*The Annals of Applied Statistics*
2020, Vol. 14, No. 2, 1030–1044
https://doi.org/10.1214/20-AOAS1342


# ANALYSES OF PREVENTIVE CARE MEASURES WITH INCOMPLETE HISTORICAL DATA IN ELECTRONIC MEDICAL RECORDS: AN EXAMPLE FROM COLORECTAL CANCER SCREENING

BY YINGYE ZHENG[1], DOUGLAS A. CORLEY[2], CHYKE DOUBENI[3], ETHAN HALM[4], SUSAN M. SHORTREED[5], WILLIAM E. BARLOW[6], ANN ZAUBER[7], TOR DEVIN TOSTESON[8], AND JESSICA CHUBAK[5]

[1]*Department of Biostatistics, Fred Hutchinson Cancer Research Center, yzheng@fredhutch.org*
[2]*Division of Research, Kaiser Permanente Northern California, douglas.corley@kp.org*
[3]*Department of Family Medicine, Mayo Clinic School of Medicine, Doubeni.Chyke@mayo.edu*
[4]*Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical School, Ethan.Halm@UTSouthwestern.edu*
[5]*Kaiser Permanente Washington Health Research Institute, Susan.M.Shortreed@kp.org*
[6]*Cancer Research and Biostatistics, williamb@crab.org*
[7]*Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, zaubera@MSKCC.ORG*
[8]*Biomedical Data Science, Dartmouth College, tor.tosteson@dartmouth.edu*

The calculation of quality of care measures based on electronic medical records (EMRs) may be inaccurate because of incomplete capture of past services. We evaluate the influence of different statistical approaches for calculating the proportion of patients who are up-to-date for a preventive service, using the example of colorectal cancer (CRC) screening. We propose an extension of traditional mixture models to account for the uncertainty in compliance which is further complicated by the choice of various screening modalities with different recommended screening intervals. We conducted simulation studies to compare various statistical approaches and demonstrated that the proposed method can alleviate bias when individuals with complete prior medical history information were not representative of the targeted population. The method is motivated by and applied to data from the National Cancer Institute–funded consortium Population-Based Research Optimizing Screening through Personalized Regiments (PROSPR). Findings from the application are important for the evaluation of appropriate use of preventive care and provide a novel tool for dealing with similar analytical challenges with EMR data in broad settings.

**1. Introduction.** Accurate assessment of preventive services in health-care systems and the factors that influence compliance with those services is critical for monitoring and improving health-care delivery in the United States (Kohn et al. (2000)). An important first step in evaluating the receipt of preventive care services is the identification of the target population, the set of patients for whom providers and organizations are accountable for the quality of care (Landon, O'Malley and Keegan (2010)). However, this population is often dynamic as patients move in and out of various health systems and therefore is difficult to define clearly. For example, patients might have received services outside the current health system without documentation in the current data systems. In addition, long windows of time are often required to assess patient adherence to recommended services. Thus, the calculation of quality-of-care measures based on electronic medical records (EMR) may be inaccurate because the history of services received is incompletely documented. Such a challenge is widely encountered when assessing a variety of preventive services, including screening for

Received June 2019; revised March 2020.
*Key words and phrases.* Cancer screening, EMR data, event-time analysis, mixture model.

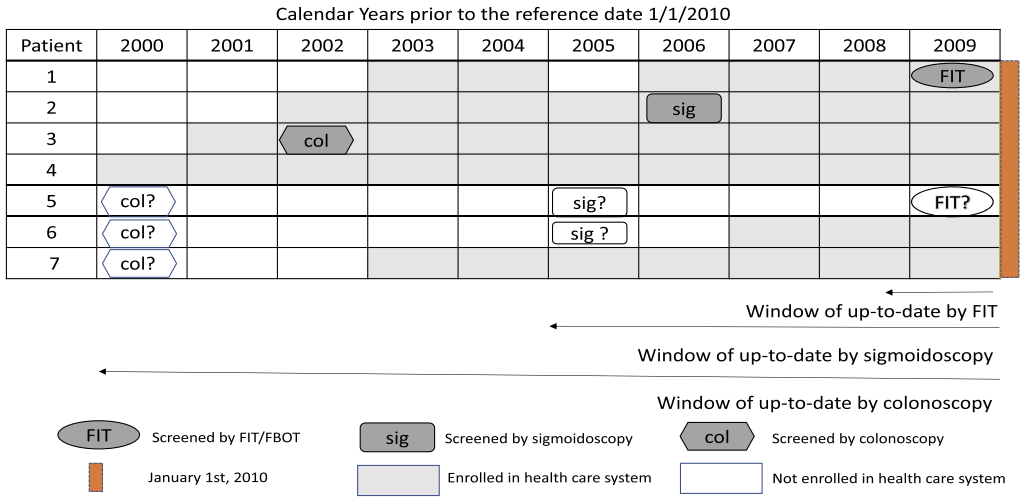


FIG. 1. *Ascertainment of UTD status for CRC screening at the reference date of January* 1*st*, 2010, *for seven hypothetical patients*. *Patient* 1: *UTD by FIT*; *Patient* 2: *UTD by flexible sigmoidoscopy*; *Patient* 3: *UTD by colonoscopy*; *Patient* 4: *not UTD*; *Patient* 5: *UTD is uncertain for all modalities*; *Patient* 6: *UTD is uncertain by sigmoidoscopy and colonoscopy*; *Patient* 7: *UTD is uncertain by colonoscopy.*

breast, cervical, colorectal or lung cancer, immunization for children and adults and those who switch to a different health system when entering Medicare at age 65.

In this manuscript we evaluate the influence of different statistical approaches for calculating the proportion of patients up to date for a preventive service, using the example of colorectal cancer (CRC) screening. CRC screening illustrates several key analytic challenges to ascertain up-to-date status. National policy groups recommend screening average-risk individuals with multiple modalities, including high-sensitivity fecal immunochemical or occult blood testing (FIT/FOBT), flexible sigmoidoscopy or colonoscopy (Lin et al. (2016)). Determining up-to-date status for screening, therefore, requires taking into account the multiple screening modalities currently available for CRC. In addition, screening intervals vary depending on the test used and its results, such as annually for a FIT/FOBT and every 10 years for a colonoscopy for an average-risk individual or more frequent colonoscopies for some higher-risk populations. For any individual, therefore, up-to-date status for screening relies upon having accurate ascertainment of prior screening history. However, even health-care systems that have been using EMRs for a long time may not have 10 years of retrospective information on all current patients which is the maximum length of CRC screening interval for any modality. Up-to-date screening status is often incomplete in EMRs for quality measure calculation, and because such incompleteness is often related to individual characteristics, we cannot assume the data are missing completely at random (Chubak and Hubbard (2016)). Figure 1 illustrates different scenarios for ascertaining patients' up-to-date screening status at the beginning of the calendar year 2010 when the proportion of the patients who are up to date for screening is calculated. The first three patients are up to date by each of the three modalities, respectively, based on historical information in EMRs prior to 2010, and the fourth patient had no screening in the prior 10 years of enrollment. For each of the remaining three patients, the status is ambiguous because the enrollment period is shorter than the interval for which a specific screening modality is recommended, and patients might, potentially, have received service prior to enrollment in the current health system. Various approaches can be considered for calculating preventive service up-to-date proportion with data from EMRs. For CRC screening, one may simply determine up-to-date status for all screening-eligible patients based on the known prior screening information available in the EMR regardless of

their length of enrollment in the health-care system (Klabunde et al. (2016)). This approach may lead to an underestimation of up-to-date proportion because individuals with prior testing done outside their current health system are not be counted. In a regression model where up to date for screening status is considered as a predictor or an outcome, this approach may also lead to biased estimation because of misclassification in this variable. Alternatively, one may specify a targeted population by focusing on people with a specific length of enrollment, for example, in a health plan for a number of consecutive years (Brunelli et al. (2013), Shortreed et al. (2016)). This approach has been shown to lead to different estimates of screening rates when inclusion criteria differ, and the selected patients, arguably, reflect the true population for which the health-care system is accountable in terms of administration of screening (Landon, O'Malley and Keegan (2010)). Further investigations on the impact of various assumptions on estimations of up-to-date status are warranted, and an alternative approach is needed to better identify individuals in need of screening services and to improve modeling of screening compliance.

To address these complications in CRC screening, we propose a statistical-modeling-based approach to account for missing EMR information on receipt of preventive services, using CRC screening as an example. By modeling an individual's probability of receiving each screening modality at various time points prior to enrollment, the approach may lead to improved estimates of screening rates and better-calibrated regression models that aim to identify factors associated with adherence to screening, compared with the aforementioned existing approaches. While there is vast literature on missing data approaches (Little and Rubin (2014)), limited methods are available to account for complicated scenarios of missing data, such as those encountered in the CRC screening setting. Appropriate statistical approaches are needed to account for the uncertainty of whether individuals screen at all during a target period which is further complicated by the choice of various screening modalities and different recommended subsequent screening intervals. Hubbard et al. (2017) considered a finite mixture model approach to account for potential misclassification in the test indication of colonoscopy using EMR data. The implications of such misclassification for EHR-based screening utilization estimates have also been investigated (Hubbard, Chubak and Rutter (2014)). Our approach estimates the probability of being up to date for any screening by calculating the probabilities of receiving each modality of screening within the modality-specific interval preceding the time of evaluation. We cast the problem into the framework of traditional mixture models (Farewell (1982), Sy and Taylor (2000)), where we consider the proportion of patients without a screening event as the "cure fraction" and extend the original model to incorporate multiple screening modalities and their various intervals into the "uncured fraction." We also provide variance estimators of the resulting screening proportions and parameters in a propensity-augmented logistic regression model estimated simultaneously as the cure-fraction model. The methods are applied to data from the NCI-funded consortium Population-Based Research Optimizing Screening through Personalized Regimens (PROSPR). The overall aim of PROSPR is to conduct multisite, coordinated, transdisciplinary research to evaluate and improve cancer screening processes. The 10 PROSPR research centers reflect the diversity of the U.S. delivery system organizations. Our study uses data derived from two PROSPR integrated health-care systems involved in studying the CRC screening process (Tiro et al. (2014)). Findings from this application are important for the evaluation of appropriate use of preventive care and provide a novel tool in dealing with similar analytical challenges using EMR data in broad settings.

The manuscript is organized as follows: We introduce notation and model assumptions in Section 2. We describe our estimation and inference procedures for the age-specific screening rates of a screening program in Section 3. We illustrate how to use estimated individual up-to-date probability in regression models for identifying factors associated with being up to

date for screening. The results of simulation studies evaluating the proposed procedures and comparisons among various analytical strategies are presented in Section 4. In Section 5 we illustrate our methods with PROSPR data.

## 2. Models for CRC screening.

2.1. *Notation and setup*. Here, we consider a broad definition of up to date for screening to mean receiving a test in the recommended time window regardless of the indication of the test (Burnett-Hartman et al. (2016)). Consider, for evaluating screening up-to-date proportion, a target population of $n$ individuals is eligible for screening. Assume there are $M$ mutually exclusive scenarios for screening. In the specific context of CRC screening, all individuals aged 50 to 75 are eligible for screening and constitute the target population, and $M = 4$ with the first three types of events representing up-to-date screening by a fecal blood test, sigmoidoscopy, colonoscopy and the $M$th category, representing the no-testing group, or screening-eligible individuals who are not up to date by any of the testing modalities.

Denote by $\mathbf{Y}$ a random matrix of $n$ by $M$ dimension with element

$$Y_{ij} = \begin{cases} 1 & \text{if the } i\text{th subject is up to date for } j\text{th screen modality at the time of evaluation;} \\ 0 & \text{otherwise;} \end{cases}$$

for $j = 1, \ldots, M-1$, $i = 1, \ldots, n$. Individuals with no prior screening are considered in the $M$th category, $Y_{iM} = 1$ if $Y_{ij} = 0$ for all $j \neq M$. $\mathbf{Y}$ thus has a multinomial distribution with $\pi_j = \mathrm{P}(Y_{ij} = 1)$ and $\sum_{j=1}^{M} \pi_j = 1$. Note that $\mathbf{Y}$ represents the true screening status and, due to incomplete information in EMR, it is not always observed for all.

Since a prior screening event is ascertained by looking back from the time of evaluation, we consider a composite backward failure time $T_{ij}^*$, which is the time from the date when screening receipt is evaluated (time 0) to the time that the $j$th type of prior test/procedure is observed, for $j = 1, \ldots, M-1$, conditioning on $Y_{ij} = 1$, and $T_{ij}^* = \infty$ for $j = M$. We note that in this notation, as time increases it indicates going further back into the past. Let $C_i$ denote the censoring time for individual $i$, that is, the farthest time in history that an individual has data available. A similar idea with backward recurrence time models was considered in estimating trends in receipt of colonoscopy before age 50 (Rutter et al. (2015)).

Time to a prior observed screening event $T_{ij}^*$ can be censored, for example, by an individual's new enrollment into a health-care setting and capacity to ascertain screening history electronically from existing medical records in the health system. In addition, let $a_j$ denote the upper limit of the time window in which the $j$th type of the test can occur for an individual to be considered up to date. $a_j$ is often related to the recommended screening interval and initial age of screening specific to the $j$th modality, for example, for colonoscopy $a_j = 10$ years. For the $j$th screening modality, let $C_{ij} = a_j \wedge C_i$. The observed time of the $j$th screening event in the past is denoted as $T_{ij} = T_{ij}^* \wedge C_{ij}$, and $\delta_{ij} = 1$ if $T_{ij}^* \leq C_{ij}$ (i.e., $Y_{ij} = 1$ and $Y_{ik} = 0$ for $j \neq k$). Let $\delta_i = \sum_{j=1}^{M-1} \delta_{ij}$, and denote $V_i$ an indicator for screening up-to-date ascertainment, $V_i = \delta_i + (1 - \delta_i) I(\max_j(a_j) \leq C_i)$. When $V_i = 1$, either one of the test modality is observed or none is observed within the upper-limit of screening interval.

2.2. *Model specification*. Let $\mathbf{Z}_{ij}$ be a covariate vector for subject $i$ that predicts the choice of screening modality $j$. Our key models of interests are multinomial incidence models (Kleinbaum and Klein (2002)) for $j = 1, \ldots, M-1$,

$$\log\left[\frac{\pi_j(\mathbf{z}_{ij}; \boldsymbol{\alpha})}{\pi_M(\mathbf{z}_{ij}; \boldsymbol{\alpha})}\right] = \boldsymbol{\alpha}_j^{\top} \mathbf{z}_{ij}, \tag{1}$$

TABLE 1
*A List of Variables used in Model Specification*

| Variable | Interpretation |
|---|---|
| $\mathbf{Y}$ | true screening status, partially known |
| V | indicator if $\mathbf{Y}$ is known from the data |
| $T_j^*$ | time to a prior observed screening test of modality $j$, partially known |
| $C$ | censoring time due to membership enrollment |
| $a_j$ | recommended screening interval for modality $j$ |
| $T_j$ | minimum of $T_j^*$, $C_i$ and $a_j$, fully known |
| $\delta_j$ | censoring indicator equals 1 if $T_j^*$ is observed. |
| $\mathbf{z}_j$ | covariates related the $j$th modality in the multinomial incidence model for $\mathbf{Y}$ |
| $\mathbf{x}_j$ | covariates related the $j$th modality in the survival model for $T$ |

where $\pi_k(\mathbf{z}_{ik};\boldsymbol{\alpha}) = \mathrm{P}(Y_{ik} = 1 \mid \mathbf{z}_{ik})$. For a specific screening modality, the likelihood of observing receipt of the test is dependent on the length of the time window that an individual remains up to date for a given test and the length of enrollment of the patient. Therefore, for each screening type $j$, $j = 1, \ldots, M-1$, a statistical model for $T_{ij}^*$ in relation to predictor $\mathbf{X}_{ij}$ can then be specified with a survival model,

$$\mathrm{P}(T_{ij}^* > t | \mathbf{X}_{ij} = \mathbf{x}_{ij}, y_{ij} = 1) = \mathcal{T}_0\{\log[H_{0j}(t)] + \boldsymbol{\beta}_j^{\mathsf{T}}\mathbf{x}_{ij}\} I(y_{ij} = 1), \tag{2}$$

where $\boldsymbol{\beta}_{xj}$ are the regression parameters associated with $\mathbf{x}_{ij}$, $\mathcal{T}_0$ is a specific distributional function and $\log[H_{0j}(t)]$ is a unknown monotonic increasing function. Model 2 includes both the popular Cox model (Cox (1975)) and other class of linear transformation model, such as the proportional odds model (Cheng, Wei and Ying (1995)), as special cases.

A brief outline of variables defined in the model specification is summarized in Table 1.

## 3. Estimation.

3.1. *Estimating up-to-date screening rates.* Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_{M-1}^{\mathsf{T}})^{\mathsf{T}}$ with length $p_\beta$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_{M-1}^{\mathsf{T}})^{\mathsf{T}}$ with length $p_\alpha$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathsf{T}}, \boldsymbol{\alpha}^{\mathsf{T}})^{\mathsf{T}}$. Note that, when $V_i = 0$, $Y_{ij}$ can be either 0 or 1 because a screening event could have occurred in the time interval between $C_i$ and $a_j$. Therefore, the corresponding observed likelihood is given by

$$\mathcal{L}(\theta) = \prod_i^n \left\{ \prod_j^{M-1} \left[\pi_j(\mathbf{z}_{ij};\boldsymbol{\alpha}_j) f_j(t_{ij}|Y_{ij}=1,\mathbf{x}_{ij};\boldsymbol{\beta}_j)\right]^{Y_{ij}} \left[1 - \sum_j^{M-1} \pi_j(\mathbf{z}_{ij};\boldsymbol{\alpha}_j)\right]^{Y_{iM}} \right\}^{V_i}$$
$$\times \left\{ \left[1 - \sum_{j=1}^{M-1} \pi_j(\mathbf{z}_{ij};\boldsymbol{\alpha})\right] + \sum_{j=1}^{M-1} \pi_j(\mathbf{z}_{ij};\boldsymbol{\alpha}_j) S_j(t_{ij}|Y_{ij}=1,\mathbf{x}_{ij};\boldsymbol{\beta}_j) \right\}^{1-V_i},$$

where $f$ and $S$ represent density and survival functions of $T_{ij}^*$, respectively. The likelihood is the product of two components: the first component of the likelihood reflects the probability of observing each test modality in the past or of observing none among patients whose screening status are observable from their EHR, while the second component reflects such probabilities among these whose status are uncertain. As a likelihood based approach, the validity of the estimation relies on the correct specification of all components of the likelihood which implies that models of $\mathbf{Y}$ and $T$ are correctly specified for each modality.

Since $Y_{ij}$ is unknown for $V_i = 0$, we consider an iterative expectation maximization (EM) estimation procedure. Specifically, in the $k$th iteration of the E-step, we calculate the conditional expectation $Y_{ij}^*(\boldsymbol{\theta}^{(k)}) \equiv \Pr\{Y_{ij} = 1 | T_{ij}, V_i, \mathbf{z}_{ij}, \mathbf{x}_{ij}; \theta^{(k)}\}$ as

$$
\begin{aligned}
&V_i Y_{ij} \\
&+ (1 - V_i) \frac{\pi_j(\mathbf{z}_{ij}; \boldsymbol{\alpha}_j^{(k)}) S_j(t_{ij} | Y_{ij} = 1, \mathbf{x}_{ij}; \boldsymbol{\beta}_j^{(k)})}{1 - \sum_{j=1}^{M-1} \pi_j(\mathbf{z}_{ij}; \boldsymbol{\alpha}_j^{(k)}) + \sum_{j=1}^{M-1} \pi_j(\mathbf{z}_{ij}; \boldsymbol{\alpha}_j^{(k)}) S_j(t_{ij} | Y_{ij} = 1, \mathbf{x}_{ij}; \boldsymbol{\beta}_j^{(k)})},
\end{aligned} \tag{3}
$$

where $S_j(t_{ij} | Y_{ij} = 1, \mathbf{x}_{ij}; \boldsymbol{\beta}_j^{(k)})$ can be estimated as $\mathcal{T}_0\{\log[\widehat{\beta}_0^{(k)}(t_{ij})] + \widehat{\boldsymbol{\beta}}_j^{(k)^\top} \mathbf{x}_{ij}\}$. Note that $S_j(a_j) = 0$ as screening beyond $a_j$ will not be considered as an up-to-date event. Here, $Y_{ij}^*$ takes the observed $Y_{ij}$ when $V_{ij} = 1$ and when $V_{ij} = 0$, we impute $Y_{ij}$ with a conditional probability based on the known information. In M-step we can obtain $\boldsymbol{\theta}^{(k+1)}$ by solving the sets of $(M-1) \times 2$ estimating equations $U^\theta(\boldsymbol{\theta}) = (U_1^\theta(\boldsymbol{\theta})^\top, \ldots, U_{M-1}^\theta(\boldsymbol{\theta})^\top)^\top = 0$, where the $j$th component is

$$
U_j^\theta(\boldsymbol{\theta}) = \begin{pmatrix} U_j^{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \\ U_j^{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{pmatrix} \equiv \begin{pmatrix} \sum_i U_{ij}^{\boldsymbol{\alpha}}(\boldsymbol{\theta}) \\ \sum_i U_{ij}^{\boldsymbol{\beta}}(\boldsymbol{\theta}) \end{pmatrix} \quad \text{with}
$$

$$
\mathbf{U}_{ij}^{\boldsymbol{\alpha}}(\boldsymbol{\theta}) = \mathbf{z}_{ij} \{ Y_{ij}^*(\boldsymbol{\theta}^{(k)}) - \pi_j(\mathbf{z}_{ij}; \boldsymbol{\alpha}_j) \},
$$

$$
\mathbf{U}_{ij}^{\boldsymbol{\beta}}(\boldsymbol{\theta}) = Y_{ij}^*(\boldsymbol{\theta}^{(k)}) \int_0^{a_j} \left\{ \mathbf{x}_{ij} - \frac{n^{-1} \sum_i Y_{ij}^* I(T_{ij} \geq s) \mathbf{x}_{ij} \exp(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij})}{n^{-1} \sum_i Y_{ij}^* I(T_{ij} \geq s) \exp(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij})} \right\} dM_{ij}(s),
$$

$M_{ij}(s) = N_{ij}(s) - \int_0^s I(T_{ij} \geq u) \exp(\boldsymbol{\beta}_j^\top \mathbf{x}_{ij}) \, dH_{0j}(u)$ and $N_{ij}(s) = I(T_{ij} \leq s)$, under the proportional hazards model. Estimators of $\boldsymbol{\theta}$, denoted as $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\alpha}}^\top)^\top$, can be obtained when the expectation maximization (EM) iterations converge by a criteria such as $|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}| \leq \epsilon$ with a prespecified $\epsilon$.

An empirical estimator of the up-to-date rate (UTD) $\mathrm{P}(Y_{iM} = 0)$ is then calculated as

$$
\widehat{\mathrm{UTD}}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{j=1}^{M-1} \sum_{i=1}^{n} \widehat{Y}_{ij}^*(\widehat{\boldsymbol{\theta}}),
$$

where $\widehat{Y}_{ij}^*(\widehat{\boldsymbol{\theta}})$ is obtained by replacing $\boldsymbol{\theta}$ in equation (3) with $\widehat{\boldsymbol{\theta}}$.

To make an inference for $\widehat{\mathrm{UTD}}(\widehat{\boldsymbol{\theta}})$, we first show that the process $\widehat{\mathcal{U}}_{\boldsymbol{\theta}} = \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically equivalent to a sum of $n$ i.i.d terms, $n^{-\frac{1}{2}} \sum_i \eta_i(\boldsymbol{\theta})$, where $\eta_i(\boldsymbol{\theta}) = \{I^\theta(\boldsymbol{\theta})\}^{-1} U_i^\theta(\boldsymbol{\theta})$ and $I^\theta(\boldsymbol{\theta})$ is the limit of the observed information matrix, a block diagonal with the $j$th diagonal element as

$$
I_j^\theta(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial U_j^\alpha(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_1} & \cdots & \dfrac{\partial U_j^\alpha(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_{M-1}} & \dfrac{\partial U_j^\alpha(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_1} & \cdots & \dfrac{\partial U_j^\alpha(\boldsymbol{\theta})}{\partial \beta_{M-1}} \\ \dfrac{\partial U_j^\beta(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_1} & \cdots & \dfrac{\partial U_j^\beta(\boldsymbol{\theta})}{\partial \boldsymbol{\alpha}_{M-1}} & \dfrac{\partial U_j^\beta(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_1} & \cdots & \dfrac{\partial U_j^\beta(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}_{M-1}} \end{pmatrix}.
$$

We can then further show that $\widehat{\mathcal{U}}_{\mathrm{UTD}} = \sqrt{n}\{\widehat{\mathrm{UTD}}(\widehat{\boldsymbol{\theta}}) - \mathrm{UTD}(\boldsymbol{\theta})\}$ is asymptotically equivalent to a sum of $n$ i.i.d terms, $n^{-\frac{1}{2}} \sum_i \xi_i(\boldsymbol{\theta})$, where $\xi_i(\boldsymbol{\theta}) = \sum_j \{Y_{ij}^*(\boldsymbol{\theta}) - \pi_j\} + A_{\mathrm{UTD}} \eta_i(\boldsymbol{\theta})$ and $A_{\mathrm{UTD}} = \frac{\partial \mathrm{UTD}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. By the central limit theorem, $\widehat{\mathcal{U}}_{\mathrm{UTD}}$ converges weakly to a normal distribution with zero mean and variance $\sigma_{\mathrm{UTD}}^2$. A consistent estimator of $\sigma_{\mathrm{UTD}}^2$ can be estimated

empirically as

$$\widehat{\sigma}^2_{\mathrm{UTD}} = \frac{1}{n}\sum_i \xi_i^2(\boldsymbol{\theta}).$$

3.2. *Estimating parameters in a regression model for failure of screening up-to-date rates.* Often, there is an interest in investigating the association of various factors $\mathbf{W}$ with not up to date for screening. One may specify a regression model of the following form:

$$\log\left[\frac{\mathrm{P}(Y_{iM}=1|\mathbf{W}_i=\mathbf{w}_i)}{\mathrm{P}(Y_{iM}=0|\mathbf{W}_i=\mathbf{w}_i)}\right] = \gamma_0 + \boldsymbol{\gamma}_z^{\mathsf{T}}\mathbf{w}_i, \tag{4}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_0, \boldsymbol{\gamma}_w^{\mathsf{T}})^{\mathsf{T}}$ of length $p_\gamma$ are coefficients measuring the associations between covariates and failure of being up-to-date for screening. Note in the presence of model (4), model (1) can be considered a working model for the purpose of deriving partially observed $Y_{ij}$. Since $Y_{iM}$ is not always observable, we show below how to address such a challenge in estimation.

In the presence of censoring due to varying lengths of enrollment, the results from the above procedure provide a probability score for up-to-date screening as $\sum_{j=1}^{M-1}\widehat{Y}^*_{ij}(\boldsymbol{\theta})$ for each individual, and it can be used to estimate parameters in a regression model for up-to-date screening, as specified in model (4) (Hubbard et al. (2017)). Estimators of $\boldsymbol{\gamma}$, $\widehat{\boldsymbol{\gamma}}$, can be obtained by solving the following estimation equation:

$$U^{\boldsymbol{\gamma}}(\boldsymbol{\theta},\boldsymbol{\gamma}) \equiv \sum_i \mathbf{w}_i\left\{1 - \sum_{j=1}^{M-1}\widehat{Y}^*_{ij}(\boldsymbol{\theta}) - \frac{\exp(\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{w}_i)}{1+\exp(\boldsymbol{\gamma}^{\mathsf{T}}\mathbf{w}_i)}\right\} = 0, \tag{5}$$

where $\widehat{Y}^*_{ij}(\boldsymbol{\theta})$ is obtained from fitting both the survival model and the incidence model as specified in equation (3). To make an inference for $\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})$, it can be shown that the process $\widehat{\mathcal{U}}_\gamma = \sqrt{n}[\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}(\boldsymbol{\theta})]$ is asymptotically equivalent to a sum of $n$ i.i.d terms, $n^{-\frac{1}{2}}\sum_i \zeta_i(\boldsymbol{\theta},\boldsymbol{\gamma})$, where $\zeta_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = \mathcal{T}\{I(\boldsymbol{\theta},\boldsymbol{\gamma})\}^{-1}U_i(\boldsymbol{\theta},\boldsymbol{\gamma})$, with $U_i(\boldsymbol{\theta},\boldsymbol{\gamma}) = (U_i^{\boldsymbol{\theta}}(\boldsymbol{\theta}), U_i^{\boldsymbol{\gamma}}(\boldsymbol{\theta},\boldsymbol{\gamma}))$, $\mathcal{T}$ is a $p_\gamma \times (p_\beta + p_\alpha + p_\gamma)$ matrix, with elements $\mathcal{T}_{jk} = 1$ for $j = p_\beta + p_\alpha + 1, \ldots, p_\beta + p_\alpha + p_\gamma$, $k = j$ and $\mathcal{T}_{jk} = 0$ otherwise, and $I^{\theta,\gamma}(\boldsymbol{\theta},\boldsymbol{\gamma})$ is the limit of the observed information matrix of the form

$$I^{\theta,\gamma}(\boldsymbol{\theta},\boldsymbol{\gamma}) = \begin{pmatrix} I^{\boldsymbol{\theta},\boldsymbol{\theta}} & 0 \\ I^{\boldsymbol{\theta},\boldsymbol{\gamma}} & I^{\boldsymbol{\gamma},\boldsymbol{\gamma}} \end{pmatrix},$$

where $I^{\boldsymbol{\theta},\boldsymbol{\gamma}}$ is a $p_\gamma \times (p_\alpha + p_\beta)$ matrix with its $jk$th element as the limit of $\frac{\partial U^{\boldsymbol{\gamma}}(\boldsymbol{\theta},\boldsymbol{\gamma})_{(j)}}{\partial \boldsymbol{\theta}_{(k)}}$, for $j = 1, \ldots, p_\gamma$, $k = 1, \ldots, p_\alpha + p_\beta$ and $I^{\boldsymbol{\gamma},\boldsymbol{\gamma}}$ is a $p_\gamma \times p_\gamma$ matrix with the $jk$th component as $I^{\boldsymbol{\gamma},\boldsymbol{\gamma}}_{jk} = \frac{\partial U^{\boldsymbol{\gamma}}(\boldsymbol{\theta},\boldsymbol{\gamma})_{(j)}}{\partial \boldsymbol{\gamma}_{(k)}}$. By the central limit theorem, $\widehat{\mathcal{U}}_\gamma$ converges weakly to a normal distribution with zero mean and variance $\sigma^2_\gamma$. A consistent estimator of $\sigma^2_\gamma$ can be estimated empirically as

$$\widehat{\sigma}^2_\gamma = \frac{1}{n}\sum_i \zeta_i^2(\boldsymbol{\theta}).$$

**4. Simulation.** We conducted simulation studies to examine the finite sample performances of our proposed procedures and the impact of different analysis strategies on estimation. We considered a screening cohort of size $n = 5000$ with three possible screening modalities. We first generated two covariates that are associated with the choice of screening

modality as well as timing of a screening event—$Z_1$, a binary variable from the Bernoulli distribution with a probability of 0.2, and $Z_2$, a standard normal random variable. The screening modalities simulated in the cohort were generated as a multinomial incidence model as specified in equation (1). Conditioning on the modality is chosen, and time from baseline (time of screening rate calculation) to prior test for each modality was generated following equation (2) with covariates $X_1 = Z_1$ and $X_2 = Z_2$. Time between baseline to when a patient enrolled in the health-care system was generated from a Gamma distribution with rate parameter set to 1 and shape parameter selected for two scenarios. In Scenario I the shape parameter is set to a constant of 10, representing a population with a majority of individuals enrolled for long periods of time (mean enrollment period = 10). In Scenario II the shape parameter was generated for each individual as a function of $Z_1$ and $Z_2$. While the mean enrollment time is similar to Scenario I (mean enrollment period = seven years), the length of enrollment was dependent on the covariates of interest. The observed time was then taken to be the minimum of time to screening, time to enrollment or years of available data recorded, whichever occurs the latest. This represented a scenario with informative censoring, and the approach excluded patients with shorter enrollment times could lead to biased results. Additionally, we consider a scenario (Scenario III) where data were generated the same way as in Scenario II, except that the model was fit with $Z_1$ and $Z_2^*$, where $Z_2^*$ was generated as $Z_2$ plus a random normal variable with mean 0 and standard deviation of 0.5. This represents a setting where all models were misspecified as the error-prone version of the covariate was used in fitting the models. Supplementary Table 1 (Zheng et al. (2020)) shows the parameters used for various models. In Scenario I the up-to-date screening status could not be determined by 34% of the patients, and in Scenarios II and III the rate is 45%.

We compared the proposed estimation procedures (A: Estimated) to alternative analytical strategies for dealing with potential missing data in estimating the rate of up-to-date screening in the population as well as in estimating parameters in a regression analysis of up-to-date screening. The first comparison method used the full population of individuals, irrespective of membership duration/time under observation, and only used documented tests/procedures for ascertaining screening up to date (B: Observed). With this approach individuals without any documented test/procedure within the time frame recommended by the guideline are considered not up to date. The second comparison approach, which might be an extreme alternate strategy, calculated up-to-date probabilities only among individuals who have been observed for a sufficient length of time so that all prior tests fall within the screening interval of interest (C: Long Enrollees). In the case of CRC screening, ideally, patients need to have at least 10 years of data availability so that their prior tests within the maximum recommended screening interval, such as for colonoscopy, are captured. Therefore, in the simulation we also calculated up-to-date rates by restricting analyses to data from individuals whose enrollment times were longer than the minimum of 10 years prior to current evaluation time.

As shown in Table 2, in Scenarios I and II, all point estimates of regression parameters for the up-to-date screening model and screening rates were very close to the true parameters under both scenarios for our proposed approach. The asymptotic-based standard error estimators approximated the empirical standard errors well, with empirical coverage levels of the 95% confidence intervals close to the nominal level for most of the parameters. A slightly conservative estimate of the standard error was observed for the up-to-date screening rate for sigmoidoscopy. For Scenario III the coverages were similar for the screening rate estimators; however, the inference procedure was not valid for the regression parameters. Table 3 provides comparisons among different analytical approaches. Under both Scenarios I and II, Approach A, our proposed model-based approach, yielded essentially unbiased estimates and the lowest mean squared error (MSE). Approach B, which ignored potential unobserved screening events due to short observable time intervals, led to substantial bias

TABLE 2

*Finite sample performance of proposed method*: *Estimates* (*Est.*), *Empirical Standard Deviation* ($SD_{emp}$), *Analytical Standard Error* ($SE_{ana}$) *and* 95% *coverage probability* (*CP*) *for screening rates* ($P_{nUTD}$ (*nUTD*: *not up to date*), $P_{colonoscopy}$, $P_{sigmoidoscopy}$, $P_{FIT/FOBT}$), *regression parameters* ($\gamma_0$, $\gamma_1$, $\gamma_2$) *in the failure of up-to-date screening model based on* 500 *simulation studies of a cohort* ($n = 5000$)

| | True | Est. | $SD_{emp}$ | $SE_{ana}$ | CP |
|---|---|---|---|---|---|
| Scenario I | | | | | |
| $P_{nUTD}$ | 0.375 | 0.377 | 0.007 | 0.008 | 0.965 |
| $P_{colonoscopy}$ | 0.299 | 0.299 | 0.007 | 0.007 | 0.963 |
| $P_{sigmoidoscopy}$ | 0.046 | 0.045 | 0.003 | 0.005 | 0.980 |
| $P_{FIT/BOBT}$ | 0.279 | 0.279 | 0.006 | 0.007 | 0.967 |
| $\gamma_0$ | −0.548 | −0.542 | 0.037 | 0.039 | 0.951 |
| $\gamma_1$ | −0.328 | −0.335 | 0.080 | 0.082 | 0.956 |
| $\gamma_2$ | −0.984 | −0.995 | 0.040 | 0.042 | 0.951 |
| Scenario II | | | | | |
| $P_{nUTD}$ | 0.375 | 0.376 | 0.008 | 0.009 | 0.950 |
| $P_{colonoscopy}$ | 0.299 | 0.299 | 0.008 | 0.009 | 0.942 |
| $P_{sigmoidoscopy}$ | 0.046 | 0.045 | 0.003 | 0.005 | 0.986 |
| $P_{FIT/FOBT}$ | 0.279 | 0.279 | 0.006 | 0.007 | 0.964 |
| $\gamma_0$ | −0.548 | −0.545 | 0.044 | 0.047 | 0.948 |
| $\gamma_1$ | −0.328 | −0.334 | 0.087 | 0.092 | 0.954 |
| $\gamma_2$ | −0.984 | −0.989 | 0.053 | 0.054 | 0.952 |
| Scenario III | | | | | |
| $P_{nUTD}$ | 0.375 | 0.375 | 0.008 | 0.039 | 0.939 |
| $P_{colonoscopy}$ | 0.299 | 0.300 | 0.008 | 0.014 | 0.945 |
| $P_{sigmoidoscopy}$ | 0.046 | 0.045 | 0.003 | 0.006 | 0.988 |
| $P_{FIT/FOBT}$ | 0.279 | 0.279 | 0.006 | 0.031 | 0.959 |
| $\gamma_0$ | −0.548 | −0.527 | 0.043 | 0.163 | 0.910 |
| $\gamma_1$ | −0.328 | −0.313 | 0.088 | 0.143 | 0.948 |
| $\gamma_2$ | −0.984 | −0.751 | 0.044 | 0.134 | 0.004 |

(underestimated screening rates and attenuated covariate effects) and high MSE in both estimation of up-to-date rates and the regression parameters of up-to-date screening. Approach C, restricting to long-term enrollees, led to unbiased estimates but higher MSE compared to Approach A under the first scenario when censoring by enrollment time was noninformative. However, bias was substantial in the second scenario when enrollment length was associated with the covariate of interest. These results suggest that, in a population where the majority of patients tend to stay enrolled within the health system for long periods of time and the enrollment length is independent of factors related to the choice of screening modality and compliance, one may consider simple methods of restricting to the long enrollees. On the other hand, in a population with many patients whose enrollment lengths may be related to screening decisions, restricting to long-term enrollees not only leads to significant loss of power but also introduces biases, since such a group might not represent the underlying screening population. Finally, under Scenario III with misspecified models, we found that our proposed estimators were quite robust compared with Approaches B and C: they still provided essentially unbiased estimates for the four screening rates, despite the biased estimate for the effect of $Z_2$ on the screening up-to-date status.

**5. Example.** We consider estimating up-to-date screening rates at three health-care systems within PROSPR: Kaiser Permanente (KP) Washington and KP Northern and Southern California (Tiro et al. (2014)). These integrated health-care systems provide health insurance

TABLE 3

*Simulation results comparing proposed methods with alternative approaches*: *Bias* (%) *and* 100× *mean squared errors* (MSE × 100) *for screening rates* ($P_{nUTD}$, $P_{colonoscopy}$, $P_{sigmoidoscopy}$, $P_{FIT/FOBT}$), *regression parameters* ($\gamma_0$, $\gamma_1$, $\gamma_2$) *in the failure of up-to-date screening model by of three estimating approaches*: (*A*) *proposed method* (*Estimated*); (*B*) *based on the observed fraction* (*observed*); *and* (*C*) *take only these enrolled for over* 10-*years*

| | A: Estimated | | B: Observed | | C: Long Enrollees | |
|---|---|---|---|---|---|---|
| Parameter | Bias% | MSE × 100 | Bias% | MSE × 100 | Bias% | MSE × 100 |
| Scenario I | | | | | | |
| $P_{nUTD}$ | 0.332 | 0.001 | 4.539 | 0.029 | 0.371 | 0.006 |
| $P_{colonoscopy}$ | 0.008 | 0.000 | −4.891 | 0.022 | 0.08 | 0.005 |
| $P_{sigmoidoscopy}$ | −2.644 | 0.000 | −4.799 | 0.001 | −3.132 | 0.001 |
| $P_{FIT/FOBT}$ | −0.021 | 0.000 | −0.059 | 0.000 | −0.066 | 0.005 |
| $\gamma_0$ | −1.050 | 0.014 | −16.572 | 0.830 | −1.169 | 0.151 |
| $\gamma_1$ | 2.146 | 0.034 | 25.444 | 0.633 | 4.155 | 0.798 |
| $\gamma_2$ | 1.067 | 0.029 | 6.794 | 0.462 | 1.746 | 0.216 |
| Scenario II | | | | | | |
| $P_{nUTD}$ | 0.165 | 0.003 | 18.807 | 0.499 | −23.433 | 0.794 |
| $P_{colonoscopy}$ | 0.030 | 0.002 | −22.114 | 0.439 | −9.074 | 0.094 |
| $P_{sigmoidoscopy}$ | −1.411 | 0.000 | −7.931 | 0.001 | 6.086 | 0.005 |
| $P_{FIT/FOBT}$ | −0.012 | 0.000 | −0.257 | 0.000 | 40.204 | 1.284 |
| $\gamma_0$ | −0.475 | 0.074 | −70.970 | 15.067 | −10.458 | 1.153 |
| $\gamma_1$ | 1.274 | 0.178 | 108.455 | 11.257 | 51.424 | 5.766 |
| $\gamma_2$ | 0.693 | 0.137 | 33.019 | 10.559 | 16.140 | 3.554 |
| Scenario III | | | | | | |
| $P_{ns}$ | 0.065 | 0.007 | 19.026 | 0.514 | −23.228 | 0.782 |
| $P_{cl}$ | 0.365 | 0.007 | −22.166 | 0.443 | −9.029 | 0.097 |
| $P_{sig}$ | −1.633 | 0.001 | −7.961 | 0.002 | 6.785 | 0.007 |
| $P_{FIT}$ | 0.147 | 0.004 | −0.141 | 0.004 | 40.132 | 1.282 |
| $\gamma_0$ | −4.735 | 0.274 | −73.361 | 16.286 | 2.714 | 0.937 |
| $\gamma_1$ | −2.811 | 0.777 | 89.264 | 9.241 | 43.129 | 5.874 |
| $\gamma_2$ | −23.082 | 5.332 | −0.357 | 0.102 | −14.349 | 2.756 |

coverage and care to over seven million members in Washington State, Northern California and Southern California. The study population includes patients aged 50–89 years who were enrolled from January 1, 2010, through December 31, 2013. Patients with a known history of partial or total colectomy or invasive CRC were excluded. Patient demographics, health status and health-care utilization data were obtained from the system's electronic clinical and administrative databases and submitted in standardized and structured formats to PROSPR's central data repository. Patient data for prior screening that occurred within the health-care system was retrospectively available in electronic databases up to 2006 for KP Washington and up to 1999 for KP California. For more details about PROSPR data and the research centers and health-care systems participating in the CRC screening component of PROSPR, see Tiro et al. (2014).

We used longitudinal data from the PROSPR data repository to calculate up-to-date screening rates. The screening rates were calculated at January 1st, 2010, among individuals who enrolled in 2010, using screening history information prior to cohort entry. We consider a patient up to date for screening if he/she had a colonoscopy within 11 years prior to cohort entry, a sigmoidoscopy within six years or a FIT/FOBT within 1.25 years. The time windows are more liberal than those recommended by USPSTF to account for the potential time lag in scheduling and reporting, and such lag is often not considered clinically significant. Due to

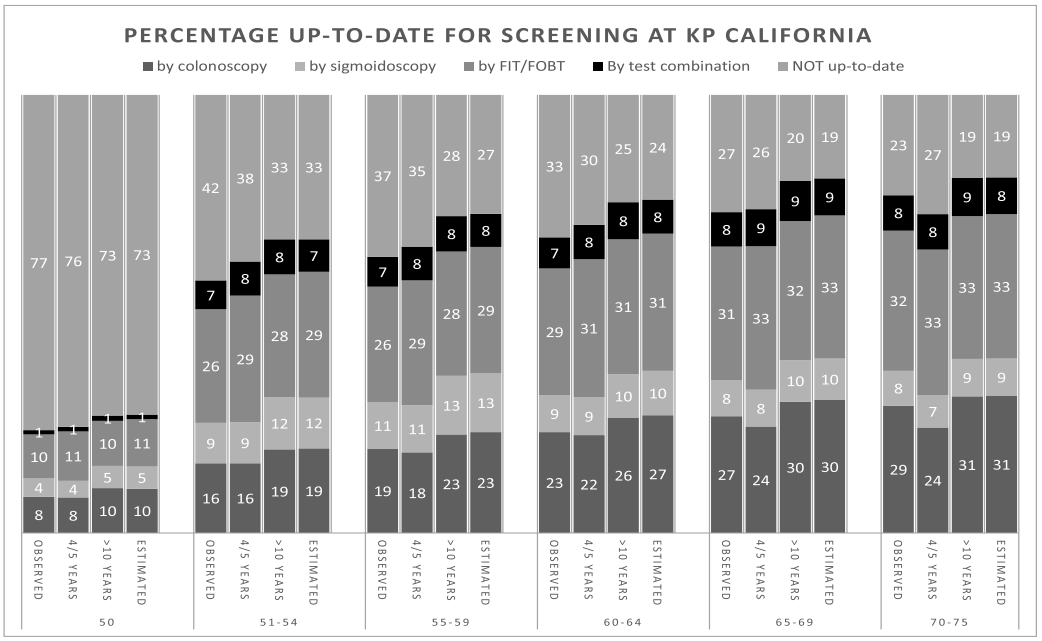


FIG. 2. *Comparing up-to-date for screening rates calculated from observed record* (*observed*), *observed record for individual enrolled four to five or* $\geq$ 10 *years*, *respectively*, *or*, *based on the proposed method* (*estimated*), *by different age groups and research centers*. *Results are for KP California*.

the lack of information on the indication of colonoscopy in our data, we did not distinguish between surveillance and screening colonoscopies. Individuals require a screening interval for colonoscopy between three to five years but received a colonoscopy between five to 10 years would still be counted as up to date for screening. This may lead to an overestimation of the rate of up to date for screening. For each screening modality event time is defined as the time from January 1st, 2010, back to a prior test within the specific time window, and such time can be censored at the time an individual initially enrolled in the health system. In addition, we created a mixed category for individuals, who had both a sigmoidoscopy within the past five years and an FOBT test within three years and were not up to date by colonoscopy, to be consistent with the current screening recommendation. We stratified the analysis by PROSPR research centers (KP Washington and KP California) and age groups of 50 years, 51–54, 55–59, 60–64, 65–69 and 70–75 years at the time of cohort entry. The stratification allows for differential rates and covariate effects by age and health-care system.

Among individuals aged 50–75 years in the 2010 cohort, 117,661 were from KP Washington and 1,663,811 were from KP California. Median length of prior available data in the repository was 48 months (IQR: 24–48 months) for KP Washington and 132 months (IQR: 102–132 months) for KP California. Predictors in models for type of screening modality, compliance and time to screening tests included sex, body mass index (BMI), race/ethnicity, Charlson comorbidity index and median household incomes from census data in the zip-code of residence, all measured at the year of cohort entry. Screening rates were calculated using the procedure described in Section 2 and data based on historical tests recorded prior to 2010. For comparison, we also calculated observed rates within the subsets of patients who had four to six years of enrollment in KP California and KP Washington and, separately in KP California, for those with 10 or more years of enrollment. Screening rates varied by different approaches, age groups and research centers (Figure 2 and Figure 3). At KP California (2), at a younger age of 51–54 years, 33% of patients were estimated as without an up-to-date screening event (42% observed). Among patients who were up to date for screening, most

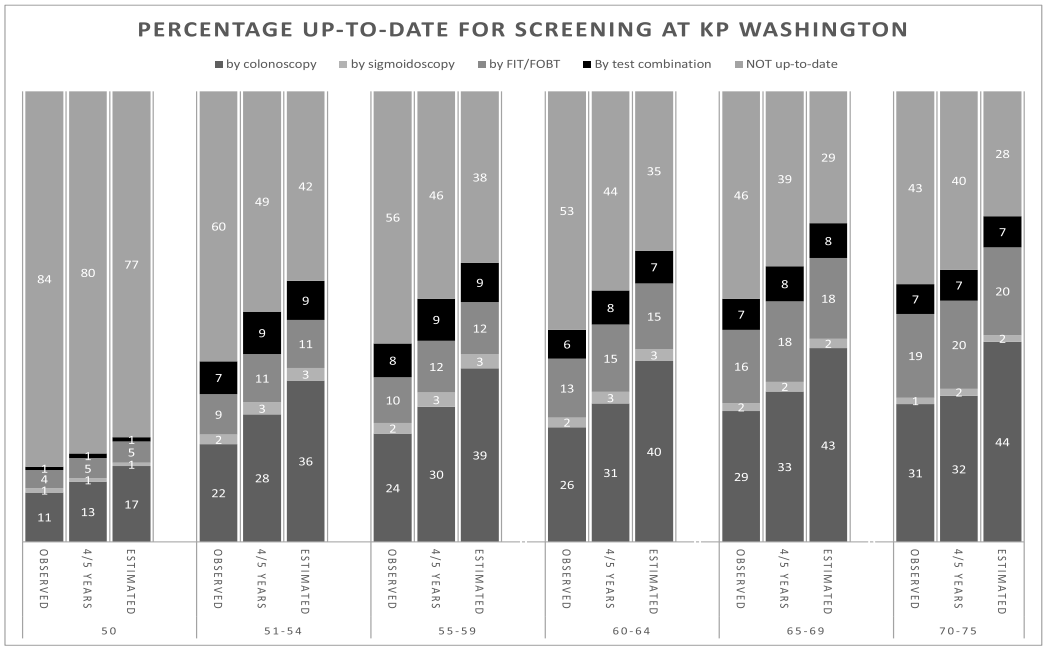


FIG. 3. *Comparing up-to-date for screening rates calculated from observed record* (*observed*), *observed record for individual enrolled four to five or* $\geq 10$ *years*, *respectively*, *or*, *based on the proposed method* (*estimated*), *by different age groups and research centers*. *Results are for KP Washington*.

individuals were estimated to have received a FIT/FOBT alone (26% observed vs. 29% estimated) or a colonoscopy (16% observed vs. 19% estimated). The remaining individuals were estimated to have received screening either from sigmoidoscopy (12%) or by a combination of tests (7%). In contrast, at KP Washington (3) 42% were estimated to not be up to date for screening (60% observed). Among individuals up to date for screening at KP Washington, most patients had received a colonoscopy (22% observed vs. 36% by estimation), followed by a FIT/FOBT (9% observed vs. 11% estimated), a sigmoidoscopy (2% observed vs. 3% estimated) or a mixed modality (7% observed vs. 9% estimated). For both KP California and KP Washington adherence was higher in older age groups; however, there were still an estimated 19% at KP California and 28% at KP Washington who were not up to date at 70–75 years of age. Across all age groups the estimated percentage of individuals without an up-to-date screening test tended to be lower than the observed percentage among individuals enrolled between four to five years. For KP California the estimated percentages were quite comparable with the observed percentages among patients enrolled for over 10 years. The results suggested that the proposed procedure was useful at recovering potential tests that might have occurred prior to patients' enrollment in the current care system and provided better estimates of screening delivery in the population. Such improvement was especially helpful in populations with shorter periods for recording prior tests.

We also evaluated associations between patient characteristics and their likelihood of failure to be up to date for screening and compared the impact of different approaches for dealing with missing data in prior test information on such evaluations. We calculated age-specific odds ratios (ORs) based on either only the recorded prior tests (A: Observed) or the expected values for screening calculated from fitted models as in equation (5) (B: Estimated). Associations varied to some degree across age groups, but trends were similar (data not shown), and stronger significant associations were more often observed with models using estimated screening status compared with observed screening status. To ease comparisons, we calculated the age-adjusted ORs from the age-specific models using meta-analytic methods. The

TABLE 4

*Odds ratio (95% Confidence Intervals) for failure of up-to-date screening results from KP California and KP Washington averaging age groups. Estimates are based on multivariate models using estimated weights for being without an up-to-date screening as the outcome*

| | KP California | | KP Washington | |
|---|---|---|---|---|
| | Observed | Estimated | Observed | Estimated |
| Gender: Male | | | | |
| Female1 | 1.11 (1.08, 1.14) | 1.00 (0.98, 1.01) | 1.00 (1.00, 1.00) | 1.00 (0.99, 1.01) |
| Charleson: 0 | | | | |
| 1 | 1.00 (0.97, 1.04) | 1.01 (0.99, 1.03) | 0.90 (0.83, 0.99) | 0.87 (0.77, 0.99) |
| 2 | 0.90 (0.86, 0.94) | 0.99 (0.97, 1.01) | 0.84 (0.73, 0.98) | 0.85 (0.74, 0.98) |
| $\geq 3$ | 0.99 (0.95, 1.05) | 1.01 (0.98, 1.04) | 1.00 (1.00, 1.01) | 1.00 (0.94, 1.07) |
| Race: White | | | | |
| Black | 1.08 (1.02, 1.16) | 1.09 (0.96, 1.23) | 0.99 (0.97, 1.01) | 1.10 (1.01, 1.19) |
| Asia | 1.10 (1.05, 1.15) | 0.98 (0.92, 1.04) | 1.01 (0.98, 1.03) | 0.90 (0.82, 0.99) |
| Pacific | 1.29 (1.18, 1.4) | 1.08 (0.98, 1.19) | 1.46 (1.04, 2.03) | 1.42 (1.04, 1.95) |
| Mix | 0.95 (0.87, 1.05) | 1.00 (0.99, 1.00) | 0.90 (0.81, 1.00) | 0.98 (0.96, 1.00) |
| Hispanic | 1.06 (0.99, 1.13) | 0.99 (0.98, 1.01) | 1.12 (1.01, 1.24) | 1.09 (1.00, 1.18) |
| BMI: $<18.5$ | | | | |
| [18.5–25) | 0.91 (0.82, 1.01) | 0.81 (0.67, 0.98) | 0.83 (0.70, 0.99) | 0.79 (0.63, 0.98) |
| [25, 30) | 0.97 (0.87, 1.08) | 0.91 (0.83, 1.00) | 0.92 (0.83, 1.01) | 0.90 (0.80, 1.01) |
| [30, 35) | 1.06 (0.95, 1.18) | 0.97 (0.91, 1.04) | 1.01 (0.97, 1.05) | 0.99 (0.92, 1.07) |
| $\geq 35$ | 1.20 (1.07, 1.33) | 1.07 (0.95, 1.20) | 1.00 (1.00, 1.01) | 0.99 (0.98, 1.01) |
| Income: $<53$k | | | | |
| 53–70k | 0.88 (0.85, 0.90) | 1.00 (0.99, 1.01) | 1.03 (1.00, 1.05) | 1.05 (1.00, 1.10) |
| $\geq 70$k | 0.93 (0.90, 0.96) | 0.99 (0.95, 1.03) | 1.00 (0.99, 1.01) | 1.00 (1.00, 1.01) |

results are presented in Table 4. For both KP California and KP Washington, patients who failed to be up to date for screening (compared to these who were up-to-date) tended to have a higher comorbidity index ($\geq 3$) (compared to comorbidity index $= 0$), be of Hispanic ethnicity or a Pacific Islander (compared to white) and were severely obese or underweight (compared to having normal weight). Up-to-date screening did not vary substantially by gender. It appeared that, while there could have been discrepancies in screening rates estimated from different approaches due to missing data in screening history, the covariate effects of regression models of up-to-date screening were more comparable in this specific study. Our proposed methods provide a way to conduct appropriate sensitivity analysis.

**6. Discussion.** Accurate assessment of the quality of preventive care is important, as it provides health-care systems with the confidence that measurements indeed reflect the quality of care delivered and that measurements are not faulty due to data quality. The identification of patient populations for which preventive care interventions should be delivered when calculating measures of quality of care can be complicated due to limitations in the availability of current and prior history information in EMR systems, which often do not communicate across institutions when patients change health plans or when an institution implements or switches to new EMR systems from a paper system. To date, the literature does not provide clear guidance on how simple approaches that ignore such missing data may impact both the estimation of the rates of care delivery and the investigation of factors that influence compliance.

In this manuscript we considered a statistical approach for evaluating whether a person is up to date with a specific preventive care procedure, using a scientific motivating example of colorectal cancer screening. The approach also allows for a more valid evaluation of factors

that may impact up-to-date status. By estimating the probability of an individual having received each of the available screening modalities during the specific time window for each specific test, our approach is potentially able to correct for the misclassification of screening status for individuals with missing data and also reduce the attenuation in regression parameters due to potential misclassification. Illustrated with CRC screening, our proposed approach can be modified for broader applications in other preventive care settings.

As a full likelihood-based approach, the validity of the estimates relies on correctly specifying the multinomial model for the choice of screening modality and the survival model for time to a prior test for each screening modality. Failure to include key predictors or mismeasured covariates in these models may lead to biased estimates. Using a proportional hazards model implies that the covariate effects are stable over time, but such a model might be violated if the screening policy changes over time and the change impacts patient subgroups differently. We also made the assumption that factors impacting screening decisions are relatively stable during the enrollment period, such that covariates collected at study entry can be used to predict the screening behaviors prior to cohort entry.

Our numerical studies and the analysis of PROSPR data provide important insights on the practical impact of different approaches for estimating screening rates. When individuals have only limited prior history information in EMR systems, the approaches that ignore potential missing information may introduce substantial bias and loss of statistical power. Calculation based on the subset of individuals with complete prior history often alleviates the bias, as long as these individuals are representative of the underlying screening population. This seemed to be the case for KP California, where a substantial fraction of the population indeed has EMRs for over 10 years and such a subgroup could yield estimates similar to these from the proposed method. When a majority of the patients only have partial prior history available to the research cohort, such as the case for KP Washington PROSPR data repository, a calculation based on individuals having only four to five years of data might underestimate screening rates especially for colonoscopy as tests may occur five to 10 years prior to the enrollment. In this situation our proposed methods can assist in overcoming these potential shortcomings.

There are future directions one may consider along this line of research. We have, in our modeling and in an example using real-world data, considered only patient-level characteristics. Often, screening delivery is influenced by factors at multiple levels, ranging from patients, health-care providers and clinics to specific health-care systems. Therefore, incorporating information about patient interactions with these various levels of care delivery could improve the fit of the model and further enhance our understanding of the quality of screening delivery in health-care systems over time.

**7. Acknowledgments.** This work was funded by NCI at the NIH (grant no. U01CA163304 and P30 CA015704, to W. E. Barlow and Y. Zheng; U54CA163261 to J. Chubak, S. Shortreed; U54CA163262, to D. A. Corley, C. A. Doubeni, and A. G. Zauber; P30CA008748 to A. G. Zauber; U54CA163308, to E. A. Halm; and U54CA163307, to T. D. Tosteson). The authors thank the participating PROSPR Research Centers for the data they have provided for this study. A list of the PROSPR investigators and contributing research staff is provided at: https://healthcaredelivery.cancer.gov/prospr/acknowledgements.html.

## SUPPLEMENTARY MATERIAL

**Supplement: "Parameters used in the simulation"** (DOI: 10.1214/20-AOAS1342SUPP; .pdf). Supplementary tables referenced in Section 4 are available with this paper at the journal's website.

## REFERENCES


BRUNELLI, S. M., GAGNE, J. J., HUYBRECHTS, K. F., WANG, S. V., PATRICK, A. R., ROTHMAN, K. J. and SEEGER, J. D. (2013). Estimation using all available covariate information versus a fixed look-back window for dichotomous covariates. *Pharmacoepidemiol. Drug Saf.* **22** 542–550.

BURNETT-HARTMAN, A. N., MEHTA, S. J., ZHENG, Y., GHAI, N. R., MCLERRAN, D. F., CHUBAK, J., QUINN, V. P., SKINNER, C. S., CORLEY, D. A. et al. (2016). Racial/ethnic disparities in colorectal cancer screening across healthcare systems. *Am. J. Prev. Med.* **51** e107–e115.

CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845. MR1380818 https://doi.org/10.1093/biomet/82.4.835

CHUBAK, J. and HUBBARD, R. (2016). Defining and measuring adherence to cancer screening. *Journal of Medical Screening* **23** 1–1.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 https://doi.org/10.1093/biomet/62.2.269

FAREWELL, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* 1041–1046.

HUBBARD, R. A., CHUBAK, J. and RUTTER, C. M. (2014). Estimating screening test utilization using electronic health records data. *EGEMs* **2**.

HUBBARD, R. A., JOHNSON, E., CHUBAK, J., WERNLI, K. J., KAMINENI, A., BOGART, A. and RUTTER, C. M. (2017). Accounting for misclassification in electronic health records-derived exposures using generalized linear finite mixture models. *Health Serv. Outcomes Res. Methodol.* **17** 101–112. https://doi.org/10.1007/s10742-016-0149-5

KLABUNDE, C. N., ZHENG, Y., QUINN, V. P., BEABER, E. F., RUTTER, C. M., HALM, E. A., CHUBAK, J., DOUBENI, C. A., HAAS, J. S. et al. (2016). Influence of age and comorbidity on colorectal cancer screening in the elderly. *Am. J. Prev. Med.* **51** e67–e75.

KLEINBAUM, D. G. and KLEIN, M. (2002). Maximum likelihood techniques: An overview. *Logistic Regression: A Self-Learning Text* 101–124.

KOHN, L. T., CORRIGAN, J. M., DONALDSON, M. S. et al. (2000). To err is human: Building a safer health system. a report of the committee on quality of health care in America, Institute of Medicine.

LANDON, B. E., O'MALLEY, A. J. and KEEGAN, T. (2010). Can choice of the sample population affect perceived performance: Implications for performance assessment. *J. Gen. Intern. Med.* **25** 104–109.

LIN, J. S., PIPER, M. A., PERDUE, L. A., RUTTER, C. M., WEBBER, E. M., O'CONNOR, E., SMITH, N. and WHITLOCK, E. P. (2016). Screening for colorectal cancer: Updated evidence report and systematic review for the us preventive services task force. *JAMA* **315** 2576–2594.

LITTLE, R. J. A. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data*. Wiley, Hoboken, NJ.

RUTTER, C. M., GREENLEE, R. T., JOHNSON, E., STARK, A., WEINMANN, S., KAMINENI, A., ADAMS, K. and DOUBENI, C. A. (2015). Prevalence of colonoscopy before age 50. *Prev. Med.* **72** 126–129.

SHORTREED, S., JOHNSON, E., RUTTER, C., KAMINEI, A., WERNLI, K. and CHUBAK, J. (2016). Cohort restriction based on prior enrollment: Examining potential biases in estimating cancer and mortality risk. *Observational Studies*. **2** 51–64.

SY, J. P. and TAYLOR, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics* **56** 227–236. MR1767631 https://doi.org/10.1111/j.0006-341X.2000.00227.x

TIRO, J. A., KAMINENI, A., LEVIN, T. R., ZHENG, Y., SCHOTTINGER, J. S., RUTTER, C. M., CORLEY, D. A., SKINNER, C. S., CHUBAK, J. et al. (2014). The colorectal cancer screening process in community settings: A conceptual model for the population-based research optimizing screening through personalized regimens consortium. *Cancer Epidemiol. Biomark. Prev.* **23** 1147–1158.

ZHENG, Y., CORLEY, D. A., DOUBENI, C., HALM, E., SHORTREED, S. M., BARLOW, W. E., ZAUBER, A., TOSTESON, T. D., and CHUBAK, J., (2020). Supplement to "Analyses of preventive care measures with incomplete historical data in electronic medical records: An example from colorectal cancer screening." https://doi.org/10.1214/20-AOAS1342SUPP