}}^{t_{k-1}} w(x)Q(t_{k-1}-x)dx\}.$$

$$I_{k,t_0} = \sum_{i=0}^{k-1}[1-\beta(t_i)]\cdots[1-\beta(t_{k-1})]$$
$$\int_{t_{i-1}}^{t_i} w(x)[Q(t_{k-1}-x)-Q(t_k-x)]dx$$
$$+\int_{t_{k-1}}^{t_k} w(x)[1-Q(t_k-x)]dx.$$

The likelihood function for this cohort of women is

$$L(\cdot\,|\,t_0) = \prod_{k=1}^{K} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1-D_{k,t_0}^{s_{k,t_0}} - I_{k,t_0}^{r_{k,t_0}})^{n_{k,t_0}-s_{k,t_0}-r_{k,t_0}} \quad (1)$$

The full likelihood for the study group across all ages is

$$L = \prod_{t_0}\prod_{k=1}^{K} D_{k,t_0}^{s_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1-D_{k,t_0}^{s_{k,t_0}} - I_{k,t_0}^{r_{k,t_0}})^{n_{k,t_0}-s_{k,t_0}-r_{k,t_0}} \quad (2)$$

The age effect was modeled in the sensitivity and the transition probability simultaneously in the following way. The sensitivity β is associated with age *t* by a logistic link,

$$\beta(t) = \frac{1}{1+\exp(-b_0 - b_1 * (t-\bar{t}))},$$

Where $\bar{t}$ is the average age at entry in the whole study group. If $b_1 > 0, \beta(t)$ will be a monotone increasing function of age *t*.

The transition probability density function *w(t)* is the instantaneous probability of a transition from S_0 to S_p. The integral $\int_0^\infty w(t)dt$ represents a lifetime risk for a healthy female to transit into the preclinical state. According to the NCI's SEER database (Ries et al. 2002), a woman's lifetime risk of being diagnosed with breast cancer is 15.7%, which is less than a women's lifetime risk of entering the preclinical disease state. Hence, 20% was chosen as a reasonable upper bound. The following was chosen

$$w(t) = \frac{0.2}{\sigma t\sqrt{2\pi}}\exp\{-\frac{(\log t-\mu)^2}{2\sigma^2}\},$$

which is the pdf of lognormal(μ, σ^2) multiplied by 20%. That is, w(t) is a sub-density function, where μ and σ^2 are parameters to be estimated.

The loglogistic distribution was adopted to model the sojourn time in the preclinical state,

$$q(x) = \frac{\kappa x^{\kappa-1}\rho^{\kappa}}{[1+(x\rho)^{\kappa}]^2}, x > 0,$$

where *x* is the sojourn time, and κ and ρ are positive parameters, represent the scale and location in the loglogistic family. An advantage of this family over the exponential is that it has two parameters and is more robust in the tails. Another advantage of this family is that its relatively simple form achieved for the survivor function and the hazard function. Its first moment can be calculated directly from

$$EX = \frac{\pi}{\rho\kappa}\csc\left(\frac{\pi}{\kappa}\right).$$

For the *r*-th moment to exist, $\kappa > r$ is needed. For justifications on how these age effect functions are chosen, see Wu et. al., 2005.

Simulation Procedure and Results

The purpose of the simulation is to check the reliability of the likelihood function as

screening sensitivity and transition probability are both age-independent. The key steps were summarized in the non-routine simulation study here. In fact, based on the steps here, one can explore other possible associated functions between age and sensitivity, age and transition density, age and sojourn time, etc.

In the proposed model, there are six unknown parameters, that is, $\theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho)$. Theoretically the parameters have a domain of either $(-\infty, \infty)$ or $(0, \infty)$. The practical meaning of these parameters will limit them to a finite range. The range for each of them was identified as: $0 < b_0 < 5$, $-0.2 < b_1 < 0.2$, $3.5 < \mu < 4.5$, $0 < \sigma^2 < 1$, $0.1 < \rho < 2.0$, and $1 < \kappa < 5$. For justifications of these ranges, see Wu, et. al., 2005.

This simulation consisted of two stages. First, age-dependent screening data based on input values of $\theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho)$ were generated, assuming that initially there are about 100,000 individuals in each age group from age 40 to 64 who will take part in the periodic screening exams. For the input values of θ, the values for $b_0, b_1, \mu, \sigma^2, \kappa$ and ρ was randomly chosen from the valid range above. Second, the MLE $\hat{\theta}$ was computed from our likelihood function using the simulated data. This procedure was repeated n = 1,000 times, then the sample mean and the sample standard deviation of the MLE were collected, and were compared with the input values of θ. If the MLE is close to the true input value of θ, then our likelihood function and the age- dependent functions work well in the modeling.

Here are more details in Step 1: Suppose there are M= 100,000 women who were born in the same year, and who will take part in the screening exam at age t_0. Their duration time spent in the disease-free state (S_0) and in the preclinical state (S_p) can be generated by the density functions *w(t)* and *q(t)* correspondingly. Since w(t) is a sub-density function, it is not obvious how to generate random variables directly from its density. The number of incident cases from disease-free into preclinical state age by age will be generated, using the probability *w(t)dt* which is binomially distributed. Then, for women in the preclinical state at age *t*, their incident time can be generated uniformly in *(t, t+1)*. See Appendix for programming details.

For details in Step 2: The log likelihood function can be implemented in C language. Then, taking the negative value of the log likelihood and calling the S-PLUS routine "nlminb" will provide a local minimum. This local minimum corresponds to a local maximum in the log likelihood. However, computer software has not been found that can find the global minimum (maximum) for a general function. To overcome this problem, the initial point of θ was chosen randomly and the procedure was repeated 5 times for each simulated data and find the global maximum.

The simulation programming code, written in C++ and S-PLUS, is attached in the Appendix. It runs well in a PC environment. Eight simulation results are listed in Table 1. For each true value of θ, the sample mean and sample standard error (S.E.) of the MLE of θ from 1000 simulations are listed. The consistency between the sample mean of the MLE and the input parameters is clearly shown.

Conclusion

The purpose of this article is to provide a simulation procedure in periodic cancer screening trials, with the computer programming code in C++ and S-PLUS. A practical issue encountered in the simulation is that it is very time consuming when MLE was calculated from the simulated data. The procedure for each MLE calculation usually takes about 20 minutes if the code is written in S-PLUS, making it impractical to repeat the procedure for 1000 times. To decrease the computation time, the likelihood part was implemented in C++, which resulted in the whole 1000 simulation procedure finishing in two or three days. The simulation and programming code can be slightly modified to fit other age effect or hormone effect models as well. Hopefully this will help other researchers in this area to carry out their simulation studies.

Table 1. Summary of the simulation results for the six parameters

	b_0	b_1	μ	σ^2	κ	ρ
True value	2.07	-0.05	4.05	0.80	4.54	0.70
MLE estimate	2.073	-0.051	4.053	0.799	4.525	0.698
S.E. of MLE	0.112	0.006	0.042	0.018	0.245	0.016
True value	0.91	-0.07	4.24	0.51	3.01	0.74
MLE estimate	0.879	-0.069	4.242	0.510	3.046	0.730
S.E. of MLE	0.093	0.004	0.019	0.015	0.150	0.029
True value	2.72	-0.12	3.65	0.55	3.73	0.65
MLE estimate	2.714	-0.120	3.652	0.551	3.750	0.647
S.E. of MLE	0.157	0.011	0.021	0.018	0.133	0.012
True value	3.14	0.12	4.42	0.86	1.16	1.23
MLE estimate	3.169	0.123	4.420	0.861	1.161	1.223
S.E. of MLE	0.308	0.029	0.024	0.034	0.015	0.025
True value	0.47	-0.17	3.59	0.15	1.67	0.76
MLE estimate	0.475	-0.170	3.591	0.150	1.667	0.752
S.E. of MLE	0.053	0.004	0.005	0.004	0.023	0.018
True value	1.64	0.02	3.93	0.08	2.37	1.05
MLE estimate	1.612	0.022	3.930	0.080	2.377	1.037
S.E. of MLE	0.150	0.004	0.003	0.001	0.054	0.037
True value	2.81	0.19	4.03	0.67	3.07	0.82
MLE estimate	2.710	0.181	4.029	0.670	3.094	0.812
S.E. of MLE	0.137	0.013	0.033	0.014	0.083	0.012
True value	3.74	-0.04	4.36	0.72	2.74	0.81
MLE estimate	3.650	-0.039	4.361	0.721	2.762	0.801
S.E. of MLE	0.538	0.030	0.024	0.027	0.075	0.021

For more details on how to combine C++ and S-PLUS code, see S-PLUS manual. Current efforts are in transporting this procedure to run on a cluster of Linux workstations. If this effort is successful, the simulation time will be shortened to a few hours.

References

Chen, T. H. H., Kuo, H. S., Yen, M. F., Lai, M. S., Tabar, L. & Duffy, S. W.(2000). Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics 56*, 167-172.

Cox, D. R. & Oakes, D. (1984). *Analysis of survival data*. Chapman & Hall/CRC.

Day, N. E. & Walter, S. (1984). Simplified models of screening for chronic disease: Estimation procedures from mass screening programs. *Biometrics 40*, 1-13.

Eddy, D. M. (1980). *Screening for cancer: Theory, analysis, and design*. Englewood Cliffs, NJ: Prentice Hall.

Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., & Ernster, V. (1996). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association 276*, 33-38.

Lee, S. J. & Zelen, M. (1998). Scheduling periodic examinations for the early detection of disease: Applications to breast cancer. *Journal of the American Statistical Association 93*, 1271-1281.

Miller, A. B., Baines C. J., To, T. & Wall, C. (1992a). Canadian national breast screening study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Canadian Medical Association Journal 147*(10), 1459-76.

Miller, A. B., Baines, C.J., To, T. & Wall, C. (1992b). Canadian national breast screening study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Canadian Medical Association Journal 147*(10), 1477-88.

National Institute of Health (2000). NIH Publication No. 00-1556, 12/12/2000.

Shapiro, S., Venet, W., Strax, P. & Venet L. (1988). *Periodic screening for breast cancer. The Health Insurance Plan Project and its Sequelae, 1963-1986*. Baltimore: The Johns Hopkins University Press.

Shen, Y., Wu, D. & Zelen, M. (2001). Testing the independence of two diagnostic tests. *Biometrics 57*, 1009-1017.

Shen, Y. & Zelen, M. (1999). Parametric estimation procedures for screening programmes: Stable and nonstable disease models for multimodality case finding. *Biometrika 86*, 503-515.

Straatman, H., Peer, P. G. M. & Verbeek, A. L. M. (1997). Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics 53*, 217-229.

Walter, S. D. & Day, N. E. (1983). Estimation of the duration of a preclinical disease state using screening data. *American Journal of Epidemiology 118*, 856-86.

Wu, D., Rosner, G. & Broemeling, L. (2005). MLE and bayesian inference of age-dependent sensitivity and transition probability in periodic screening. *Biometrics, 61*, 1056-1063.

Zelen, M. (1993). Optimal scheduling of examinations for the early detection of disease. *Biometrika 80*, 279-93.