Title

A multivariate cure model for left-censored and right-censored data with application to colorectal cancer screening patterns

Journal

Statistics in Medicine, 35(19)

ISSN

0277-6715

Authors

Hagar, Yolanda C
Harvey, Danielle J
Beckett, Laurel A

Publication Date

2016-08-30

DOI

10.1002/sim.6934

Peer reviewed

HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2017 August 30.

Published in final edited form as:

Stat Med. 2016 August 30; 35(19): 3347–3367. doi:10.1002/sim.6934.

A multivariate cure model for left-censored and right-censored data with application to colorectal cancer screening patterns

Yolanda C. Hagar[a,*,†], **Danielle J. Harvey**[b], and **Laurel A. Beckett**[b]

[a]Department of Applied Mathematics, University of Colorado Boulder, Boulder, CO, U.S.A

[b]Department of Public Health Sciences, University of California, Davis, Davis, CA, U.S.A

Abstract

We develop a multivariate cure survival model to estimate lifetime patterns of colorectal cancer screening. Screening data cover long periods of time, with sparse observations for each person. Some events may occur before the study begins or after the study ends, so the data are both left-censored and right-censored, and some individuals are never screened (the 'cured' population). We propose a multivariate parametric cure model that can be used with left-censored and right-censored data. Our model allows for the estimation of the time to screening as well as the average number of times individuals will be screened. We calculate likelihood functions based on the observations for each subject using a distribution that accounts for within-subject correlation and estimate parameters using Markov chain Monte Carlo methods. We apply our methods to the estimation of lifetime colorectal cancer screening behavior in the SEER-Medicare data set.

Keywords

left-censoring; cure model; multivariate survival; colorectal cancer; SEER-Medicare

1. Introduction

Colorectal cancer (CRC) is the third most common cancer in the USA, both in incidence and mortality rates [1]. Because this cancer is largely asymptomatic, it is important for individuals to be screened regularly. Not only can screening detect CRC at earlier stages, but it can also detect pre-cancerous polyps, which can be removed [1]. The effectiveness of CRC screenings led the United States Preventive Task Force to set screening guidelines in 1996 [2]. In this work, we focus on colonoscopy screenings, which are currently recommended to be performed once every 10 years for average-risk individuals starting at age 50. While more expensive and risky, a colonoscopy is the most thorough form of screening as it can examine the entire colon with few false negatives or false positives and can remove polyps and even some cancers during an examination [3]. Little is known about lifetime colonoscopy screening behavior, as it is challenging to estimate due to incomplete data. Many individuals are not observed for the entirety of their eligibility, possibly censoring observations that

*Correspondence to: Yolanda C. Hagar, Department of Applied Mathematics, University of Colorado Boulder, 526 UCB, Boulder, CO 80309, U.S.A.

†yolanda.hagar@colorado.edu

occur before or after the study period. Additionally, individuals can have zero to many screenings in a lifetime and can have a screening immediately upon becoming due or can delay varying lengths of time.

Previous studies have examined rates of CRC screening in different populations, including the Medicare population (such as those found in Winawer *et al.* [3], Seeff *et al.* [4], Smith *et al.* [5], and Gross *et al.* [6]). However, while these studies report that screening and adherence rates are low, the methods used do not account for screening behavior that may have occurred before or after the study observation period. In addition, these studies do not quantify the average amount of time individuals wait between screenings or how many screenings individuals receive in a lifetime. Because of this lack of information, it has been difficult for researchers to confirm optimal screening guidelines. While screening for CRC reduces cancer risk [1–3], the colonoscopy procedure itself can be risky, requires a trained specialist [7, 8], and unnecessary screenings put an avoidable financial burden on the Medicare system (for example, see Vijan *et al.* [9]). Without knowledge of lifetime screening patterns, it has been difficult to perform long-term cost-benefit analyses for outcomes in CRC. Given the importance of determining lifetime colonoscopy screening behavior, we develop a multivariate survival model that allows for a proportion of subjects to never be screened, and we use our model to estimate patterns in lifetime colonoscopy screening behavior.

2. SEER-Medicare data

We used the Surveillance, Epidemiology and End Results (SEER)-Medicare data set to quantify lifetime colonoscopy screening behavior. This large, public data set is a linkage between the SEER program of cancer registries and Medicare claims files and is one of the largest and most complete data sets containing colonoscopy screening information [10]. However, subjects in the SEER-Medicare data set are age 65 years or older and were only observed between 1991 and 2003. Possible screenings occurred before age 65, before 1991, or after 2003 and were not observed, so lifetime behavior was left-censored and/or right-censored for many individuals. Additionally, some individuals were never screened through colonoscopy, while some individuals were screened more than once in a lifetime. Examples of the complexities of screening behavior can be observed in Figure 1. Note that although the actual colonoscopy screening behaviors of hypothetical subjects A and B are different, the observed trajectories are the same. Similarly, hypothetical subjects C and D have identical observed screening patterns but different true lifetime behaviors. Statistical methods that can account for screening patterns that occur outside the observation window are necessary for proper estimation of lifetime screening behavior.

In addition to estimation of the time and rates of colonoscopy screenings, we are also interested in the impact of health policy changes on screening behavior. Medicare changed insurance coverage policy rules in 1998 to provide increased coverage for colonoscopy screenings. Before 1998, no colonoscopy screenings were covered by Medicare (‘phase 0’). Between 1998 and 30 June 2001 (‘phase 1’), colonoscopy screenings were covered for high-risk individuals (e.g., those with family history of CRC), and starting 1 July 2001 (‘phase 2’), coverage was provided to all Medicare patients, regardless of risk level. Previous work

has shown that an increase in Medicare colonoscopy coverage led to an increase in screenings [6]; however, this has not been examined in the context of lifetime screening patterns or in the quantification of the time to being screened. Understanding the impact of changes in guidelines is important to understanding barriers and patterns of screening behaviors.

The model we propose will answer a number of open questions; we will be able to quantify the number of times individuals are screened in a lifetime and the length of time individuals wait between screenings, while quantifying how insurance policy changes affect screening behaviors. To do this, we develop and implement a multivariate cure model that accounts for both left-censoring and right-censoring, within-subject correlation, the estimation of multiple event times, and the average number of events per person. This model is particularly well suited for the estimation of lifetime CRC screening, as these events are sparse over the course of an individual's lifetime and cover long periods of time. The resulting estimates are robust, despite the left-censoring, right-censoring, and number of subjects who are never screened. The rest of this article is organized as follows: In Section 3, we discuss the multivariate survival methodology we have developed for left-censored and right-censored data, including the Gibbs sampler algorithm used for parameter estimation. We also discuss how covariates are incorporated into the model through the parameters. In Section 4, we present results of a simulation to validate our approach for settings similar to the SEER-Medicare data. In Section 5, we use our methodology to estimate colonoscopy screening behavior in the Medicare population, using the SEER-Medicare data set, assuming individuals can have up to two screenings in a lifetime. Section 6 contains a discussion and concluding remarks.

3. Model

Many methods have been used to examine colonoscopy screening behavior (e.g., [3–6]). These studies examined screening rates using simple approaches, such as counting the number of colonoscopy screenings that occur each year or more sophisticated approaches using Poisson regression [6]. While these statistical approaches provide basic information on colonoscopy screening trends, to quantify lifetime screening behavior, a model is needed that can account for both the left-censoring and right-censoring inherent in the SEER-Medicare data set, as well as allow for some people to never be screened. In the succeeding sections, we formulate a multivariate survival model that determines how long people wait between screenings, as well as the average number of screenings individuals get in a lifetime.

3.1. Background

Cure models allow for the estimation of time to an event when a subset of the population is risk-free and will never experience the event. In the estimation of lifetime screening behavior, the event of interest is a colonoscopy screening, and those individuals who will never be screened are part of the population that will never experience the event and hence are 'cured' in traditional model terminology. The time to event is calculated as the time an average-risk individual waits between becoming due for a screening (occurs at age 50 or 10

years after the previous colonoscopy screening) and actually getting screened, which we refer to as the 'lag time'. (The lag time is depicted by solid lines in Figure 1.)

Cure models have been examined at length. Initial work by Boag [11] presents the mixture model $S_{pop}(t; \psi) = \pi + (1 - \pi)S(t; \psi)$, where $S(t; \psi)$ represents the survival function for individuals who will experience the event and $\lim_{t\to\infty} S_{pop}(t; \psi) = \pi$. This model has been studied extensively and can be seen in Berkson and Gage [12], Haybittle [13, 14], Farewell [15], Goldman [16], Maller and Zhou [17, 18], and others, but can be complex computationally and is difficult to extend to the multivariate case.

In addition to the possibility that some individuals may never be screened, some individuals may receive many screenings in a lifetime, so the possibility of multiple, dependent events (i.e., colonoscopy screenings) must be accounted for. To this end, there are many existing multivariate survival models that have been studied. A large body of work has been devoted to using Cox models and a marginal hazards approach to investigate the effects of covariates on the hazard rate(s), such as that by Wei *et al.* [19], Liang *et al.* [20], Lin [21], Prentice and Hsu [22], Spiekerman and Lin [23], and others. These models obtain population-averaged covariate effects but are mainly attractive when the correlation between observations is not of interest. Hougaard has done much work with a frailty term in multivariate survival and competing-risks models [24–29]; however, these models do not include the possibility of a cured population. Extensive work on multivariate survival models has been performed by Chen *et al.* [30, 31] and perhaps matches our work most closely as some of the models allow for a cured population for right-censored data. In this work, the authors integrate over latent variables representing the number of risks for each subject, as well as a frailty parameter (to account for within-subject correlation) to get a likelihood function that can accommodate multiple events as well as a proportion of subjects who are cured. However, in addition to right-censoring, we are also interested in the case of left-censoring, and require a model that incorporates this type of missingness.

We introduce a type of multivariate cure model that allows for the estimation of multiple lag times for each individual and the probability that an individual will receive zero through many lifetime screenings. Correlation between lag times is accounted for with a frailty term.

3.2. Notation

Assume an individual has M screenings in his or her lifetime, where M is a random variable such that $0 \le M \le \ell < \infty$, with ℓ denoting the maximum number of lifetime screenings possible for any individual. (The time frame for CRC screening is finite, so the assumption $\ell < \infty$ is natural.) The probability that an individual will have m lifetime screenings, that is, $P(M = m)$, is equal to θ_m, $m = 1, \ldots, \ell$, and the probability an individual will never be screened is equal to $\theta_0 = 1 - \sum_{j=1}^{\ell} \theta_j$. Individuals are left-censored if they enter their study after eligibility begins, and individuals are right-censored if observation stops or the study ends before their eligibility is over. In these instances, we may only observe k of the M screenings, where $k \le M \le \ell$, possibly obscuring part of the lifetime screening pattern. We estimate the lag times and probability of individuals receiving M lifetime screenings via the likelihood function and the multivariate survival distribution developed by Hougaard [25]:

For individuals who receive at least one screening, let Y_{ij} represent the j^{th} lag time for the i^{th} subject ($j = 1, \ldots, M_i$, $i = 1, \ldots, n$). The lag times for each individual are correlated through a subject-specific frailty quantity, and conditioned upon this frailty quantity, the lag times are independent within each subject. Let Z_i be the subject-specific quantity for the i^{th} subject. Assume the Z_i's are independent and follow a positive stable distribution with parameter α. The distribution is given by the Laplace transform

$$E\{\exp(-sZ)\} = \exp(-s^{\alpha}),$$

with $\alpha \in (0, 1]$. The case of $\alpha = 1$ represents the case of independent observations within each subject. The choice of the positive stable distribution for the frailty parameter is common (for examples, see Clayton [32], Oakes [33, 34], and Chi and Ibrahim [35]), is a smooth distribution that allows for varying degrees of skewness, and is mathematically convenient. For the i^{th} subject, conditional on M and z, the joint distribution for the lag times is then

$$P(Y_1 > y_{i1}, \ldots, Y_M > y_{iM} | M, Z = z) = \exp\{-z(\Lambda_1(y_{i1}) + \cdots + \Lambda_M(y_{iM}))\},$$

where $\Lambda_j(.)$ is the cumulative hazard of the j^{th} lag time. For all individuals, the multivariate survival distribution, given M, then becomes

$$\begin{aligned} P(Y_1 > y_1, \ldots, Y_M > y_M | M) &= \int_z \exp\{-z(\Lambda_1(y_1) + \cdots + \Lambda_M(y_M))\} p(z) dz \\ &= \exp\{-(\Lambda_1(y_1) + \cdots + \Lambda_M(y_M))^{\alpha}\} \\ &= S(y_1, \ldots, y_M | M) \end{aligned} \quad (1)$$

Using this multivariate survival function, we can calculate the probability of a colonoscopy screening occurring before, at, or after certain time points. For example, the probability of a colonoscopy screening occurring at time y_1 is calculated as

$$P(Y_1 = y_1, Y_2 > y_2, \ldots, Y_M > y_M | M) = -\frac{\partial}{\partial y_1} S(y_1, \ldots, y_M | M), \quad (2)$$

and the probability of observing all M screenings can be calculated as

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_M = y_M | M) = (-1)^M \frac{\partial^M}{\partial y_1 \ldots \partial y_M} S(y_1, \ldots, y_M | M) \quad (3)$$

These calculations are akin to finding the cdf $f(y) = -d/dyS(y)$ in the univariate case. We use the notation $P(Y = y)$ here rather than the standard notation $f(y)$ for the probability density function for the continuous random variable Y with time of colonoscopy screening y, in order to simplify presentation of the joint distribution of all lifetime screening times.

3.3. Likelihood function

Denote the observed data for individual i by $\overrightarrow{W}_i = (t_{Li}, t_{Ri}, t_1, \ldots, t_{k_i})$, where t_{Li} and t_{Ri} represent the left-censoring and right-censoring times (with $t_{Li} = 0$ if an individual is not left-censored and t_{Ri} equal to the end of the observation period if the individual is not right-censored) and $t_1, \ldots, t_{k_i}$ denoting the k_i observed screening times, with $0 \leq k_i \leq M_i \leq \ell$. The case of k_i equal to zero represents the case of no colonoscopy screenings observed in the study period. Using the multivariate survival distribution in Equation (1), we can then write a complete data likelihood function as follows:

$$L(\mathbf{W}, \overrightarrow{\eta} | \phi, \overrightarrow{\theta}, \alpha) = \prod_{i=1}^{n} \prod_{j=0}^{\ell} \left(P(M=j) p_{ij}(\overrightarrow{W}_i | \overrightarrow{\phi}_j, \alpha, M=j) \right)^{\eta_{ij}} = \prod_{i=1}^{n} \prod_{j=0}^{\ell} \left(\theta_j p_{ij}(\overrightarrow{W}_i | \overrightarrow{\phi}_j, \alpha, M=j) \right)^{n_{ij}}, \quad (4)$$

where $\overrightarrow{\phi}_j$ denotes the parameter vector for the survival distribution for $M = j$ lag times and $\eta_{ij} = I_i(M = j)$ is an indicator variable that equals one if subject i has j true colonoscopies in a lifetime (with $\eta_{i0} = 1 - \sum_{j=1}^{\ell} \eta_{ij}$) and 0 otherwise. The $p_{ij}\left(\overrightarrow{W}_i | \overrightarrow{\phi}_j, \alpha, M=j\right)$ are probabilities associated with the i^{th} person having j lifetime screenings and are calculated using the multivariate survival function in Equation (1) based on the screening pattern observed for the i^{th} individual. An example of the calculation of these probabilities is shown in Section 3.4. For the case of no observed screenings (i.e., $M = 0$), the probability of zero screenings, $p_{i0}(.)$, is not defined and the only likelihood contribution in this instance is θ_0.

In using the likelihood function presented in Equation (4), we assume that the left-censoring and right-censoring patterns are independent of the screening process. While this is a reasonable assumption for left-censored subjects, right-censoring could possibly be dependent on the screening process, particularly if an individual is no longer observed because of diagnosis of or death due to CRC. However, our analysis is primarily concerned with the screening pattern for average-risk subjects. Because CRC is slow to progress and is easy to prevent or catch early with regular screenings [3], it seems unlikely that many of the average-risk subjects in this data set are dependently right-censored.

For subjects who are left-censored and/or right-censored and who do not have ℓ observed screenings, the value of M (the true number of lifetime screenings) may be unknown and therefore some or all of $\overrightarrow{\eta}_i = (\eta_{i0}, \ldots, \eta_{i\ell})$ may be unobserved. To overcome this obstacle and to estimate parameters in the presence of a complex likelihood function, we use a Bayesian approach. Within a Markov chain Monte Carlo (MCMC, [36, 37]) routine, for subjects without a fully observed lifetime screening pattern, we sample the number of lifetime screenings using the individual-level probability of the subject receiving 0,…, ℓ screenings in a lifetime. The $\overrightarrow{\eta}_i$ values are then set corresponding to the sampled number of lifetime screenings. Conditioned on these sampled values, the parameters are then sampled using Gibbs sampler [38, 39]. This method for estimation allows us to determine posterior

densities for the parameters and then perform statistical inference on our model. Examples of likelihood contributions and probability calculations are shown in Section 3.4.

3.4. Examples

To fix ideas, we provide example likelihood contributions and functions for the univariate case, the case of two maximum lifetime screenings (i.e., $\ell = 2$). Following, we provide a general algorithm for deriving the likelihood for more than two maximum lifetime screenings.

3.4.1. Example: univariate likelihood—To illustrate our model in its simplest form, we first cover the univariate case of $\ell = 1$ (i.e., individuals can only get one colonoscopy in a lifetime). In the univariate instance, the likelihood function in (4) can be simplified to

$$L(\mathbf{W}, \overrightarrow{\eta} | \overrightarrow{\phi}, \theta) = \prod_{i=1}^{n} \prod_{j=0}^{1} \left(\theta_j p_{ij}(\overrightarrow{W}_i | \overrightarrow{\phi}_j, M=j) \right)^{\eta_{ij}} = \prod_{i=1}^{n} (1-\theta)^{1-\eta_i} \left(\theta p_{i1}(\overrightarrow{W}_i | \overrightarrow{\phi}_1, M=1) \right)^{\eta_i}. \quad (5)$$

In the univariate model, subject i who has an observed screening at time t_{i1} contributes the probability $\theta \times p_{i1}(\overrightarrow{W}_i | \overrightarrow{\phi}_1, M=1) = \theta \times f(t_{i1} | \phi_1)$ to the likelihood function (as η_i is known and equal to 1). Conversely, subjects who have no observed screenings and who are not left-censored or right-censored contribute $(1 - \theta)$ to the likelihood function, with a known $\eta_i = 0$. However, a left-censored subject with no observed screening, who enters the study at time t_{iL}, does not have complete information, and it is not known whether a screening occurred before time t_{iL} or did not occur at all. On this occasion, $p_{i1}(\overrightarrow{W}_i | \overrightarrow{\phi}_1, M=1) = F(t_{iL} | \overrightarrow{\phi}_1)$, accounting for a possible screening before study entry. Whether the individual received a screening or not is unknown, and thus the value of the true M (equal to either 0 or 1) is sampled at each iteration of the MCMC routine with the probability π_{i1} of a screening at the r^{th} iteration calculated as

$$\pi_{i1}^{(r)} = \frac{\theta^{(r)} F\left(t_{iL} | \overrightarrow{\phi}_1^{(r)} \right)}{(1 - \theta^{(r)}) + \theta^{(r)} F\left(t_{iL} | \overrightarrow{\phi}_1^{(r)} \right)}.$$

The probability of zero screenings for subject i can be calculated as $1 - \pi_{i1}^{(r)}$. If the sampled value is equal to 1, then η_i is set to 1, otherwise it is set to 0. Similar probability calculations can be made for right-censored or left-censored and right-censored subjects. In the univariate likelihood, the within-subject correlation parameter α is not necessary, as each subject only has one screening. Note that the probability that individual i receives one screening is not the same as $P(M = 1) = \theta$, which represents the probability of one screening for the entire population.

The univariate likelihood is similar to the complete data likelihood presented in Sy and Taylor [40] for the univariate case of right-censored subjects. This formulation of the cure model does not reduce to the standard cure model as $t \to \infty$; however, this definition allows for a simpler likelihood function in the estimation of left-censored screening behavior.

3.4.2. Example: two possible lifetime screenings—We now extend the univariate example to the multivariate case of $\ell = 2$. In the instance of cure models, one issue that arises is identifiability, where distinguishing between models with low cure rates and long survival times and models with high cure rates and short survival times can become difficult. The identifiability of cure models been examined extensively (for example, see Farewell [41], Li *et al.* [42], Yu *et al.* [43]). In this manuscript, we take a simple approach and employ the following reasonable assumptions to mitigate this issue:

- A maximum recommended screening age exists. Because colonoscopy is not risk-free and benefits are long-term (due to the slow progression of CRC) [3], colonoscopy screening is generally not recommended for average-risk older patients [2].
- The maximum number of lifetime screening colonoscopies, ℓ is fixed. In this example, we set $\ell = 2$. (This is the maximum number observed in the SEER-Medicare data set for average-risk patients).

Using these assumptions, we examine the following subject: An individual i is left-censored at time t_{iL}, is observed until the maximum screening age (i.e., is not right-censored), and has one screening observed at time t_{i1}. Because the individual is left-censored and less than two screenings were observed, there are two possible true trajectories for this subject (Figure 2):

1. The one observed screening is the only lifetime screening (i.e., $M = 1$).
2. A screening also occurred before observation of the individual began, and so the subject was screened twice in his or her lifetime (i.e., $M = 2$).

If the one observed screening is the only lifetime (case 1 earlier, represented by trajectory 1 in Figure 2), then the individual would contribute the following to the likelihood function:

$$\begin{aligned} p_{i1}(\overrightarrow{W}_i|\overrightarrow{\phi}_1, M{=}1) &= P(Y_1{=}t_{i1}|\overrightarrow{\phi}_1, M{=}1) \\ &= -\tfrac{\partial}{\partial y_1} S(y_1|\overrightarrow{\phi}_1, M{=}1)\Big|_{y_1=t_{i1}} . \end{aligned} \quad (6)$$

The likelihood contribution for the second case (represented by trajectory 2 in Figure 2) is calculated under the assumption that one screening occurred before the left-censoring time, and the observed screening is the second lifetime screening (i.e., $M = 2$) and is written as

$$\begin{aligned} p_{i2}(\overrightarrow{W}_i|\overrightarrow{\phi}_2, \alpha, M{=}2) &= P(Y_1{<}t_{iL}, Y_2{=}t_{i1} - 10 - y_1|\overrightarrow{\phi}_2, \alpha, M{=}2) \\ &= -\int_0^{t_{iL}} \left(\frac{\partial}{\partial y_1} S(y_1, y_2|\overrightarrow{\phi}_2, \alpha, M{=}2)\right)\Big|_{y2=t_{i1}-10-y_1} dy_1. \end{aligned} \quad (7)$$

Note that we do not actually observe the length of y_1 (the length of the first lag time), we only know that it is less than the left-censoring time t_{iL}. Similarly, we do not observe y_2, the length of the second lag time, and only know that y_2 is equal to the length of time between the first screening at time y_1 and the second screening at t_{i1}, minus the 10-year waiting period.

In this instance, the true number of lifetime screenings is unknown for subject i is unknown, so at each iteration of the MCMC routine, the value of M is sampled. In this example, the individual-level probabilities at the r^{th} MCMC iteration of either one or two lifetime screenings are calculated as

$$\pi_{i1}^{(r)}=\frac{\theta_1^{(r)}p_{i1}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_1^{(r)},M=1\right)}{\theta_1^{(r)}p_{i1}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_1^{(r)},M=1\right)+\theta_2^{(r)}p_{i2}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_2^{(r)},\alpha^{(r)},M=2\right)} \quad (8)$$

$$\pi_{i2}^{(r)}=\frac{\theta_2^{(r)}p_{i2}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_2^{(r)},\alpha^{(r)},M=2\right)}{\theta_1^{(r)}p_{i1}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_1^{(r)},M=1\right)+\theta_2^{(r)}p_{i2}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_2^{(r)},\alpha^{(r)},M=2\right)}, \quad (9)$$

where $p_{i1}(.)$ and $p_{i2}(.)$ are calculated using Equations (6) and (7). Note that for this subject, because at least one screening was observed, $\pi_{i0} = 0$ and $\eta_{i0} = 0$. If the sampled value of M at the r^{th} iteration is equal to 1, then $\eta_{i1} = 1$, and $\eta_{i2} = 0$, with the reverse occurring if the sampled value of M is equal to 2. As mentioned previously, π_{i0}, π_{i1} and π_{i2} are the individual-level probabilities of 0, 1, or 2 lifetime screenings, while θ_0, θ_1, and θ_2 represent the population probabilities of 0, 1, or 2 lifetime screenings.

3.5. Case of more than two lifetime screenings

Our methodology can be extended to instances where more than two lifetime screenings are possible. In this section, we examine the general algorithm for more than two lifetime screenings (i.e., $\ell > 2$). We employ the same assumptions as with the previous multivariate model, fixing a maximum screening age. In the general case, subjects can fall into one of four categories, which are outlined in the succeeding sections.

3.5.1. Case 1, all screenings observed—For subjects who are not left-censored or right-censored, or the maximum number of lifetime screenings (ℓ) are observed, all information is known. For these subjects, we know the value of M ($M = 0,\ldots, \ell$), the true number of lifetime screenings for subject i. The likelihood contribution for this subject is

$$\theta_M \times p_{iM}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_M,\alpha,M\right)=P(Y_1=t_{i1},Y_2=t_{i2}-t_{i1}-10,\ldots,Y_M=t_{iM}-t_{i,M-1}-10). \quad (10)$$

If no screenings are observed for subject i (and there is no censoring), the equation earlier is reduced to θ_0. Because all information is known for this subject, $\eta_{iM} = 1$ and all other values of $\eta_{ij\neq M}$ are set equal to 0. The probabilities (i.e., $P(Y_1 = y_1,\ldots, Y_\ell = y_\ell)$) can be calculated using techniques akin to those in Equations (2) and (3).

3.5.2. Case 2, subjects only left-censored—Assume subject i is only left-censored, with k observed screenings occurring after t_{iL} ($k = 0, \ldots, M - 1 \leq \ell - 1$). In addition, subject i has p screenings before time t_{iL}, with $p = 0, \ldots, p_{\max}$, with $p_{\max} = \lfloor t_{i1} \div 10 \rfloor$ (where $\lfloor . \rfloor$ is the floor function) is the maximum number of possible screenings that can occur before the first observed screening. (If no screening is observed, than $p_{\max}$ is calculated based on the maximum age for screening or the left-censoring time.) The likelihood contribution for subject i is then

$$\theta_{p+k} \times p_{i,k+p}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_{k+p}, \alpha, M=k+p\right) = \theta_{p+k} \times \int_0^{a_1}\int_0^{a_2}\cdots\int_0^{a_p} P\left(Y_1=y_1,\ldots,Y_p=y_p, Y_{p+1}=t_{i1}-10p-\sum_{j=1}^{p}y_p, Y_{p+2}=t_{i2}-t_{i1}-10,\ldots,Y_{p+k}=t_{ik}-t_{ik-1}-10\right)dy_p\ldots,dy_1, \quad (11)$$

where $a_s = \min\left\{10, t_{iL} - 10p - \sum_{j=1}^{s-1} y_j\right\}$, $s = 1,\ldots,p$. When there are no observed screenings, and if $p = 0$, the equation is reduced to θ_0. In this instance, the true value of M is unknown for subject i, but can be sampled. The probability of subject i getting j ($j = k, \ldots, k + p \leq \ell$) lifetime screenings is calculated as

$$\pi_{ij}^{(r)} = \frac{\theta_j^{(r)} \times p_{ij}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_j^{(r)}, \alpha, M=j\right)}{\sum_{s=k}^{k+p_{\max}} \theta_s^{(r)} \times p_{is}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_s^{(r)}, \alpha^{(r)}, M=s\right)}. \quad (12)$$

3.5.3. Case 3, subjects only right-censored—Assume subject i is only right-censored, with k observed screenings occurring before t_{iR}, with $k = 0, \ldots, M - 1 \leq \ell - 1$. If p screenings occur after the right-censoring time, with $p = 0, \ldots, p_{\max}$, with $p_{\max} = \lfloor (t_{\max} - t_{ik}) \div 10 \rfloor$ (the maximum number of screenings that can occur between the last observed screening and the maximum time for screening), and $t_{\max}$ representing the maximum time a person can get screened. (In the instance of $k = 0$, $p_{\max}$ is instead calculated based on the length of the screening period or the right-censoring time.) The likelihood contribution for subject i is

$$\theta_{p+k} \times p_{i,k+p}\left(\overrightarrow{W}_i|\overrightarrow{\phi}_{k+p}, \alpha, M=k+p\right) = \theta_{p+k} \times \int_{\max\{0, t_{iR}-10-t_{ik}\}}^{a_1} \cdots \int_0^{a_p} P(Y_1=t_{i1},\ldots,Y_k=t_{ik}-t_{ik-1}-10, Y_{k+1}=y_{k+1},\ldots,Y_{k+p}=y_{k+p})dy_{k+1}\ldots dy_{k+p}, \quad (13)$$

where $a_s = \min\{10, t_{\max} - t_{ik} - 10p - \sum_{j=1}^{s} y_{k+j}\}$. If no screenings are observed and $p = 0$, subjects contribute θ_0 to the likelihood function. The probability of subject i getting j lifetime screenings is calculated as in Equation (12).

3.5.4. Case 4, subjects left-censored and right-censored—Assume subject i is left-censored and right-censored, with k observed screenings occurring between t_{iL} and t_{iR}, with $k = 0, \ldots, M - 1 \leq \ell - 1$. If p_L screenings occur before the left-censoring time ($p_L = 0, \ldots, p_{L,\max} = \lfloor t_{i1} \div 10 \rfloor$) and p_R screenings occur after the right-censoring time ($p_R = 0,\ldots, p_{R,\max} = \min\{\lfloor (t_{\max} - t_{iR}) \div 10 \rfloor, \ell - k - p_L\}$), then the likelihood contribution for subject i is

$$\theta_{p_L+p_R+k} \times p_{i,p_L+p_R+k}\left(\vec{W}_i | \vec{\phi}_{p_L+p_R+k}, \alpha, M = p_L + p_R + k\right) = \theta_{p_L+p_R+k} \times \int_0^{a_{1L}} \int_0^{a_{2L}} \cdots \int_0^{a_{pL}} \int_{max\{0, t_{iR}-10-t_{ik}\}}^{a_{1R}} \int_0^{a_{2R}} \cdots \int_0^{a_{pR}} P(Y_1 = y_1, \ldots, Y_{p_L} = y_{p_L}, Y_{p_L+1} = t_{i1} - 10p_L - \sum_{j=1}^{p_L} y_p, Y_{pL+1} = t_{i,p_L+1} - t_{i,pL} - 10, \ldots, Y_{p_L+k} = t_{ik} - t_{ik-1} - 10, Y_{p_L+k+1} = y_{p_L+k+1}, \ldots, Y_{p_L+p_R+k} = y_{p_L+p_R+k}) dy_{p_L+p_R+k} \ldots dy_{p_L+k+1} dy_{p_L} \ldots dy_1, \quad (14)$$

where $a_{s_L} = \min\left\{10, t_{iL} - 10p_L - \sum_{j=1}^{s_L - 1} y_j\right\}$, $s_L = 1,\ldots,p_L$ and $a_{s_R} = \min\{10, t_{\max} - t_{ik} - 10p_R - \sum_{j=1}^{s_R - 1} y_{p_L+k+j}\}$, $s_R = 1, \ldots, p_R$. At the $r^{(\text{th})}$ iteration of the MCMC routine, the probability subject i receives j lifetime screenings that can then be calculated as

$$\pi_{ij}^{(r)} = \frac{\theta_j^{(r)} \sum_{s \in (k+p_L+p_R=j)} p_{is}\left(\vec{W}_i | \vec{\phi}_s^{(r)}, \alpha^{(r)}, M = s\right)}{\sum_{t=k}^{k+p_{L,\max}+p_{R,\max}} \theta_t^{(r)} \sum_{s \in (k+p_L+p_R=t)} p_{is}\left(\vec{W}_i | \vec{\phi}_s^{(r)}, \alpha^{(r)}, M = s\right)}. \quad (15)$$

In the instance of $p_R = 0$ (i.e., no screenings occur after the right-censoring time), the equation is reduced to the form shown in (11). Similarly, when $p_L = 0$ (no screenings occur before the left-censoring time), the equation is reduced to the form shown in (13). For the instance where no screenings are observed (i.e., $k = 0$) and both p_L and p_R are equal to 0, the likelihood contribution is θ_0.

3.6. Parameter estimation

Because the likelihood is high-dimensional and the observed screening colonoscopies are sparse, we use Gibbs sampler [38] to estimate the posterior distributions of each parameter, iterating through the following steps:

1. Draw $\overrightarrow{\theta}$ from the conditional posterior distribution

$$\Pi(\overrightarrow{\theta}|\overrightarrow{\gamma},\boldsymbol{\eta}) \propto \left\{\prod_{i=1}^{n}\prod_{j=0}^{\ell}\theta_j\right\}\times Dir\left((\gamma_0,\ldots,\gamma_\ell)^T\right)=Dir\left(\left(\sum_i \eta_{i0}+\gamma_0,\ldots,\sum_i \eta_{i\ell}+\gamma_\ell\right)^T\right),$$

where Dir(.) denotes the Dirichlet distribution [44].

2. Sample each γ_j, $j = 0, \ldots, \ell$, the parameters of the Dirichlet prior for $\overrightarrow{\theta}$, assuming an exponential hyperprior with parameter s_j, from the conditional posterior distribution

$$\Pi(\gamma_j|\theta_j,\overrightarrow{\gamma}_{j-}) \propto \frac{\Gamma\left(\sum_{k=1}^{\ell}\gamma_k\right)}{\Gamma(\gamma_j)}\theta_j^{\gamma_j-1}\exp\{-s_j\gamma_j\},$$

where $\overrightarrow{\gamma}_{j-}$ denotes the vector or γ parameters without the j^{th} element.

3. Sample each element in $\overrightarrow{\phi}_j=(\phi_{j1},\ldots,\phi_{jq})$, for all $\overrightarrow{\phi}_j$, $j = 1, \ldots, \ell$. If $\phi_{jk}\sim g_{jk}(.;\overrightarrow{\kappa}_{jk})$, $k = 1, \ldots, q$, the conditional distribution is as follows:

$$\Pi(\phi_{jk}|\mathbf{W},\boldsymbol{\eta},\overrightarrow{\kappa}_j) \propto \prod_{i=1}^{n} p_{ij}\left(\overrightarrow{W}_i|\phi_j,\alpha,M=j\right)^{\eta_{ij}} g_{jk}(\phi_{jk};\overrightarrow{\kappa}_{jk}),$$

where $\overrightarrow{\kappa}_{jk}$ is the parameter vector for the prior distribution of φ_{jk}.

In an exploratory examination of colonoscopy screening patterns observed in the SEER-Medicare data set, the hazard rate of the lag times is very flat (Figure 3), so an exponential distribution for the lag times was used in our analysis of lifetime colonoscopy screening patterns (i.e., $f(t) = \lambda \exp\{-\lambda t\}$). In the exponential case, $\overrightarrow{\phi}_j=(\lambda_{j1},\ldots,\lambda_{jj})$, as each of the j screenings has one associated parameter. The cumulative hazard $\Lambda_{jk}(t) = \lambda_{jk}t$. Let $\lambda_{jk} \sim$ Gamma (κ_{jk1}, κ_{jk2}), with the following posterior distribution:

$$\Pi(\lambda_{jk}|\mathbf{W},\boldsymbol{\eta},\overrightarrow{\kappa}_{jk}) \propto \prod_{i=1}^{n} p_{ij}\left(\overrightarrow{W}_i|\lambda_{jk},\alpha,M=j\right)^{\eta_{ij}} \lambda_{jk}^{\kappa_{jk1}-1}\exp\{-\lambda_j/\kappa_{jk2}\},$$

for $k = 1, \ldots, j$ and $j = 1, \ldots, \ell$.

4. Sample each element of $\overrightarrow{\kappa}_{jk}$, the prior parameters for φ_{jk} from the conditional posterior distributions. In the exponential case, $\kappa_{jk} = (\kappa_{jk1}, \kappa_{jk2})$, with $\kappa_{jk1} \sim Exp(b_{jk})$ and $\kappa_{jk2} \sim IG(c_{jk}, d_{jk})$, where $IG(.)$ is the Inverse Gamma distribution. In this instance, the conditional posterior distributions become:

$$\Pi\left(\kappa_{jk1}|\lambda_{jk},\kappa_{jk2},b_{jk}\right)\propto\left(\kappa_{jk2}^{\kappa_{jk1}}\Gamma(\kappa_{jk1})\right)^{-1}\lambda_{jk}^{\kappa_{jk1}-1}\exp\{-b_{jk}\kappa_{jk1}\}$$
$$\Pi\left(\kappa_{jk2}|\lambda_{jk},\kappa_{jk1},c_{jk},d_{jk}\right)\propto IG(\kappa_{jk1}+c_{jk},\lambda_{jk}+d_{jk})$$

5. Sample the correlation parameter, α, from the conditional posterior

$$\Pi(\alpha|\mathbf{W},\boldsymbol{\eta},\boldsymbol{\phi},\vec{\tau})\propto\prod_{i=1}^{n}\left(p_{ij}\left(\vec{W}_i|\vec{\phi}_j,\alpha,M{=}j\right)\right)^{\eta_{ij}}\times\alpha^{\tau_1-1}(1-\alpha)^{\tau_2-1},$$

where $\alpha \sim Beta(\tau_1, \tau_2)$.

6. Sample the prior parameters for α, τ_1 and τ_2, with both parameters distributed *Exp*(1), such that the conditional posterior distributions are

$$\Pi(\tau_1|\alpha,\tau_2)\propto\frac{\Gamma(\tau_1+\tau_2)}{\Gamma(\tau_1)}\alpha^{\tau_1-1}\exp\{-\tau_1\}$$
$$\Pi(\tau_2|\alpha,\tau_1)\propto\frac{\Gamma(\tau_1+\tau_2)}{\Gamma(\tau_2)}(1-\alpha)^{\tau_2-1}\exp\{-\tau_2\}$$

7. For subjects with an unknown number of lifetimes screenings (i.e., unknown *M*), sample the number of lifetime screenings using their corresponding probabilities π_{ij} and the parameter values of the $t^{(th)}$ iteration of the MCMC routine. After this value is sampled, update the unknown η_{ij} correspondingly.

3.7. Covariates

Following previous work performed by Ghitany and Maller [45] and others [30, 31, 45], covariates are added to the model by incorporating them into the parameter(s) of interest or by stratifying the model. The elements of the parameter vector $\vec{\phi}_j$ can be modeled using an appropriate link function relating the parameter(s) to the covariate(s) of interest. If $\vec{\phi}_j$ only has one element, then $\phi_{ij}{=}h^{-1}(Z_i'\vec{\omega}_j)$, where Z_i is the covariate vector for the i^{th} subject, $\vec{\omega}_j$ are the effects of the covariates on the lag time, and $h(\cdot)$ is an appropriate link function. (If $\vec{\phi}_j$ has more than one element, each element may be modeled with the same covariates and link function, or this may vary based on the constraints on the parameters in $\vec{\phi}_j$ and the biological rationale behind the covariate modeling.) In the exponential example, the link function $h(.)$ needs to be such that the parameter λ_j is positive. A natural function that ensures this is the exponential link function $\lambda_j{=}\exp\{Z_i'\vec{\omega}_j\}$, so that the resulting λ_j are all greater than 0.

In the univariate case, the probability that an individual i has a lifetime colonoscopy screening can be modeled as $\theta_{ij}{=}\text{expit}(X_i'\vec{\beta})$, where X_i is a $1 \times p$ covariate vector for the i^{th} subject and $\vec{\beta}{=}(\beta_1,\ldots,\beta_p)$ are the effects of the p covariates on the probability of a lifetime colonoscopy screening. The expit(.) function, defined as expit(a) = exp{a}/(1 + exp{a}), is used to ensure that the resulting θ_i estimates will be between 0 and 1. However,

because of the constraint that $\sum_{j=0}^{\ell}\theta_j=1$, in the instance of multiple screenings, it is easier to stratify the covariate and sample the θ values separately for each covariate level.

The likelihood function in Equation (4) can then be rewritten as follows:

$$L(\mathbf{W},\mathbf{X},\boldsymbol{Z},\boldsymbol{\eta},|\boldsymbol{\omega},\boldsymbol{\beta},\alpha)=\prod_{j=0}^{\ell}\prod_{s=1}^{\mathscr{S}}\prod_{i\in s}(\theta_{js})^{\eta_{ij}}\times\prod_{i=1}^{n}\prod_{j=0}^{\ell}\left(p_{ij}\left(\overrightarrow{W}_i|h^{-1}\left(Z_i'\overrightarrow{\omega}_j\right),\alpha,M=j\right)\right)^{\eta_{ij}},$$

where s is the covariate stratum with $s=1,\ldots,\mathscr{S}$, the total number of covariate strata. A similar Gibbs sampler routine to that presented in 3.6 is used for estimation. However, instead of sampling all $\overrightarrow{\theta}$, at once, the vectors are sampled by covariate strata from a Dirichlet posterior:

$$Dir\left(\left(\sum_i\eta_{i0s}+\gamma_{0s},\ldots,\sum_i\eta_{i\ell s}+\gamma_{\ell s}\right)^T\right).$$

For the $\overrightarrow{\omega}$ values, elements can be sampled from the posterior:

$$\Pi(\omega_{jk}|\boldsymbol{X},\boldsymbol{\eta},\overrightarrow{\omega}_j^-)=\prod_{i=1}^{n}\prod_{j=0}^{\ell}\left(p_{ij}\left(\overrightarrow{W}_i|h^{-1}\left(Z_i'\overrightarrow{\omega}_j\right),\alpha,M=j\right)\right)^{\eta_{ij}}\times\frac{(\omega_{jk}-\mu_{\omega_{jk}})^2}{2\sigma^2_{\omega_{jk}}}.$$

Each element of $\overrightarrow{\beta}_j$, $j=1,\ldots,\ell$ and $k=1,\ldots,p_j$, with $\omega_{jk}\sim\mathscr{N}\left(\mu_{\omega_{jk}},\sigma^2_{\omega_{jk}}\right)$.

4. Simulation studies

To determine the efficiency, accuracy, and consistency of our method and algorithm in the SEER-Medicare data context, we conducted a simulation study for the multivariate screening case. We set the maximum number of possible lifetime screenings at two ($\ell=2$), which is the maximum number of observed colonoscopy screenings in the SEER-Medicare data set and is a value consistent with medical practice in the oldest old. Data were generated varying the percentages of 0, 1, or 2 lifetime screenings, and assuming different lag times for subjects with only one screening when compared with subjects with two screenings. As is suggested by the SEER-Medicare data (Figure 3), we assumed an exponential distribution for the lag times. Under the exponential distribution for the lag times, the multivariate survival distribution for subjects with two screenings becomes

$$P\left(T_1>t_1,T_2>t_2|\overrightarrow{\lambda}_2,\alpha,M=2\right)=\exp\left\{-(\lambda_{21}t_1+\lambda_{22}t_2)^{\alpha}\right\}, \quad (16)$$

where $\overrightarrow{\lambda}_2=(\lambda_{21},\lambda_{22})$ and $\lambda_{21}t_1$ is the cumulative hazard for the lag time to the first of two screenings, and $\lambda_{22}t_2$ is the cumulative hazard for the lag time to the second of two

screenings. For subjects who only receive one lifetime screening, the survival function reduces to the standard survival function for the exponential distribution and is given by

$$P(T>t|\lambda_1, M=1)=\exp\{-\lambda_1 t\}.$$

To prevent the identifiability issues between the probability of a screening and the lag times (previously discussed in Section 3.4), we assumed that the maximum possible lag time was 10 years (as subjects who are overdue by more than 10 years are no longer 'average-risk' because of the rate of CRC progression [3]) and that subjects were only eligible for colonoscopy screenings between the age of 50 and 90 years. It is rare that a colonoscopy screening would be recommended for a patient over 90 years old because the risks associated with the colonoscopy screening procedure outweigh the long-term benefits of colonoscopy [2]. The length of the simulated study was 25 years. Three different lag time scenarios were used to generate data and are denoted as 'LT1', 'LT2', and 'LT3' (Table I). The three different lag time scenarios were paired with two different scenarios for the number of lifetime screenings (denoted as 'NLS1' and 'NLS2') and can be seen in Table II. The correlation parameter a was set at 0.9 (light correlation between screenings) as a biologically plausible value, consistent with descriptive summaries for SEER-Medicare data, where lag times when known are neither independent nor strongly correlated. The parameter values used to simulate data were chosen based on observed lag times in the SEER-Medicare data set and the possible true number of lifetime screenings. Paired together, there were six different types of simulated data sets, each containing 1000 subjects and generated 200 times. Left-censoring and right-censoring percentages ranged from approximately 40% to 50%. About 40% of subjects had at least one observed screening and about 15% of subjects had two observed screenings. In addition, we examined results for a model with covariates using simulations. The parameter values and results are outlined in Section 4.2.

Markov Monte Carlo (MCMC) chains were run on each data set, each for 50 000 iterations, with the first 10 000 iterations burned and thinned by 10 to reduce autocorrelation. Point estimates were calculated as the median of the marginal posterior distribution of each parameter.

4.1. Simulation results

Performance of the algorithm was assessed through the bias and the square root of the mean square error (RMSE). The bias was calculated as the average difference between the parameter estimate and the true value of the parameter, and the RMSE was calculated as the square root of the average squared difference between the parameter estimate and the true value of the parameter.

Table III presents the bias (RMSE) of the parameter estimates across all simulated data sets as percentages of the total parameter range (0 to 1 for a and all θ parameters, and 0–10 for median lag times). Across all data set variations, the average bias (RMSE) of the median lag times, $\tilde{Y}_{11}$, $\tilde{Y}_{21}$, and $\tilde{Y}_{22}$, were −1.75% (2.14%), −0.08% (0.85%), and 0.25% (0.64%) (corresponding to approximately 9.1 (11.1), 0.4 (4.4), and 1.3 (3.3) weeks) for the only lifetime screening, the first of two screenings, and the second of two screenings,

respectively. The bias (RMSE) for the percentage of screenings is also small, with values equal to −1.67% (2.49%), 1.69% (2.49%), and −0.05% (1.42%) for θ_0, θ_1, and θ_2, respectively. The largest bias and RMSE percentages were for the correlation parameter a, but numbers remained relatively small and ranged from about 4.5% to 6.3%.

4.2. Simulations with covariates

A subset of our simulations were performed incorporating a binary insurance covariate, representing whether the subject had colonoscopy coverage or not. For subjects who did have coverage, the probability or 0, 1, or 2 screenings was equal to 20%, 35%, and 45% respectively. For subjects who did not have coverage, these probabilities were 60%, 25%, and 15%. The lag times were modeled as a function of β_0 and β_1, such that

$$\lambda_{11}=\exp\{\beta_{0,11}+\beta_{1,11}I\{\text{coverage}\}\},$$

were $I\{\text{coverage}\}$ is an indicator variable denoting whether the individual had colonoscopy coverage or not. For the group with coverage, the average time to the only screening was 3 years, and the average time to the first and second of two screenings was 1.5 and 1 year. For the group without coverage, the average time to the only screening was 4 years, and the average time to the first and second of two screenings was 2 and 1 year. Results of the simulations can be observed in Table IV. As with the previous simulations, the bias and RMSE percentages were low for both covariate groups (less than 4%). The results for the $\vec{\beta}$ parameters were similar, with RMSE percentages ranging from 0.90% to 5.34%. The correlation parameter, a, behaved similarly to the previous set of simulations.

5. Application to SEER-medicare data

We applied our multivariate survival model to the SEER-Medicare data set to investigate colonoscopy screening behavior between 1991 and 2003, assuming the maximum number of lifetime screenings was equal to 2 (i.e., $\ell = 2$), as that is the maximum number we observed in our data set. This data set contains 403 842 individuals age 65 years or older at study entry after the removal of subjects who used other CRC screening methods (such as fecal occult blood tests or sigmoidoscopy) or who did not have average-risk screening patterns (such as very frequent colonoscopy screenings). Individuals were considered eligible for screening colonoscopy in 1991; while current screening guidelines recommend screening starting at age 50, very few people received colonoscopies before the early 1990s (as the USPSTF did not even provide official guidelines until 1996 [2]), and therefore the probability of an unobserved screening on an average-risk individual occurring before 1991 is very small. Among these subjects, 62% were left-censored, with average left-censoring times equal to 4.5 years (range: 0.1–9.9 years). In addition, 22% had one observed colonoscopy and 0.11% had two observed colonoscopies. Among individuals who had at least one colonoscopy, the median lag time before the first observed screening was 5.7 years (range: 0.01–10 years), and among individuals with two colonoscopies, the median lag time before the second observed screening was 0.1 years (range: 0.01–2.7 years). An approximated hazard rate showed the rate of screening was constant (Figure 3), so we assumed an exponential distribution for all lag times.

We examined both univariate models (estimating the time to the first screening), including a covariate to account for insurance coverage levels, as well as multivariate models estimating parameters for zero to two screenings. The univariate models were created to provide initial estimates of the median time to the first screening (regardless of whether it was the only screening or the first of two), as well as the probability an individual would never be screened. In addition, we were able to include an insurance level coverage covariate in the univariate model, which allowed us to quantify the effects of at least some insurance coverage on the probability of receiving at least one lifetime colonoscopy screening. We then examined multivariate models with two possible lifetime screenings. This allowed us to quantify differences between the lag time to the first lifetime screening and the lag time to the first of two lifetime screenings. We were also able to examine if the lag time to the first screening was longer or shorter than the lag time to the second screening. Among both the univariate and multivariate models, we compared the results when individuals were eligible for screening until age 75 or eligible for screening until age 80, which are commensurate with current screening guidelines [2]. In the multivariate model, we assumed a maximum lag time of 10 years to prevent issues with identifiability.

Three separate MCMC chains were run for each model for a total of 50 000 iterations, each with a burn-in of 10 000 and thinned by 10 to reduce autocorrelation, leaving a total of 4000 thinned iterations in each chain for each parameter for analysis. Convergence was determined through the Geweke diagnostic [46], graphical diagnostics (such as trace plots and density plots), and Gelman–Rubin tests [47, 48]. Point estimates were calculated as the median of the posterior marginal distributions for each parameter, and 95% central credible intervals were used for inference.

5.1. Univariate model results

We first examined univariate models, which provided us with initial estimates on the probability of never receiving a colonoscopy screening and the time to the first screening (based on the likelihood function in equation 5). The simple univariate model shows that before age 75 years, approximately 38% (95% CI: 37.6–37.9%) of the Medicare population gets a colonoscopy screening, with a median lag time (calculated based on λ) equal to about 5.2 years (95% CI: 5.16–5.20 years). These numbers change slightly when the maximum screening age is raised to age 80 years with slight increases in screening rates as well as increases in median lag times (Table V).

To determine the impact of changes in levels of insurance coverage for colonoscopy screenings (i.e., differences between coverage phase 0, phase 1, and phase 2) on colonoscopy screening rates, we included a covariate in the estimate of θ in the following manner:

$$\theta = \text{expit}(\beta_0 + \beta_1 I\{\text{study entry after} 1998\}).$$

In the covariate model, the baseline group (represented by β_0) were subjects who became eligible for screening when no colonoscopy coverage was offered, and β_1 represents the change in this probability when some or all coverage was available. (A covariate was not

included in the parameter for the lag time, as there was not enough information to reliably run the Gibbs Sampler for this particular data set.) Results from the covariate model show that screening rates increased almost 15 percentage points for subjects age 75 years and younger when at least some insurance coverage was offered, and increased almost 17 percentage points for subjects 80 and younger when at least some insurance coverage was offered. These results show that providing at least some insurance coverage for colonoscopy screenings dramatically improves the rate of screening (Table V).

Figure 4 shows the estimated 'survival curves' (i.e., the probability of no lifetime screening colonoscopy) for subjects who have no colonoscopy insurance coverage compared with subjects who have some or all colonoscopy insurance coverage. In our analysis of CRC screening, a higher survival curve indicates a worse screening pattern (i.e., lower numbers and longer lag times), and it can be observed that (not surprisingly) the subjects who had no colonoscopy coverage had lower rates of screening. Among patients eligible for screening up to age 75 years, 26% of patients without coverage were screened by age 60 years and 33% of patients without coverage were screened by age 70 years. However, when at least some coverage for colonoscopy was available, 36% of patients were screened by age 60 years and 45% of patients were screened by age 70 years. This can also be observed in Figure 5, which shows the densities of the survival curves (i.e., probability of no colonoscopy screening) for subjects at age 55, 60, and 70 years. The figure shows that at age 55, the two densities are the closest together, and each density is narrow. However, by age 70, the densities are farther apart from each other, meaning that differences in screening patterns between subjects with and without insurance coverage become bigger with increasing age. Note that in all three graphs, the densities do not overlap, providing evidence that the probability of never being screened via colonoscopy is statistically significantly different when subjects have some insurance coverage compared with those who have no insurance coverage.

5.2. Multivariate model results

In the multivariate case, we examined the case of two maximum possible lifetime screenings, as we had no individuals with three or more observed colonoscopies in our data set, and we assumed both lag times were distributed exponentially. No covariates were included in the multivariate model; by the nature of the multiple screenings model, the lag times and screening percentages at different points in the study inherently include temporal changes such as insurance coverage levels.

Results show that up to age 75, the probability of never being screened is approximately 68% (95% CI: 67.6–67.9%). The probability of one lifetime screening is about 27% (95% CI: 26.8–27.0%), and the probability of two lifetime screenings is about 5% (95% CI: 5.3% – 5.4%) . Among subjects who are only screened once, the estimated median lag time is 2.9 years (95% CI: 2.89–2.92 years). Among subjects who are screened twice, the median lag time for the first screening is 1 year (95% CI: 1.01–1.07 years), and the median lag time for the second screening is 1.6 years (95% CI: 1.57–1.63 years) . The parameter α, which represents the correlation between screenings, is equal to 0.92 (95% CI: 0.912–0.923), which means the within-subject correlation between screenings is low. Numbers changed

little when subjects were eligible for screening up to age 80 (see Table VI for model results). Note that the probability of never being screened through colonoscopy before age 75 is similar in the multivariate and univariate models. Lag times between the univariate and multivariate models differ because of the unrestricted possible maximum lag time in the univariate models.

Estimated marginal survival curves (i.e., the probability of not being screened) for each of the lag times of the multivariate model are shown in Figure 6. As with the univariate model, high survival curves indicate a poor screening rate. On the left, it is observed that the number of two or more screenings is very low, and differences between the first and second of two screenings are minor. The number of individuals receiving one lifetime colonoscopy screening is higher, but still poor. On the right, the survival curves are again shown, but *conditional* on getting one or two lifetime screenings (i.e., θ is not used in the calculation of the survival curve). These estimates show that among individuals who get two screenings, the time to the first screening is shorter than the time to the second screening, with 96% of these subjects acquiring the first screening within 5 years of becoming due, and 88% of subjects acquiring the second screening within 5 years of becoming due. Subjects who only get one lifetime colonoscopy screening waited longer, with 77% of these subjects getting screened within the first 5 years of becoming due. Our results show that while the actual rates of screening are poor, those who are getting screened are diligent, with a large majority of individuals getting screened within 5 years after becoming due.

Figure 7 shows the posterior distributions for the probability of at least one screening by age 55, 65, and 75 years. We see that by age 55, about 25% of individuals have had at least one screening. By age 65, at least 31.5% of individuals have had at least one screening, and by age 75, at least 32% of individuals have had at least one screening. The densities of the probability of at least one screening are very narrow, and the curve for 55-year old subjects is far from the other two curves, indicating there is a significant difference in the probabilities. There is less difference between the 65-year-olds and 75-year-olds.

Figure 8 shows the bivariate survival distribution in a contour plot. While the figure is fairly symmetric (meaning there is little difference between the lag time to the first screening and second screening), the gray shading extends slightly higher up the y-axis (the axis that denotes the time to the second screening), which means that the time to the second screening is delayed slightly longer when compared with the first screening. The bivariate survival distribution is only shown for the first 5 years, as the probabilities for years 5 through 10 are very small, and it is difficult to discern differences in the distribution after this time point.

6. Discussion

We have proposed a cure rate model for multivariate survival data that can account for both left-censored and right-censored data. We have demonstrated theory that works for the general case of multiple lifetime screenings and then applied it to the case of two colonoscopy screenings. The case of two colonoscopy screenings in a lifetime is common, but more than that is perhaps unrealistic, as beyond a certain age the risks outweigh the long-term benefits of screenings and are often not recommended in the later stages of life.

Our approach and sampling of missing information allows us to provide robust estimates with narrow credible intervals, even in the difficult setting of considerable left-censoring and right-censoring and with the inclusion of subjects who never get screened. In addition, our estimates are sufficiently accurate to detect both demographic differences and the time-varying impact of policy shifts.

Using this method, we have shown that many individuals are never being screened for CRC, with overall estimates of at least one screening at only 30%. However, screening behavior was dramatically improved following increases in Medicare payments, with an estimated reduction in the probability of never being screened for CRC of around 15% or more when colonoscopy coverage was provided. These results agree with previous work, which has shown that screening incidence is generally low but can be improved with increased levels of coverage [49, 50]. In addition, among subjects who do get screened, they are diligent and do not wait long periods of time after becoming due for a screening. We have extended these results to quantify the exact rates of incidence and how adherent individuals are to current screening guidelines.

Future work with this model and the SEER-Medicare data set includes linking lifetime screening behavior to cancer incidence rates, as well as the inclusion of other screening modalities, such as sigmoidoscopy and fecal occult blood testing (FOBT). This link will greatly inform the debate on optimal screening guidelines, as well as improve current cost-benefit analyses of CRC screening and Medicare expenditures.

We have only presented simulations for the case of two lifetime screenings, which are reasonable for analysis of the SEER-Medicare data set. The extension to three or more lifetime screenings is more difficult computationally, although it can be performed with time and care. Another extension would be an exploration of the assumption that the frailty term follows a positive stable distribution, such as that found in Qiou *et al.* [51], or using Bayesian modeling to sample the frailty parameter at every iteration of the MCMC sampling routine. In addition, we make the assumption that the right-censoring process is independent of the screening process, which is likely true for most individuals, but it would be interesting to examine how this assumption impacts the parameter estimates using strategies like imputation (for example, see Daniels and Hogan [52]) for dependently censored data.

Our model not only has answered very important questions about CRC screening behavior but it also has broad applicability to situations with multiple events where there may be patterns unobserved before study entry or after study exit. These types of analyses will become more prevalent as time progresses, particularly with major changes in health care coverage because of the Affordable Care Act. Accurate assessment of patterns of lifetime preventive medical care will become more necessary as government-funded health care becomes more prevalent, and this information is required to determine the effectiveness of different medical procedures.

Acknowledgments

This study used the linked SEER-Medicare database. The interpretation and reporting of these data are the sole responsibility of the authors. The authors acknowledge the efforts of the National Cancer Institute; the Office of

Research, Development and Information, CMS; Information Management Services (IMS), Inc.; and the Surveillance, Epidemiology, and End Results (SEER) Program tumor registries in the creation of the SEER-Medicare database.

Dr Beckett acknowledges the support from the UC Davis Cancer Center Support Grant, P30CA093373-06. Statistical support was made possible by Grant number UL1 RR024146 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NCRR or NIH. Information on Re-engineering the Clinical Research Enterprize can be obtained from http://nihroadmap.nih.gov/clinicalresearch/overview-translational.asp).

The authors would like to thank Dr Joshua Fenton at the UC Davis Medical Center for his guidance and support with this research.

References

1. American Cancer Society. Cancer facts and figures. 2014.
2. Pignone M, Rich M, Teutsch S, Berg A, Lohr K. Screening for colorectal cancer in adults at average risk: a summary of evidence of the U.S. Preventive services task force. Annals of Internal Medicine. 2002; 137(2):E132–E141.
3. Winawer S, Schottenfeld D, Flehinger B. Colorectal cancer screening. Journal of the National Cancer Institute. 1992; 83:243–251. [PubMed: 1994053]
4. Seeff L, Nadel M, Klabunde C, Thompson T, Shapiro J, Vernon S, Coates R. Patterns and predictors of colorectal cancer test use in the adult U.S. population. Cancer. 2004; 100:2093–2103. [PubMed: 15139050]
5. Smith R, Cokkinides V, von Eschenbach A, Levin B, Cohen C, Runowicz C, Sener S, Saslow D, Eyre H. American Cancer Society guidelines for the early detection of cancer. CA: A Cancer Journal for Clinicians. 2002; 52:8–22. [PubMed: 11814067]
6. Gross C, Andersen M, Krumholz H, McAvay G, Proctor D, Tinetti M. Relation between medicare screening reimbursement and stage at diagnosis for older patients with colon cancer. Journal of the American Medical Association. 2006; 296(23):2815–2822. [PubMed: 17179458]
7. Levin T, Zhao W, Conell C, Self L, Manninen D, Shapiro J, Schulman J. Complications of colonoscopy in an integrated health care delivery system. Annals of Internal Medicine. 2006; 145:880–886. [PubMed: 17179057]
8. Dominitz J, Eisen G, Baron T, Goldstein J, Hirota W, Jacobson B, Johanson J, Leighton J, Mallery J, Raddawi H, Vargo J II, Waring J, Fanelli R, Wheeler-Harbough J, Faigel D. Complications of colonoscopy. Gastrointestinal Endoscopy. 2003; 57:441–445. [PubMed: 12665750]
9. Vijan S, Hwang E, Homer T, Hayward R. Which colon cancer screening test? A comparison of costs, effectiveness, and compliance. The American Journal of Medicine. 2001; 111:593–601. [PubMed: 11755501]
10. National Cancer Institute (NCI). Seer-medicare linked database. 2015. Available at: http://appliedresearch.cancer.gov/seermedicare/ [Accessed on 7 jan 2015]
11. Boag J. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. Journal of the Royal Statistical Society Series B (Methodological). 1949; 11(1):15–33.
12. Berkson J, Gage R. Survival curve for cancer patients following treatment. Journal of the American Statistical Association. 1952; 47:501–515.
13. Haybittle J. The estimation of the proportion of patients cured after treatment for cancer of the breast. British Journal of Radiology. 1959; 32:725–733. [PubMed: 14400306]
14. Haybittle J. A two-parameter model for the survival curve of treated cancer patients. Journal of the American Statistical Association. 1965; 60:16–26.
15. Farewell V. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics. 1982; 38:1041–1046. [PubMed: 7168793]
16. Goldman A. Survivorship analysis when cure is a possibility: a Monte Carlo study. Statistics in Medicine. 1984; 3:153–163. [PubMed: 6463452]
17. Maller R, Zhou S. Estimating the proportion of immunes in a censored sample. Biometrika. 1992; 79:731–739.

18. Sposto R, Sather H, Baker S. A comparison of tests of the difference in the proportion of patients who are cured. Biometrics. 1992; 48:87–99. [PubMed: 1581495]
19. Wei L, Lin D, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. Journal of the American Statistical Association. 1989; 84:1065–1073.
20. Liang K, Self S, Chang Y. Modelling marginal hazards in multivariate failure time data. Journal of the Royal Statistical Society (Series B). 1993; 55:441–453.
21. Lin D. Cox regression analysis of multivariate failure time data: the marginal approach. Statistics in Medicine. 1994; 13:2233–2247. [PubMed: 7846422]
22. Prentice R, Hsu L. Regression on hazard ratios and cross ratios in multivariate failure time analysis. Biometrika. 1997; 84:349–363.
23. Spiekerman C, Lin D. Marginal regression models for multivariate failure time data. Journal of the American Statistical Association. 1998; 93:1164–1175.
24. Hougaard P. Modelling multivariate survival. Scandinavian Journal of Statistics. 1987; 14:291–304.
25. Hougaard P. A class of multivariate failure time distributions. Biometrika. 1986; 73(3):671–678.
26. Hougaard P. Life table methods for heterogeneous populations: distributions describing the heterogeneity. Biometrika. 1984; 71:75–83.
27. Hougaard P. Survival models for heterogeneous populations derived from stable disributions. Biometrika. 1986; 73:387–96.
28. Hougaard P. Frailty models for survival data. Lifetime Data Analysis. 1995; 1:255–273. [PubMed: 9385105]
29. Hougaard, P. Analysis of Multivariate Survival Data. Springer; New York: 2000.
30. Chen M, Ibrahim J, Sinha D. Bayesian inference for multivariate survival data with a cure fraction. Journal of Multivariate Analysis. 2002; 80:101–126.
31. Ibrahim, J.; Chen, M.; Shinha, D. Bayesian Survival Analysis. Springer; New York: 2001.
32. Clayton D. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978; 65:141–151.
33. Oakes D. Bivariate survival models induced by frailties. Journal of the American Statistical Association. 1989; 84:487–493.
34. Oakes D. Multivariate survival distributions. Journal of Nonparametric Statistics. 1994; 3:343–354.
35. Chi Y, Ibrahim J. Joint models for multivariate longitudinal and multivariate survival data. Biometrics. 2006; 62:432–445. [PubMed: 16918907]
36. Hastings W. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970; 57:97–109.
37. Gilks, W.; Richardson, S.; Spiegelhalter, D. Markov Chain Monte Carlo in Practice. Chapman and Hall; Great Britain: 1996.
38. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1984; 6:721–741. [PubMed: 22499653]
39. Gilks W, Wild P. Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society Series C (Applied Statistics). 1992; 41(2):337–348.
40. Sy J, Taylor J. Estimation in a Cox proportional hazards cure model. Biometrics. 2000; 56(1):227–236. [PubMed: 10783800]
41. Farewell V. Mixture models in survival analysis: are they worth the risk? The Canadian Journal of Statistics. 1986; 14:257–262.
42. Li C, Taylor J, Sy J. Identifiability of cure models. Statistics and Probability Letters. 2001; 54:389–395.
43. Yu B, Tiwari R, Cronin K, Feuer E. Cure fraction estimation from the mixture cure models for grouped survival data. Statistics in Medicine. 2004; 23:1733–1747. [PubMed: 15160405]
44. Gelman, A.; Carlin, J.; Stern, H.; Rubin, D. Bayesian Data Analaysis. Chapman and Hall; Boca Raton, FL: 1995.

45. Ghitany M, Maller R, Zhou S. Exponential mixture models with long-term survivors and covariates. Journal of Multivariate Analysis. 1994; 49:218–241.

46. Geweke, J. Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments in Bayesian Statistics 4. Clarendon Press; Oxford, UK: 1992.

47. Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. Statistical Science. 1992; 7:457–511.

48. Brooks S, Gelman A. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics. 1998; 7:434–455.

49. Zapka J, Puleo E, Vickers-Lahti M, Luckmann R. Healthcare system factors and colorectal cancer screening. American Journal of Preventive Medicine. 2002; 23:28–35. [PubMed: 12093420]

50. Messina C, Lane D, Grimson R. Colorectal cancer screening attitudes and practices. American Journal of Preventive Medicine. 2005; 28:439–446. [PubMed: 15894147]

51. Qiou Z, Ravishanker N, Dey D. Multivariate survival analysis with positive stable frailties. Biometrics. 1999; 55:637–644. [PubMed: 11318227]

52. Daniels, M.; Hogan, J. Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis. CRC Press; Boca Raton, FL: 2008.

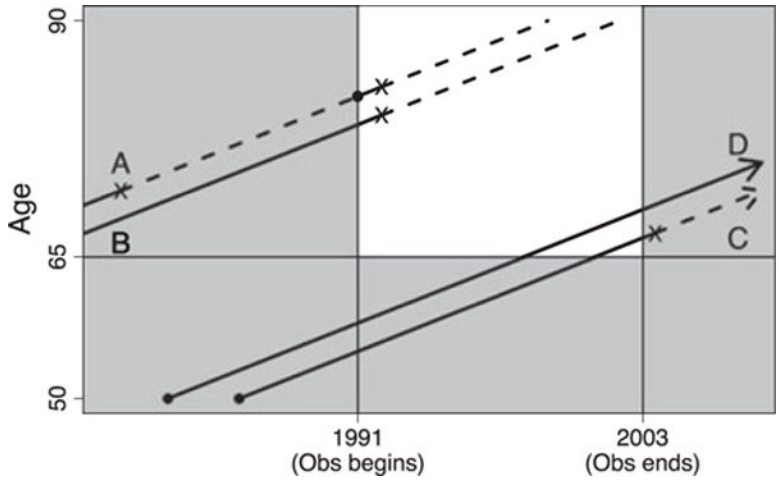

Figure 1.

Four hypothetical lifetime colonoscopy screening trajectories and the observed information (unshaded region) available from the SEER-Medicare data set. In the figure, a circle denotes the time an individual becomes due for a screening (occurs either at age 50 or 10 years after the previous screening), and an 'X' denotes the time of a colonoscopy screening. The solid lines represent periods of time where the subject is overdue for a colonoscopy (called a 'lag time'), and the dashed lines represent the 10-year period during which average-risk subjects are not due for a screening. Among the hypothetical trajectories, subject *A* is screened twice, once before 1991 and then again in the observation window. Subject *B* has only one screening, which is observed. However, based on information provided in the observation window, it is not possible to tell if the lifetime trajectory for subject *A* is different than it is for subject *B.* Similarly, neither subject *C* or *D* has an observed screening, but subject *C* does get a colonoscopy screening after the study ends, while subject *D* does not. In this example, all four hypothetical subjects are both left-censored and right-censored.

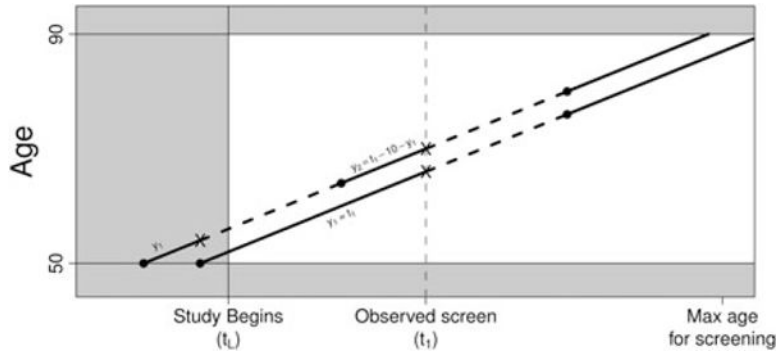

Figure 2.

The figure shows two possible trajectories for a left-censored subject who enters the the observation period at time t_L and has one observed screening at time t_1. Lag times are denoted with solid lines, and represent periods of time when the subject is due for a screening, and the 10-year post-screening period is denoted with a dashed line. Screenings are marked with 'X', and time points when the subject becomes due for a screening are marked with a circle. In this example, it is possible that (1) the one observed screening is the only lifetime screening or (2) the subject may have had two lifetime screenings, the first one occurring before the left-censoring time t_L and the second being observed. In the first case, the length of the first lag time, y_1, is equal to the time to the first screening, t_1, and can be written as $P(Y_1{=}t_1|\overrightarrow{\phi}_1, M{=}1)$. In the second case, we can only determine that the time to the first screening, y_1, is less than the left-censoring time, and the second lag time, y_2, is the remaining period of time between the observed screening and the previous screening ($t_1 - 10 - y_1$). This can be written as $P(Y_1{<}t_L, Y_2{=}t_1 - 10 - y_1|\overrightarrow{\phi}_2, \alpha, M{=}2)$. This subject is not right-censored because he or she reaches the maximum screening age before the observation period ends.

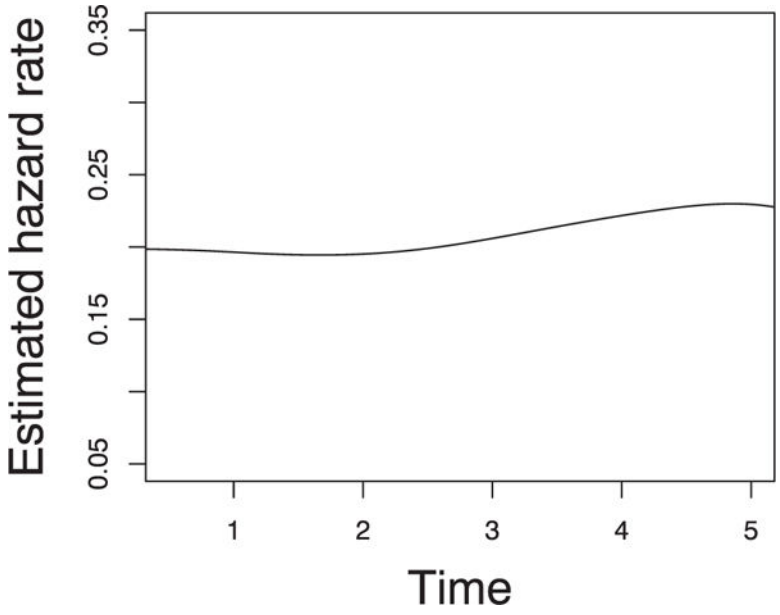

Figure 3.

Approximated hazard rate of the time to the first screening between years 1991 and 1996 (before Medicare insurance coverage changes or guidelines were set). The hazard rate is very flat, providing evidence that an exponential parametric distribution is appropriate. The hazard rate is approximated dividing the number of observed failures by the number of subjects at risk (provided by survfit() in R) and then smoothed using smooth.spline() in R.

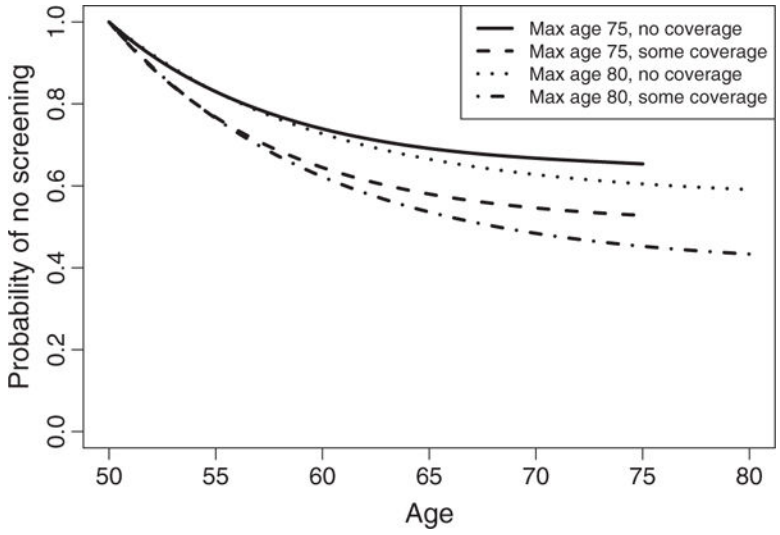

Figure 4.

Univariate model results showing estimated survival curves (i.e., the probability of not being screened via colonoscopy) for the time to the first screening, comparing subjects with no colonoscopy insurance coverage to those with at least some colonoscopy insurance coverage. In the colonoscopy screening context, a high survival curve indicates a poor rate of screening. It can be observed that the subjects with the higher survival curve are those without colonoscopy coverage.

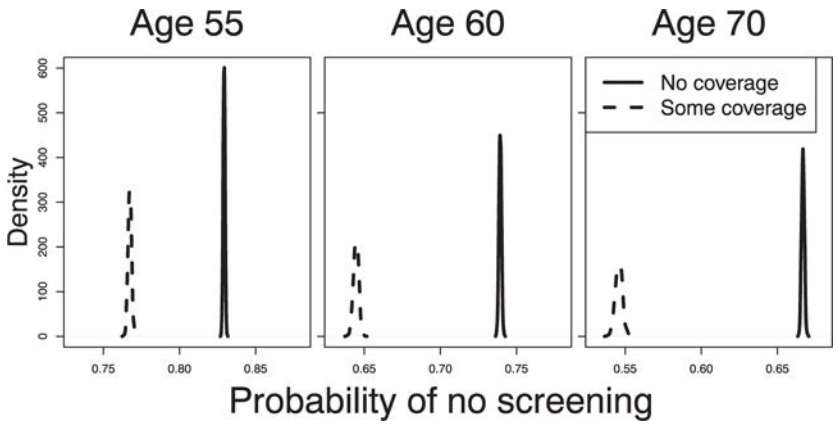

Figure 5.

Univariate model results showing the densities for the probability of not being screened for colorectal cancer via colonoscopy by age 55, 60, and 70 years, comparing those with no coverage (solid line) and those with at least some coverage (dashed line). Note that while the *x*-axes on all three graphs cover a different range of probability values, the size of the range is 15% for all three graphs. We can observe that the two densities are closest together at 55 years and are farthest apart at 70 years. Note that in all three figures, the densities do not overlap, providing evidence that the probability of screening under some and no coverage is statistically significantly different across the different time points. The densities are calculated using the density() function in R on the MCMC chain of survival probabilities, calculated at each iteration of the thinned and burned chains. Results are shown for the model with a maximum screening age of 75 years.

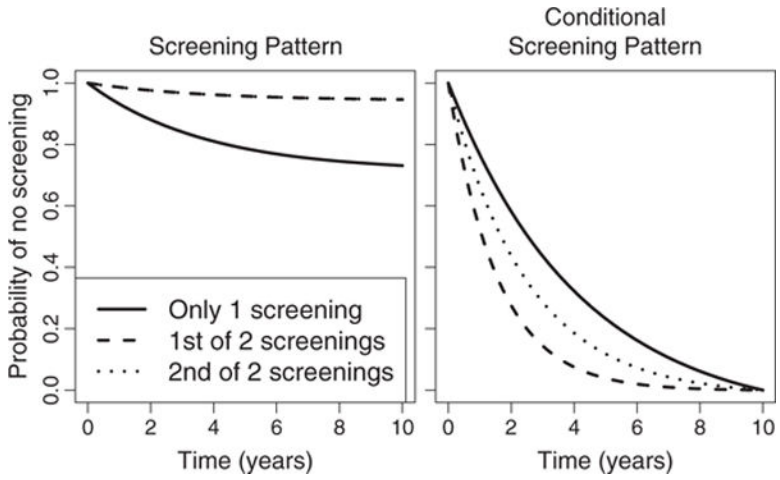

Figure 6.

Left: Multivariate model results showing the marginal estimated survival curves (i.e., probability of not being screened for colorectal cancer via colonoscopy) for the only lifetime screening (solid line) or the first and second of two screenings (dashed and dotted lines). As expected, more people get only one screening in a lifetime rather than two lifetime screenings, and therefore that survival curve is lowest. Right: The estimated survival curves conditional on the number of lifetime screenings. These curves show that among subjects who will receive two screenings, the first screening happens quickly when compared with the second screening. Subjects who only get one colonoscopy take the longer than those who get two colonoscopies. As with the univariate models, a high survival curve indicates poor screening rates. Estimates were calculated from the multivariate model that assumes 75 is the maximum eligible age for screening.

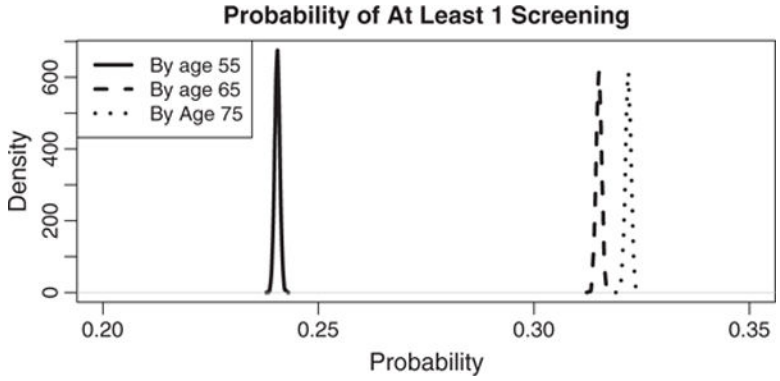

Figure 7.

Left: Multivariate model results showing the densities of the probability of getting at least one screening before the age of 55, 65, or 75 years. (Densities calculated using the density() function in R.) The densities are narrow, and the probability of at least one screening for 55-year-olds is far from the other two age groups, indicating that there may be a statistically significant difference in the probability of screening for individuals who are 55 years old when compared with 65- or 75-year-olds. The probability does not increase as rapidly between age 65 and 75 years. Estimates were calculated from the multivariate model that assumes 75 is the maximum eligible age for screening.

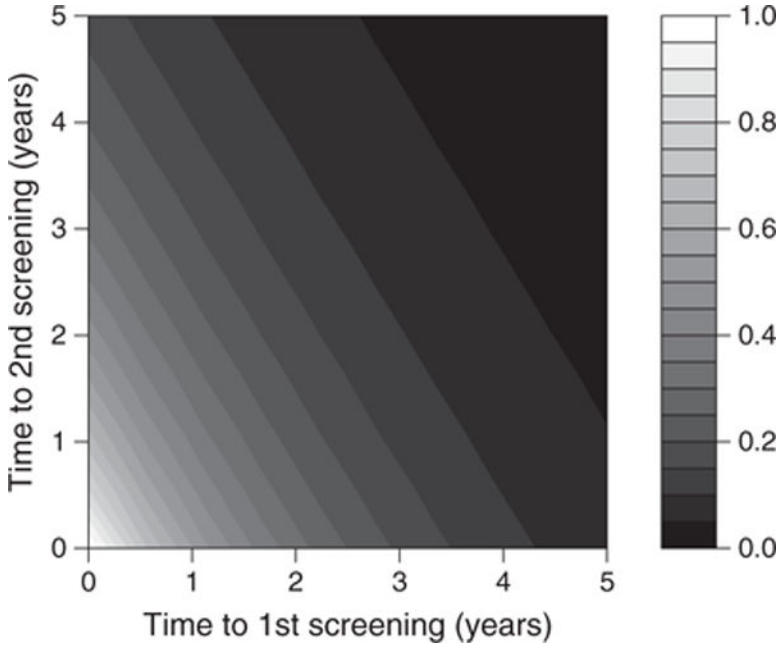

Figure 8.

A contour plot of the multivariate model results showing the joint survival curve (i.e., the probability of not being screened via colonoscopy), conditional on subjects who get two screenings (i.e., the probability of two screenings, θ_2, is not used in the calculation of the survival probabilities) for the first 5 years. The figure shows minor differences between the time to the first screening and time to the second screening, as the contour plot is fairly symmetric. However, the gray shading extends slightly higher up the *y*-axis (which represents the time to the second screening), meaning that the probability of not being screened is higher for a longer time period before the second screening. As with the univariate models, a high survival curve indicates poor screening rates. Years 5 through 10 are omitted from the figure as the probabilities are very small and difficult to discern in the figure. Estimates were calculated from the multivariate model that assumes 75 is the maximum eligible age for screening.

Table I

Three different lag time ('LT') scenarios used to generate data sets used in our simulation study. The lag times shown are the median lag times (in years) from the survival distribution that is used to generate the lag time to colonoscopy screenings for individuals who are screened at least once in their lifetime.

	Only one screening	Two screenings	
Scenario	**Lag time (λ_{11})**	**1st Lag time (λ_{21})**	**2nd Lag time (λ_{22})**
LT1	4.3 (0.02)	1 (0.70)	1 (0.70)
LT2	3.5 (0.09)	4/3 (0.50)	2/3 (1.05)
LT3	2.25 (0.35)	4/3 (0.50)	2/3 (1.05)

Table II

Two different scenarios for the number of lifetime screenings ('NLS') used to generate data sets of the number of lifetime screenings (0, 1, or 2) for each subject in the simulated data sets.

Scenario	Probability of 0 screenings (θ_0)	Probability of 1 screening (θ_1)	Probability of 2 screenings (θ_2)
NLS1	1/3	1/3	1/3
NLS2	1/2	1/4	1/4

Table III

Summary bias and RMSE results for the simulation studies across all data sets (not including covariates) as percentages of the total parameter value (0–1 for $\vec{\theta}$ and a parameters and 0–10 for median lag times). Results of the simulations show the estimates of the screening rates to be close to the true parameter values, with RMSE values less than or equal to 3% for all possible values of θ and median lag times (calculated as a function of $\vec{\lambda}$ at each MCMC iteration), and less than or equal to 6% for the a parameter, which had the highest bias and RMSE percentages.

		LT1		LT2		LT3		Total (over LTs)	
Model		**Bias (%)**	**RMSE (%)**	**Bias (%)**	**RMSE (%)**	**Bias (%)**	**RMSE (%)**	**Bias (%)**	**RMSE (%)**
	θ_0	−2.09	2.66	−1.01	1.94	−2.22	2.77	−1.77	2.49
	θ_1	2.19	2.82	0.91	1.94	2.12	2.73	1.74	2.53
	θ_2	−0.14	1.44	0.06	1.44	0.06	1.45	−0.01	1.44
NLS1	$\tilde{Y}_{11}$	−2.15	2.42	−1.01	1.42	−1.95	2.22	−1.70	2.07
	$\tilde{Y}_{21}$	0.33	0.72	−0.32	0.80	−0.32	0.82	−0.10	0.78
	$\tilde{Y}_{22}$	0.15	0.69	0.29	0.50	0.29	0.57	0.24	0.59
	a	4.52	5.30	5.04	5.71	5.11	5.73	4.89	5.58
	θ_0	−2.07	2.72	−0.82	2.10	−1.85	2.59	−1.58	2.48
	θ_1	2.17	2.70	0.89	2.06	1.84	2.53	1.63	2.44
	θ_2	−0.14	1.42	−0.11	1.36	−0.02	1.40	−0.09	1.39
NLS2	$\tilde{Y}_{11}$	−2.21	2.51	−1.11	1.64	−2.09	2.39	−1.80	2.21
	$\tilde{Y}_{21}$	0.30	0.83	−0.31	0.97	−0.17	0.93	−0.06	0.91
	$\tilde{Y}_{22}$	0.16	0.80	0.30	0.61	0.31	0.60	0.26	0.68
	a	5.26	5.94	5.59	6.31	5.64	6.33	5.50	6.20
	θ_0	−2.08	2.69	−0.91	2.02	−2.03	2.69	−1.67	2.49
	θ_1	2.18	2.76	0.90	2.00	1.98	2.63	1.69	2.49
	θ_2	−0.14	1.43	−0.02	1.40	0.02	1.42	−0.05	1.42
Total (over θ)	$\tilde{Y}_{11}$	−2.18	2.46	−1.06	1.53	−2.02	2.31	−1.75	2.14
	$\tilde{Y}_{21}$	0.31	0.77	−0.31	0.89	−0.25	0.87	−0.08	0.85
	$\tilde{Y}_{22}$	0.16	0.75	0.30	0.56	0.30	0.59	0.25	0.64
	a	4.89	5.63	5.31	6.01	5.37	6.04	5.19	5.90

Note: RMSE, root-mean-square error; MCMC, Markov chain Monte Carlo; LT, lag time; NLS, number of lag time screenings.

Table IV

Summary estimates, 95% credible intervals, bias, and RMSE results for the simulation studies across the 200 data sets that included an insurance coverage covariate. Estimates are calculated as the median of the marginal posterior distribution for each parameter. Bias and RMSE values are shown as percentages of the total parameter value (0–1 for $\vec{\theta}$ and a parameters, and varying ranges (approximately −6 to 6) for the $\vec{\beta}$ parameters. Bias and RMSE percentages are not shown for $\vec{\beta}_1$ parameters because the range is $(-\infty, \infty)$ for those parameters. Results of the simulations show the estimates of the screening rates to be close to the true parameter values, with RMSE values less than or equal to 4% for all possible values of θ and less than or equal to 6% for the a and $\vec{\beta}$ parameters.

	Parameter	(true value)	Estimate	(95% Interval)	Bias%	(RMSE%)
	θ_0	(0.6)	0.58	(0.52, 0.63)	−2.12	(3.45)
	θ_1	(0.3)	0.27	(0.22, 0.32)	2.02	(3.23)
	θ_2	(0.2)	0.15	(0.12, 0.18)	0.04	(1.60)
No insurance coverage	$\beta_{0,11}$	(−1.35)	−1.32	(−1.41, −1.14)	4.57	(5.34)
	$\beta_{0,21}$	(−0.73)	−0.68	(−0.94, −0.43)	0.68	(0.90)
	$\beta_{0,22}$	(−0.05)	0.01	(−0.26, 0.28)	−0.37	(2.09)
	θ_0	(0.2)	0.19	(0.15, 0.24)	−0.79	(2.15)
	θ_1	(0 35)	0.36	(0.31, 0.41)	0.62	(2.49)
	θ_2	(0.45)	0.45	(0.40, 0.49)	0.09	(2.17)
Insurance coverage	$\beta_{0,11}$	(0.29)	0.23	(−0.01, 0.49)		
	$\beta_{0,21}$	(0.35)	0.30	(0.01, 0.59)		
	$\beta_{0,22}$	(0.00)	0.02	(−0.28. 0.35)		
	a	(0.90)	0.94	(0.89, 1.00)	4.57	(5.34)

Note: RMSE, root-mean-square error.

Table V

Univariate model results showing the median and 95% credible intervals (calculated as the 2.5% and 97.5% of the MCMC chain for each parameter) of the posterior distributions for the probability of never being screened for colorectal cancer (θ) and the median lag time (as a function of λ) to the first screening for colorectal cancer in the SEER-Medicare data set.

Univariate		Cap at age 75		Cap at age 80	
Model	**Parameter**	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Simple model	P (not being screened)	0.623	(0.621, 0.624)	0.554	(0.551, 0.556)
	Median time to screening	5.166	(5.133, 5.199)	6.532	(6.486, 6.576)
Covariate model	P (not being screened), no coverage	0.639	(0.637, 0.641)	0.570	(0.568, 0.572)
	P (not being screened), some coverage	0.508	(0.502, 0.513)	0.404	(0.398, 0.411)
	Median time to screening	4.179	(4.172, 4.185)	6.896	(6.844, 6.945)

Note: MCMC, Markov chain Monte Carlo; CI, confidence interval; SEER, Surveillance, Epidemiology and End Results.

Table VI

Estimates and 95% credible intervals (calculated as the median and 2.5% and 97.5% of the marginal posterior for each parameter) for the probability of receiving none, one, or two screenings in a lifetime and the median time to the only screening or the first and second of two screenings for colorectal cancer in the SEER-Medicare data set, as well as the parameter α, which represents the correlation between screenings. Results are similar regardless of the maximum eligible age for screening.

	Cap at age 75		Cap at age 80	
Covariate	**Estimate**	**95% CI**	**Estimate**	**95% CI**
Probability of no screenings	0.677	(0.676, 0.679)	0.686	(0.685, 0.688)
Probability of one screening	0.269	(0.268, 0.270)	0.224	(0.222, 0.225)
Probability of two screenings	0.054	(0.053, 0.054)	0.090	(0.089, 0.091)
Median time to only 1 screening	2.905	(2.886, 2.924)	3.176	(3.154, 3.202)
Median time to 1st of two screenings	1.038	(1.010, 1.067)	1.270	(1.249, 1.293)
Median time to 2nd of two screenings	1.597	(1.571, 1.625)	1.564	(1.541, 1.589)
Correlation parameter α	0.917	(0.912, 0.923)	0.960	(0.956, 0.965)

Note: CI, confidence interval; SEER, Surveillance, Epidemiology and End Results.