Published in final edited form as:

Cancer Epidemiol Biomarkers Prev. 2009 March ; 18(3): 712–717. doi:10.1158/1055-9965.EPI-08-0866.

Analysis of Colorectal Adenoma Recurrence Data Subject to Informative Censoring

Chiu-Hsieh Hsu^{1,2}, Jeremy M. G. Taylor³, Qi Long⁴, and David S. Alberts¹

¹Arizona Cancer Center, College of Medicine, University of Arizona, Tucson AZ 85724

²Division of Epidemiology and Biostatistics, Mel and Enid Zuckerman College of Public Health, University of Arizona, Tucson AZ 85724

³Department of Biostatistics, University of Michigan, Ann Arbor MI 48109

⁴Department of Biostatistics, Emory University, Atlanta GA 30322

Abstract

The treatment effect of a colorectal polyp prevention trial is often evaluated from the colorectal adenoma recurrence status at the end of the trial. However, early colonoscopy from some participants complicates estimation of the final study end recurrence rate. The early colonoscopy could be informative of status of recurrence and induce informative differential follow-up into the data. In this paper we use mid-point imputation to handle interval-censored observations. We then apply a weighted Kaplan-Meier method to the imputed data to adjust for potential informative differential follow-up, while estimating the recurrence rate at the end of the trial. In addition, we modify the weighted Kaplan-Meier method to handle a situation with multiple prognostic covariates by deriving a risk score of recurrence from a working logistic regression model and then use the risk score to define risk groups to perform weighted Kaplan-Meier estimation. We argue that mid-point imputation will produce an unbiased estimate of recurrence rate at the end of the trial under an assumption that censoring only depends on the status of early colonoscopy. The method described here is illustrated with an example from a colon polyp prevention study.

Keywords

current status data; mid-point imputation; weighted Kaplan-Meier estimator

Introduction

Colorectal cancer is one of the most common malignancies in the United States, with a projected 148,610 new cases and 68,840 deaths from this disease in 2006 (1). It is believed that most colorectal carcinomas arise from adenomas (2). Hence, most of colorectal cancer prevention trials use colorectal adenomas to study preventive agents. For a typical colorectal polyp prevention trial, individuals who had undergone removal of a colorectal adenoma within six months prior to study are randomly assigned to the treatment group or the matched placebo group and treated for three or more years. The treatment effect is then evaluated on occurrence of newly discovered adenomas by performing colonoscopy at follow-up to remove all new colorectal polyps and then check if any of the polyps is adenomatous. The follow-up

Reprint requests should be addressed to: Chiu-Hsieh Hsu Arizona Cancer Center PO Box 245024 1515 N Campbell Tucson AZ 85724
–5024 Phone (520) 626–5054 Fax: (520) 626–2467 phsu@azcc.arizona.edu.

This original work has not been presented or submitted elsewhere.

colonoscopy is scheduled to be performed once at the end of the trial (e.g. at three years after start of the intervention) to evaluate the recurrence status. The actual event time for each participant is then only known as occurring either before three years or not. A reasonable statistical method to analyze this type of data will be logistic regression, which simply analyzes the binary outcomes, i.e. recurrence status at the end of the trial.

Because of family history of colorectal cancer and other health conditions, some participants could have their only follow-up colonoscopy before three years (the scheduled time) or even have more than one follow-up colonoscopy. The follow-up colonoscopy performed before three years is considered as “early colonoscopy”. As a result, the participants with early colonoscopy could have differential follow-up lengths in contrast to those participants who adhered strictly to the schedule of follow-up colonoscopy. Due to differential follow-up, the recurrent adenoma data can be considered as current status data (“case 1” interval-censored data), in which each participant was observed only once and the recurrence time was known only to be either less than or greater than the observation time. In addition to interval-censored observations, there could be right censored observations as well if a participant did not have any recurrent adenomas at all follow-up colonoscopies. Logistic regression could produce a biased estimate of the recurrence rate at the end of the study when there are right censored observations existing before the end of the study (3).

A potential estimator is the nonparametric maximum likelihood estimator (NPMLE) (4,5), which is often used to analyze interval-censored data. The NPMLE requires a computationally intensive algorithm to obtain measures of uncertainty. Our previous work modified logistic regression using a weight function, a function of follow-up length, to propose a less computationally intensive approach to account for censoring due to differential follow-up, while still estimating the recurrence rate at the end of the trial (3). Both the weighted logistic regression and the NPMLE approach assume a non-informative differential follow-up, i.e. the reasons a participant had early colonoscopy are not associated with risk of recurrence. However, often the reasons (e.g. family history of colorectal cancer and previous polyp history) are associated with the risk of recurrence and, therefore, induce informative differential follow-up. When the early colonoscopy is informative of risk of recurrence, a method, e.g. the weighted logistic regression approach, which simply treats informative differential follow-up as non-informative, could produce a biased estimate of the recurrence rate at the end of the trial.

In this paper, we assume that every participant with early colonoscopy had the same distribution of follow-up colonoscopy and propose to use mid-point imputation to simplify interval-censored data to right-censored data. We then combine this with the weighted Kaplan-Meier (WKM) estimator (6,7), which incorporates prognostic factors into survival analysis, to adjust for informative differential follow-up, to estimate the recurrence rate at the end of the trial. The mid-point imputation approach is attractive because it does not require a distribution for the imputation. Such a distribution would be hard to estimate in a typical colorectal polyp prevention trial because often over 50% of the participants had their only follow-up colonoscopy performed at the end of trial and those participants provided very little information with regard to their actual time of recurrence. Although it has been shown that mid-point imputation could produce biased survival estimates (8-10), we will show it will not produce a biased estimate when the primary interest is in estimating the recurrence rate at the end of the trial. In addition, we will also modify the WKM method, which is limited to a situation with a single categorical prognostic factor, to handle a situation with multiple prognostic factors. We propose to modify the WKM method by deriving a risk score of recurrence from a working logistic regression model with all prognostic factors as the covariates and then categorize the risk score to perform the WKM method. The risk score summarizes the complex structure of prognostic factors into a scalar. In this paper, we will focus on estimating the recurrence rate at the end of the trial and are interested in comparing the performance of the WKM method

derived from the mid-point imputed data with NPMLE under a situation of informative early colonoscopy. In addition, logistic regression (Logit), which is often used for this type of study, and weighted logistic regression (WLogit) will be also explored.

This paper is organized as follows. In the methods section, we show the theoretical properties of midpoint imputation in estimating the recurrence rate at the end of the trial, review and describe the WKM estimator derived from mid-point imputed data for interval-censored data and modify the WKM estimator to handle a situation with multiple prognostic covariates. In the application section, we apply the WKM method to a data set from a ursodeoxycholic acid colorectal polyp prevention (UDCA) study. In the discussion section, we discuss the performance and limitations of the proposed WKM estimator for estimating the recurrence rate at the end of a colon polyp prevention trial.

Methods

Notation

Let X denote the time to first occurrence of a polyp after randomization, T_k denote the k^{th} follow-up colonoscopy time, where $k=1, \dots, K$, τ (e.g. three years) denote maximum follow-up time, M denote an early colonoscopy indicator variable showing whether the first follow-up colonoscopy was performed before the end of the trial (i.e. $I(T_1 < \tau)$) and Z denote a baseline covariate. We only know X falls in some interval (L, R) , where $L < X < R$. Right censoring is equivalent to $R = \infty$. Let (L_i, R_i) denote the observable random interval, (l_i, r_i) denote the observed time interval and $\Delta_i = I(R_i < \infty)$ denote the recurrence indicator for each subject. Suppose there are n participants in a study. The observed data are thus $O = (l_1, r_1, \delta_1, m_1, z_1), \dots, (l_n, r_n, \delta_n, m_n, z_n)$. We assume that these n subjects come from a random sample and are independent. Each participant could have only one follow-up colonoscopy (i.e. $K=1$) or more than one follow-up colonoscopy (i.e. $K>1$). For a participant with only one follow-up colonoscopy, the recurrence time is either right-censored at $L=T_1$ or interval-censored into an interval $(0, R=T_1)$, where $T_1=\tau$ if $M=0$ and $T_1<\tau$ if $M=1$. Of the participants with multiple colonoscopy, some could have their final follow-up colonoscopy at the end of the trial (i.e. $T_K=\tau$). Therefore, the recurrence time is either right-censored at $L=T_K$ (last follow-up colonoscopy time) or interval-censored into an interval (L, R) , where $L < X < R = T_K \leq \tau$.

For participant i with an interval-censored recurrence time, i.e. (l_i, r_i) , where $r_i \leq \tau$, mid-point imputation is used to impute time to recurrence by $(l_i + r_i)/2$, the mid-point of (l_i, r_i) . For participant j with a right-censored recurrence time, i.e. $r_j = \infty$, time to recurrence is treated as right censored at l_j , where $l_j \leq \tau$. Let X^* denote the observed time to recurrence derived from mid-point imputation. That is, $X^* = X$ if $\delta = 0$ and $X^* = (l + r)/2$ if $\delta = 1$. The KM and WKM estimates can then be derived from the imputed data set.

Properties of mid-point imputation

The survival rate at the end of the trial based on the imputed data set can be written as

$$\begin{aligned} S^*(\tau) &= P(X^* > \tau) = P(X^* > \tau | M=0) P(M=0) + P(X^* > \tau | M=1) P(M=1) \\ &= P(X > \tau | M=0) P(M=0) + P(X > \tau | M=1) P(M=1) \\ &= P(X > \tau) = S(\tau) \end{aligned}$$

Under a condition that some of the participants with multiple colonoscopy could have their final follow-up colonoscopy at the end of the trial (i.e. $P(T_K = \tau | M=1, K>1) > 0$), the third equality holds because mid-point imputation only affects $S(x)$, where $x < \tau$. Therefore, under an assumption that differential censoring only depends on the status of early colonoscopy, i.e. M , the imputed data X^* can be used to give an unbiased estimate of the recurrence rate at the

end of the trial. This provides a theoretical foundation for using mid-point imputed data to replace the interval-censored data and can be generalized to handle a situation that censoring depends on more than one covariate, when the main interest is in estimating the recurrence rate at the end of the trial.

Weighted Kaplan-Meier (WKM) estimator

For illustration we assume Z is a categorical covariate and takes on values $1, \dots, K$. The survival function derived from the imputed data can be written as,

$$S^*(t) = P(X^* > t) = \sum_{k=1}^K P(X^* > t | Z=k) P(Z=k) = \sum_{k=1}^K S_k^*(t) \theta_k$$

where θ is the probability a subject has covariate value k , ($k=1, \dots, K$) and $S_k^*(t)$ is the probability of survival conditional on having covariate value k . Based on the above expression, the WKM

estimator (6,7) is defined as $WKM(t) = \sum_{k=1}^K \widehat{S}_k^*(t) \frac{n_k}{n}$, where $\widehat{S}_k^*(t)$ is the KM estimator among those with covariate value k , n_k is the number of subjects with covariate value k , and

$n = \sum_{k=1}^K n_k$. The recurrence rate at time t is then equal to $1 - WKM(t)$. The associated variance is equal to the sum of the weighted averages of within-variation and between-variation (see below),

$$\text{var}(\sqrt{n}WKM(t)) = \sum_{k=1}^K \theta_k S_k^*(t)^2 \int_0^t \frac{\lambda_k(u) d(u)}{H_k(u) S_k^*(u)} + \sum_{k=1}^K \theta_k \left(S_k^*(t) - \bar{S}^*(t) \right)^2$$

where $\lambda(\cdot)$ and $H(\cdot)$ are hazard and cumulative hazard functions, respectively. The first term of the variance can be easily estimated by calculating the weighted average of the variances derived from the Greenwood's formula for those K groups and the second term can be estimated by plugging in estimates of each component (7).

In a situation with early colonoscopy, regarding M as the only covariate, the WKM estimator (denoted as WKM^C) at the end of the trial can be expressed as,

$$WKM^C(\tau) = \frac{n_e}{n} \widehat{S}_e^*(\tau) + \frac{n_{ne}}{n} \widehat{S}_{ne}^*(\tau)$$

where $\widehat{S}_e^*(\tau)$ and $\widehat{S}_{ne}^*(\tau)$ are the KM estimators for the participants with early colonoscopy and without early colonoscopy, respectively, n_e and n_{ne} are the number of the participants with early colonoscopy and without early colonoscopy, respectively, and $n = n_e + n_{ne}$. Because all of the participants without early colonoscopy had their only follow-up colonoscopy at the end of the trial, $\widehat{S}_{ne}^*(\tau)$ reduces to $1 - \widehat{p}_{ne}$, where $\widehat{p}_{ne}$ is the sample proportion of recurrence among those participants. All of the participants with early colonoscopy had their first follow-up colonoscopy before τ . If the largest observed time among those participants was censored, we propose to complete the tail of $\widehat{S}_e^*$ by an exponential curve to estimate S_e^* (11).

WKM with multiple covariates

In addition to early colonoscopy, often there is more than one covariate associated with the risk of recurrence in a colorectal polyp prevention trial. They could be either categorical or

continuous. Let $Z=(z_1,...,z_p)$ denote the p covariates associated with risk of recurrence. The WKM method, because it requires a single categorical covariate, cannot directly incorporate those p covariates into estimation. In order to use the information from the covariates to improve the marginal survival estimate, Hsu et al. (12) considered a situation of possibly multiple time-independent or time-dependent continuous covariates and proposed deriving risk scores. These risk scores summarize the associations between the covariates and the failure and censored times, from two working proportional hazards (PH) models, one for the failure time and one for the censoring time. By incorporating predictive covariates into survival analysis, one can both increase efficiency and reduce bias due to dependent censoring of the estimate of the marginal survival distribution.

In this paper we adapt and modify the ideas in Hsu et al. (12) to incorporate multiple covariates into the WKM method. We propose to fit a working logistic regression model for recurrence of the form, $\text{logit}[Pr(\Delta=1)|Z]=Z^*\beta$, to reduce the covariates to a risk score, which provides an indicator of an individual's risk of recurrence. We propose to fit one working model for the participants with early colonoscopy and one for the participants without early colonoscopy, since we believe they might have a different association between the covariates and risk of recurrence. The risk scores are then defined as $\widehat{RS}_e=Z\widehat{\beta}_e$ for the participants with early colonoscopy and $\widehat{RS}_{ne}=Z\widehat{\beta}_{ne}$ for the participants without early colonoscopy, where $\widehat{\beta}_e$ and $\widehat{\beta}_{ne}$ denote the estimates of the regression coefficients for the logistic regression models. The risk scores will be continuous and can be categorized into groups based on dichotomization or quartiles. The WKM estimator can then be easily derived based on the categorized groups of the risk scores for both participants with or without early colonoscopy (denoted as WKM_e and WKM_{ne} , respectively). The WKM method using both early colonoscopy and prognostic

covariates of recurrence is denoted as $WKM^{R+C}(\tau)=\frac{ne}{n}WKM_e(\tau)+\frac{n_{ne}}{n}WKM_{ne}(\tau)$. Note if there is only one prognostic covariate, then there is no need to fit the working logistic regression model.

Both WKM^C and WKM^{R+C} treat the status of early colonoscopy, i.e. M , as a baseline covariate rather than a post hoc variable and might underestimate the variabilities of estimators of recurrence rate. In addition, midpoint imputation simplifies the complexity of the interval-censored data and Greenwood's variance formula is approximate and known to underestimate the variance, especially with heavy censoring and in the right tail of the survival distribution. Hence, we propose to use the bootstrap technique to estimate the standard errors of the estimators for WKM^C and WKM^{R+C} .

Application to UDCA data

In 1996, the Arizona Cancer Center initiated a multi-center trial to determine whether ursodeoxycholic acid (UDCA) can prevent the recurrence of colorectal adenomas (13). A total of 1285 subjects with identified colorectal adenomas at the qualifying examination were recruited and randomly assigned to one of the two treatment groups, placebo and UDCA (8–10 mg/kg/day). Of 1,285 subjects, a total of 1,192 subjects underwent at least one follow-up colonoscopy and were thus considered for the endpoint analysis, 579 in the placebo group and 613 in the UDCA group. For each of the 1,192 subjects, his/her recurrent status was measured, as well as the baseline covariates, such as age (mean: 66.2; standard deviation: 8.5), gender (67.4% male), BMI (mean: 27.4; standard deviation: 4.6), family history of colorectal cancer (27.4% with family history of colorectal cancer) and previous polyp history (before the qualifying examination) (47.3% with previous polyp history). According to the baseline covariates, on average the UDCA participants were slightly overweight and had a higher risk of recurrence compared to the general population.

Initially, the follow-up colonoscopy was planned to be performed only once, no earlier than six months before the three-year anniversary date after randomization (i.e. 30 months). However, some participants went through their follow-up colonoscopy before the planned time. The number of participants who had early colonoscopy are 233 (40.2%) in the placebo group and 260 (42.4%) in the UDCA group. Some of those participants had multiple follow-up colonoscopies. Table 1 displays the frequency and recurrence results of follow-up colonoscopy for the participants with early colonoscopy and indicates that the participants with multiple follow-up colonoscopy tended to have their first follow-up colonoscopy performed earlier compared to those who had only one follow-up colonoscopy. Of the participants with multiple follow-up colonoscopy (N=327), 297 (90.8%) had at least one follow-up colonoscopy at least 30 months after randomization and 138 (42.2%) had at least one recurrent adenoma at their first follow-up colonoscopy. Based on Table 1, a participant could have recurrent adenomas at the first colonoscopy and no recurrent adenomas at the second colonoscopy. This is because at each colonoscopy the participant's colorectal polyps were removed and tested to see if any of them is adenomatous. Instead of fixing the end of the trial exactly at three years, for each participant the actual time of the colonoscopy was used to define the interval of time to first recurrence. The midpoint imputation method was then conducted on the interval-censored observations.

Table 2 explores the covariates associated with having early colonoscopy and risk of recurrence. According to the table, early colonoscopy is highly associated with risk of recurrence and marginally associated with gender (male). The participants with early colonoscopy had a significantly higher recurrence rate (49.3%) compared to the participants without any early colonoscopy (37.5%) with an odds ratio of 1.621 and a 95% confidence interval (CI) of (1.283, 2.048). This indicates informative early colonoscopy for the UDCA study. Age, BMI, gender, early colonoscopy and previous polyp history are significantly associated with risk of recurrence. In this paper, we calculate the WKM^C estimator using early colonoscopy as the only covariate and the WKM^{R+C} estimator using both early colonoscopy and a prognostic categorical covariate derived from a linear combination (risk scores) of the covariates (age, gender, BMI and previous polyp history), which are associated with risk of recurrence. The risk scores of recurrence are obtained by fitting logistic regression models for the participants with and without early colonoscopy separately. Each risk score is then dichotomized into two groups (low vs. high) to calculate the WKM estimate. We repeated the analyses using 4 groups instead of 2 and it gave similar results.

In this paper we are interested in estimating the recurrence rate of adenomas at three years for both the placebo and UDCA groups based on the UDCA study protocol. A sample proportion of recurrence (Logit), WLogit (with an exponential weight function truncated at three years (3)), NPMLE, WKM^C and WKM^{R+C} methods are calculated from the data. The results are provided in Table 3. Logit produces a lower recurrence rate compared to the NPMLE and WKM methods for both placebo and UDCA groups. This supports our previous findings (3). WLogit produces a higher recurrence rate compared to the other methods (Logit, NPMLE and WKM) for both placebo and UDCA groups. The WKM^C method, which incorporates early colonoscopy directly into the analysis to adjust for informative differential follow-up, produces a slightly higher recurrence rate for the placebo group and a similar recurrence rate for the UDCA group compared to the NPMLE method. This results in a lower odds ratio 0.747 with a 95% CI of (0.737, 0.992). The WKM^{R+C} method, which incorporates both early colonoscopy and prognostic covariates into analysis, produces a much higher recurrence rate for the placebo group and a similar recurrence rate for the UDCA groups compared to the NPMLE method. As a result, the WKM^{R+C} method yields the lowest odds ratio 0.725 with a 95% CI of (0.541, 0.981) compared to the other methods. The 95% CI for both WKM^C and WKM^{R+C} does not cover one and indicates that UDCA is associated with a lower risk of recurrence in contrast to the results from the NPMLE, Logit and WLogit methods. In addition, we observe a counter-

intuitive phenomenon that the WKM^{R+C} method has a slightly higher estimate of standard error compared to the WKM^C method. This could be due to one (previous polyp history) of the covariates used in the WKM^{R+C} method having missing observations and, as a result, a smaller dataset (only the non-missing data) used for the WKM^{R+C} method compared to the WKM^C method or because the WKM^{R+C} method gives a more accurate estimate of standard error compared to the WKM^C method. In summary, the WKM method could provide an adjustment for informative differential follow-up due to early colonoscopy. We also perform simulations to investigate the properties of the WKM methods. The simulation results will be published as supplementary material, including supplementary tables 1 and 2.

Discussion

The research in this paper uses mid-point imputation to handle interval-censored observations and then combines with a weighted Kaplan-Meier approach to adjust for informative early colonoscopy through the use of the status of early colonoscopy and gains efficiency by incorporating prognostic covariates of recurrence, when estimating the recurrence rate at the end of the trial for a colorectal polyp prevention trial. This approach can handle a situation with multiple prognostic covariates by deriving a risk score from a logistic regression model. Although the idea of this approach might appear simple, the results based on simulations (not provided here) do show that the weighted Kaplan-Meier approach can provide a reasonable recurrence rate estimate under an informative differential follow-up and can gain efficiency when prognostic covariates exist. In contrast, the conventional statistical methods such as logistic regression, which simply ignores differential follow-up, and the WLogit and NPMLE methods, which depend on the assumption of non-informative differential follow-up, could produce biased estimates. Hence, the method that does not incorporate informative differential follow-up into estimation of the recurrence rate could produce misleading conclusions as indicated in the data analysis section.

In this paper, we treat the early colonoscopy status as known at baseline and use it directly to define risk groups to perform weighted Kaplan-Meier estimator to adjust for informative differential follow-up while estimating the recurrence rate. This might appear to be unrealistic and underestimate variability of the weighted Kaplan-Meier estimator. However, based on the guidelines for screening colorectal cancer (14), the chance of a participant having early colonoscopy during follow up was very likely to be decided by his/her family history of colorectal cancer, previous polyp history and baseline polyp characteristics (e.g. size ≥ 1 cm), which were known at baseline. Hence, treating the early colonoscopy status as known at baseline might not be unrealistic. In addition, the risk score approach in this paper can be generalized to handle a situation that the early colonoscopy status and time to the first follow-up colonoscopy are treated as endpoints and known to be associated with some baseline covariates (e.g. family history, previous polyp history and baseline polyp characteristics) where a working proportional hazards model can be fitted to time to the first follow-up colonoscopy data to derive a risk score to summarize the association between time to the first follow-up colonoscopy and the covariates. This risk score and the risk score derived from the working logistic regression model for recurrence can then be jointly used to define the risk groups to perform the weighted Kaplan-Meier estimator, instead of using the status of early colonoscopy directly to define the risk groups. Although the approach is motivated by the data from colorectal polyp prevention trials, the WKM approach can also be generalized to handle the data from other types of clinical studies where each participant is only scheduled to be followed for a fixed standard time period instead of regular monitoring throughout the study and the main interest is in estimating the event rate at the end of the study.

Simply using mid-point imputation to handle interval-censored observations highly depends on the lengths of intervals and might produce biased survival estimates and misleading

results⁹, especially at early time points. However, in this paper, we focus on estimating the recurrence rate at the end of the trial and have shown that mid-point imputation will not produce a biased estimate of recurrence rate at the end of the trial under an assumption that censoring only depends on the status of early colonoscopy.

The recurrence rate at the end of the study corresponds to the tail of the survival curve. It is well recognized that the survival estimate in the tail is often unstable. Hence, one might feel that survival analysis techniques do not seem good choices for estimating the recurrence rate at the end of the study under this setting. However, in a colorectal polyp prevention trial often over 50% of participants had their only follow-up colonoscopy at the end of the trial. Those participants were either interval-censored or right-censored at the end of the study. For those participants, mid-point imputation is less likely to contribute additional variation to the estimate of recurrence rate at the end of the study and the information they provide towards estimation of recurrence rate at the end of the study simply reduces to a binary outcome and their follow-up lengths provide little information with regard to the actual recurrence time. We suspect this might stabilize the tail problem.

In a situation with multiple covariates, we fit a working logistic regression model for recurrence to reduce multiple covariates into a scalar. The model is only used as a convenience in calculating the risk scores to create a categorical variable, which is predictive of risk of recurrence, to implement the weighted Kaplan-Meier method. More sophisticated and computationally intensive approaches for fitting the working model could be used, such as a proportional hazard model for interval-censored data, but we suspect that would not lead to a significant reduction in the bias, which is the major concern under a situation with informative differential follow-up for the WKM method. In addition, parametric assumptions connected with the statistical model are only employed to define the risk scores. As a result, the reliance on the logistic regression model is weaker for the WKM approach. However, the performance of the WKM method using predictive covariates of recurrence to improve efficiency in estimation of the recurrence rate in an informative differential follow-up situation will depend on the strength of the association between these covariates and recurrence.

The research in this paper assumes that every participant with early colonoscopy had the same distribution of follow-up colonoscopy. However, the distribution of follow-up colonoscopy might depend on a participant's health condition or family history of colorectal cancer. In addition, the research mainly focuses on estimating the recurrence rate at the end of the study. Often one of the main interests is in testing the prevention effect. Future research can focus on developing approaches that can handle complex distributional assumptions for the follow-up colonoscopy and perform two sample tests, as well.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by National Institute of Health (CA23074; CA41108) and American Cancer Society (IRG7400128).

Reference

1. Jemal A, Siegel R, Ward E, et al. Cancer statistics 2006. *CA: a Cancer Journal for Clinicians* 2006;56:106–30. [PubMed: 16514137]
2. Leslie A, Carey FA, Pratt NR, Steele RJ. The colorectal adenoma-carcinoma sequence. *British Journal of Surgery* 2002;89:845–60. [PubMed: 12081733]

3. Hsu C-H, Green SB, He Y. A weighted logistic regression model for estimation of recurrence of adenomas. *Statistics in Medicine* 2007;26:1567–78. [PubMed: 16850435]
4. Peto R. Experimental survival curves for interval-censored data. *Applied Statistics* 1973;22:86–91.
5. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society* 1976;38:290–5. Series B
6. Malani HM. A modification of the redistribution to the right algorithm using disease markers. *Biometrika* 1995;82:515–26.
7. Murray S, Tsiatis AA. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics* 1996;52:137–51. [PubMed: 8934589]
8. Rucker G, Messerer D. Remission duration: an example of interval-censored observations. *Statistics in Medicine* 1988;7:1139–45. [PubMed: 3201039]
9. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* 1992;11:1569–78. [PubMed: 1439361]
10. Odell PM, Anderson KM, D'Agostino RB. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* 1992;48:951–9. [PubMed: 1420849]
11. Brown, JB.; Hollander, M.; Korwar, RM. Nonparametric tests of independence for censored data, with application to heart transplant studies.. In: Proschan, F.; Serfling, RJ., editors. *Reliability and Biometry: Statistical Analysis of Lifelength*. SIAM; Philadelphia: 1974. p. 327-54.
12. Hsu C-H, Taylor JMG, Murray S, Commenges D. Survival analysis using auxiliary variables via nonparametric multiple imputation. *Statistics in Medicine* 2006;25:3503–17. [PubMed: 16345047]
13. Alberts DS, Martinez ME, Hess LM, et al. Phase III Trial of Ursodeoxycholic Acid To Prevent Colorectal Adenoma Recurrence. *Journal of the National Cancer Institute* 2005;97:846–53. [PubMed: 15928305]
14. Barclay L. New guidelines for colorectal cancer screening. *Gastroenterology* 2003;124:544–60. [PubMed: 12557158]

Table 1
Summary of follow-up colonoscopy for the participants with early colonoscopy.

No. of colonoscopy	N(%)	T ₁ (months) [*]	T _K (months) [†]	No. of (T _K ≥30) [§]	Recurrence Results [‡]
1	166 (33.7)	22.4	22.4	0	-; 118; +; 48
2	246 (49.9)	10.6	43.2	217	(-,-): 117; (-,+): 37 (+,-): 57; (+,+): 35
3	60 (12.2)	9.6	59.0	59	(-,-,-): 12; (+,-,-): 11 (-,-,+): 9; (-,-,+): 2 (+,-,-): 11; (+,-,+): 9 (-,-,+): 3; (+,-,+): 3
4	13 (2.6)	8.8	58.0	13	(-,-,-,-): 2 (+,-,-,-): 1; (-,-,-,-): 1 (-,-,-,-): 1; (+,-,-,-): 2 (+,-,-,-): 1; (-,-,-,-): 1 (+,-,-,-): 3; (+,-,-,-): 1
5	8 (1.6)	8.7	67.2	8	(-,-,-,-,-): 1; (-,-,-,-,-): 1 (-,-,-,-,-): 1; (-,-,-,-,-): 1 (+,-,-,-,-): 1; (+,-,-,-,-): 1 (+,-,-,-,-): 1; (+,-,-,-,-): 1

^{*} median time to the first follow-up colonoscopy
[†] median time to the last follow-up colonoscopy
[§] Number of participants with follow-up colonoscopy after 30 months
[‡] -; no recurrent adenomas detected; +; at least one recurrent adenoma detected

Table 2

Univariate analysis for the UDCA study based on logistic regression

Early Colonoscopy			
Variable	Odds ratio	95% CI	p-value
Age	1.011	(0.997,1.025)	0.13
BMI	1.001	(0.976,1.027)	0.92
Male	0.809	(0.633,1.033)	0.09
Treatment	1.094	(0.868,1.378)	0.45
Recurrence	1.621	(1.283,2.048)	0.00
Previous Polyp History	1.026	(0.809, 1.300)	0.84
Family History of CRC*	1.013	(0.783,1.312)	0.92
Risk of Recurrence			
Age	1.022	(1.008,1.036)	0.00
BMI	1.031	(1.006,1.058)	0.02
Male	1.381	(1.077,1.770)	0.01
Treatment	0.887	(0.705,1.117)	0.31
Early Colonoscopy	1.621	(1.283,2.048)	0.00
Previous Polyp History	1.319	(1.041,1.671)	0.02
Family History of CRC*	1.080	(0.835,1.396)	0.56

* Colorectal cancer

Table 3
Estimation of recurrence rate for placebo and UDCA groups at three years*

Method	Placebo		UDCA		Odds Ratio (95% CI)
	Estimate	Standard Error	Estimate	Standard Error	
NPMLE	0.467	0.031	0.414	0.032	0.807 (0.545, 1.136)
Logit	0.439	0.020	0.409	0.020	0.887 (0.705, 1.122)
WLogit	0.539	0.023	0.504	0.023	0.869 (0.674, 1.157)
WKM ^C	0.488	0.026	0.416	0.022	0.747 (0.558, 0.972)
WKM ^{R+C} [†]	0.496	0.027	0.416	0.023	0.725 (0.541, 0.981)

* Standard errors and 95% CI were derived from 500 bootstrap samples

[†] 65 missing observations for previous polyp history