Estimating the prevalence of atrial fibrillation from a three-class mixture model for repeated diagnoses

Liang Li[*,1], **Huzhang Mao**[2], **Hemant Ishwaran**[3], **Jeevanantham Rajeswaran**[4], **John Ehrlinger**[4], and **Eugene H. Blackstone**[5]

[1] Department of Biostatistics, MD Anderson Cancer Center, Houston, Texas
[2] Department of Biostatistics, The University of Texas Health Science Center at Houston, Houston, Texas
[3] Department of Biostatistics, University of Miami, Miami, Florida
[4] Department of Quantitative Health Sciences, Cleveland Clinic, Cleveland, Ohio
[5] Department of Cardiothoracic Surgery, Cleveland Clinic, Cleveland, Ohio

Received 22 April 2016; revised 2 September 2016; accepted 28 October 2016

Atrial fibrillation (AF) is an abnormal heart rhythm characterized by rapid and irregular heartbeat, with or without perceivable symptoms. In clinical practice, the electrocardiogram (ECG) is often used for diagnosis of AF. Since the AF often arrives as recurrent episodes of varying frequency and duration and only the episodes that occur at the time of ECG can be detected, the AF is often underdiagnosed when a limited number of repeated ECGs are used. In studies evaluating the efficacy of AF ablation surgery, each patient undergoes multiple ECGs and the AF status at the time of ECG is recorded. The objective of this paper is to estimate the marginal proportions of patients with or without AF in a population, which are important measures of the efficacy of the treatment. The underdiagnosis problem is addressed by a three-class mixture regression model in which a patient's probability of having no AF, paroxysmal AF, and permanent AF is modeled by auxiliary baseline covariates in a nested logistic regression. A binomial regression model is specified conditional on a subject being in the paroxysmal AF group. The model parameters are estimated by the Expectation-Maximization (EM) algorithm. These parameters are themselves nuisance parameters for the purpose of this research, but the estimators of the marginal proportions of interest can be expressed as functions of the data and these nuisance parameters and their variances can be estimated by the sandwich method. We examine the performance of the proposed methodology in simulations and two real data applications.

Keywords: Atrial fibrillation; Latent class model; Mixture model; Two-part model; Zero-inflated binomial.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Modeling the atrial fibrillation

Atrial fibrillation (AF) affects approximately 2.2 million individuals in the United States (Fuster, 2006), particularly those with structural heart disease and the elderly. It is characterized by rapid and irregular heart beat. Each AF episode may last between a few seconds to a few days, and the frequency also varies with subjects. If the AF episodes occur intermittently, they are called paroxysmal AF; if the patient is constantly in AF, this condition is called permanent AF. Most AF episodes occur

*Corresponding author: e-mail: LLi15@mdanderson.org

without any symptoms, while some may be accompanied by perceived heart palpitations, weakness, shortness of breath, or chest pain. AF patients may suffer from tachycardia, low cardiac output from loss of atrial function, atrial and ventricular remodeling, and significantly elevated risk of stroke. Accurately detecting the start and end of the AF episodes is difficult due to the possible intermittent recurrence of the episodes. While devices such as the insertable cardiac monitors can record the AF episodes continuously over a long period of time, these devices require a surgery to be implanted on the patient, and another surgery to be extracted out at the end of the monitoring. Hence, they are not suitable for use in the general at-risk population. In the current clinical practice, physicians often rely on electrocardiogram (ECG) or Holter monitoring to capture the AF episodes. The ECG is usually performed at inpatient visits and lasts for less than 15 minutes; the Holter monitoring device may be carried at home by the patient and provides 24–72 hours of continuous monitoring. Since the gap times between consecutive AF episodes may vary in length between a few minutes to many days, in many situations such monitoring methods can be viewed as only providing a "snapshot" of the underlying continuous on-and-off process of AF. Consequently, the AF episodes are easily missed in these snapshots, leading to underdiagnosis or false-negative diagnosis. It helps if the monitoring frequency and duration are increased (Arya et al., 2007), for example, with the seven-day ECG or daily telephonic ECG, but these options are not always feasible in the current clinical practice.

In this paper, we investigate the diagnosis of AF under the following study design. Suppose n patients, indexed by $i = 1, 2, \ldots, n$, satisfy the inclusion criteria and are enrolled in a study on the risk of AF after a surgical procedure such as the Cox-Maze, with the goal of evaluating whether the surgery reduces or eliminates the AF (Gillinov et al., 2006). After the surgery, each patient is scheduled for n_i recurrent clinical visits; the presence or absence of AF at the j-th visit ($j = 1, 2, \ldots, n_i$) is denoted by Y_{ij}. The patients may be free of AF, have paroxysmal AF with recurrent episodes of arrythmia, or have permanent AF so that the heart is always in abnormal arrythmia. Within the time frame of the study, a patient may be in one of three states, denoted by C_i, with $C_i = 0$ indicating no AF, $C_i = 1/2$ indicating paroxysmal AF, and $C_i = 1$ indicating permanent AF. C_i is not directly observable, but may be inferred from the observed data $\boldsymbol{Y}_i = (Y_{i1}, \ldots, y_{in_i})^T$. The objective of this paper is to estimate the marginal prevalence probabilities

$$p_0 = Pr(C_i = 0), \quad p_{1/2} = Pr(C_i = 1/2), \quad p_1 = Pr(C_i = 1). \tag{1}$$

For a given patient population receiving AF treatments, these estimands quantifies the effectiveness of the surgery in reducing or eliminating AF in that population. An effective procedure is expected to increase the probability of no AF (p_0) and decrease the probabilities of paroxysmal AF ($p_{1/2}$) and permanent AF (p_1). If we can estimate p_0, $p_{1/2}$, and p_1 for a single population, extension to randomized clinical trial or propensity score matched observational studies with multiple treatment groups is straightforward because these marginal prevalence probabilities can be estimated for each group under comparison and the estimators and their estimated variances can be compared or used to form a statistical test for group differences.

Let $M_i = \sum_{j=1}^{n_i} Y_{ij}$ denote the number of positive ECG results. A simple method to estimate the marginal prevalence is to claim that a patient is free of AF if $M_i = 0$, in permanent AF if $M_i = n_i$, and in paroxysmal AF if $0 < M_i < n_i$. The idea of this approach conforms to some clinical practice where a patient is diagnosed with AF if the AF episodes are captured by the ECGs, and non-AF otherwise. According to this approach, the estimated marginal prevalence probabilities are:

$$\tilde{p}_0 = \frac{1}{n}\sum_{i=1}^{n} 1\{M_i = 0\}, \ \tilde{p}_{1/2} = \frac{1}{n}\sum_{i=1}^{n} 1\{0 < M_i < n_i\}, \ \tilde{p}_1 = \frac{1}{n}\sum_{i=1}^{n} 1\{M_i = n_i\}. \tag{2}$$

These prevalence estimates may be biased. By definition, if a patient is free of AF, $Y_{ij} \equiv 0$ and $M_i = 0$; if a patient is in permanent AF, $Y_{ij} \equiv 1$ and $M_i = n_i$; if a patient is in paroxysmal AF, Y_{ij} may

be 0 or 1 with nonzero probabilities. By chance, some paroxysmal AF patients may have all the Y_{ij}'s being 0, which leads them to be incorrectly classified into the non-AF group. Similarly, by chance some paroxysmal AF patients may have all the Y_{ij}'s being 1, which leads them to be incorrectly classified into the permanent AF group. Mathematically, $p_0 \leq \tilde{p}_0$, $p_{1/2} \geq \tilde{p}_{1/2}$, $p_1 \leq \tilde{p}_1$. Unless n_i is very large for every subject, the equalities in these expressions are unlikely to hold even with large sample size n. This analysis explains a widely observed phenomenon in AF research that AF detection depends on the monitoring intensity: the more frequent the AF is monitored, the higher the estimated AF prevalence becomes (Senatore et al., 2006; Arya et al., 2007; Gaillard et al., 2010).

The estimands of interest (1) cannot be identified without additional auxiliary information. There are well-established prognosis risk factors of AF. In this paper, we consider exploiting their relationship with C_i and $\boldsymbol{Y}_i$ (i.e., M_i) to improve the estimation of these marginal prevalence probabilities. Let $\boldsymbol{X}_i$ be a vector of baseline prognostic factors. Suppose we can specify a hierarchical two-stage regression model for the data,

$$Pr(C_i = c|\boldsymbol{X}_i; \boldsymbol{\theta}) \quad \text{and} \quad Pr(M_i|C_i = c, \boldsymbol{X}_i; \boldsymbol{\theta}),$$

where $c = 0, 1/2, 1$ and $\boldsymbol{\theta}$ is a generic notation for all the unknown parameters in this model. For the purpose of this research, $\boldsymbol{\theta}$ are nuisance parameters, because the estimands of interest are $p_0, p_{1/2}, p_1$. We first estimate $\boldsymbol{\theta}$ by maximizing the log-likelihood of the data $\{M_i, \boldsymbol{X}_i; i = 1, 2, \ldots, n\}$:

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log Pr(M_i|\boldsymbol{X}_i) = \sum_{i=1}^{n} \log \left\{ \sum_{c\in\{0,1/2,1\}} Pr(M_i|C_i = c, \boldsymbol{X}_i) Pr(C_i = c|\boldsymbol{X}_i) \right\}. \tag{3}$$

Denote the maximum-likelihood estimator of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$. Then we propose the following estimator for the marginal prevalence probabilities $p_0, p_{1/2}, p_1$:

$$\begin{aligned}
\hat{p}_0 &= \widehat{Pr}(C_i = 0) = \frac{1}{n}\sum_{i=1}^{n} Pr(C_i = 0|M_i, \boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) \\
\hat{p}_{1/2} &= \widehat{Pr}(C_i = 1/2) = \frac{1}{n}\sum_{i=1}^{n} Pr(C_i = 1/2|M_i, \boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) \\
\hat{p}_1 &= \widehat{Pr}(C_i = 1) = \frac{1}{n}\sum_{i=1}^{n} Pr(C_i = 1|M_i, \boldsymbol{X}_i; \hat{\boldsymbol{\theta}}).
\end{aligned} \tag{4}$$

These are the sample averages of each subject's conditional probability of C_i given the observed data, evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. We can prove that these estimators are unbiased for $p_0, p_{1/2}, p_1$, using $E\{Pr(C_i = c|M_i, \boldsymbol{X}_i; \boldsymbol{\theta})\} = E\{1\{C_i = c\}\} = Pr(C_i = c)$ and the consistency of $\hat{\boldsymbol{\theta}}$ for the nuisance parameter $\boldsymbol{\theta}$.

In Section 2, we provide details on the model and estimation procedure. Section 3 presents the simulation results. Section 4 includes two real data examples to illustrate the proposed method. Discussion is in Section 5.

2 Estimation

We first describe the estimation of the nuisance parameter $\boldsymbol{\theta}$ through an EM algorithm, and then discuss the point and variance estimation of the estimands of interest p_0, $p_{1/2}$, p_1. Throughout this paper, we use the following model for $Pr(C_i|\boldsymbol{X}_i)$:

$$\begin{aligned}
Pr(C_i = 0|\boldsymbol{X}_i) &= \frac{1}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})} \\
Pr(C_i = 1/2|\boldsymbol{X}_i) &= Pr(C_i = 1/2 \text{ or } 1|\boldsymbol{X}_i) \times Pr(C_i = 1/2|\boldsymbol{X}_i, C_i = 1/2 \text{ or } 1) = \\
&= \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})}\frac{1}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})} \\
Pr(C_i = 1|\boldsymbol{X}_i) &= Pr(C_i = 1/2 \text{ or } 1|\boldsymbol{X}_i) \times Pr(C_i = 1|\boldsymbol{X}_i, C_i = 1/2 \text{ or } 1) = \\
&= \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})}\frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\beta})}.
\end{aligned}$$

This is a nested logistic regression model, in which we first model the probability of $C_i = 0$ versus $C_i = 1/2$ or 1, and then model the probability of $C_i = 1$ given $C_i = 1/2$ or 1. For notational convenience, we assume that the first element in $\boldsymbol{X}_i$ is 1, corresponding to the intercept. With one more set of parameters, the nested logistic regression model is in general more flexible than some widely used models such as the proportional odds model, but is also less parsimonious and less transparent to interpret. Since we are interested in the marginal probabilities, and all the regression parameters $\boldsymbol{\theta}$ are nuisance, we sacrifice some parsimony and interpretability to gain flexibility when making the model choice.

The conditional distribution of M_i given C_i and $\boldsymbol{X}_i$ may be specified as:

$$\begin{aligned}
&Pr(M_i|C_i = 0, \boldsymbol{X}_i) = 1\{M_i = 0\} \\
&Pr(M_i|C_i = 1/2, \boldsymbol{X}_i) = \binom{n_i}{M_i} p_i^{M_i}(1-p_i)^{n_i - M_i} = \\
&\qquad = \binom{n_i}{M_i} \exp\left\{M_i\boldsymbol{X}_i^T\boldsymbol{\gamma} - n_i \log\left(1+e^{X_i^T\gamma}\right)\right\} \\
&\qquad p_i = \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\gamma})}{1+\exp(\boldsymbol{X}_i^T\boldsymbol{\gamma})} \\
&Pr(M_i|C_i = 1, \boldsymbol{X}_i) = 1\{M_i = n_i\}.
\end{aligned} \tag{5}$$

When the patient does not have AF ($C_i = 0$), M_i can only be 0; when the patient is in permanent AF ($C_i = 1$), M_i can only be n_i; when the patient is in paroxysmal AF ($C_i = 1/2$), M_i follows a binomial distribution with probability p_i, which depends on $\boldsymbol{X}_i$. An alternative way to model (5) is to specify the conditional distribution of $\boldsymbol{Y}_i$ given $C_i = 1/2$ and $\boldsymbol{X}_i$ through a regression model for longitudinal binary outcomes, incorporating the time effect and intrasubject correlation. Under that model specification, the estimation of marginal prevalence probabilities via (4) is still valid. However, the purpose of this article is to study the diagnosis of AF during a period of time when the patient's AF status is stable. For that purpose, the binomial model above is more parsimonious and interpretable.

The nuisance parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ can be estimated by maximizing the log-likelihood (3) in an EM algorithm. We define the complete data set as $\{C_i, M_i, \boldsymbol{X}_i; i = 1, 2, \ldots, n\}$, where C_i is unobserved. There are only three possibilities in the complete data set: (1) $C_i = 0, M_i = 0$; (2) $C_i = 1/2$,

$0 \le M_i \le n_i$; (3) $C_i = 1, M_i = n_i$; the case with $C_i = 0, M_i > 0$, or $C_i = 1, M_i < n_i$ does not exist. Hence, the complete data log-likelihood is:

$$\sum_{i=1}^{n}\Big\{1\{C_i=0\}\log Pr(C_i=0|\boldsymbol{X}_i)+1\{C_i=1/2\}\log Pr(C_i=1/2|\boldsymbol{X}_i)Pr(M_i|\boldsymbol{X}_i,C_i=1/2)\Big\}+$$

$$+1\{C_i=1\}\log Pr(C_i=1|\boldsymbol{X}_i)\Big\}=$$

$$=\sum_{i=1}^{n}\Bigg\{1\{C_i=0\}(-1)\log\left(1+e^{X_i^T\boldsymbol{\alpha}}\right)+1\{C_i=1/2\}\Bigg[\log\left(\frac{e^{X_i^T\boldsymbol{\alpha}}}{1+e^{X_i^T\boldsymbol{\alpha}}}\frac{1}{1+e^{X_i^T\boldsymbol{\beta}}}\right)+$$

$$+\log\binom{n_i}{M_i}+M_i\boldsymbol{X}_i^T\boldsymbol{\gamma}-n_i\log\left(1+e^{X_i^T\boldsymbol{\gamma}}\right)\Bigg]+1\{C_i=1\}\log\left(\frac{e^{X_i^T\boldsymbol{\alpha}}}{1+e^{X_i^T\boldsymbol{\alpha}}}\frac{e^{X_i^T\boldsymbol{\beta}}}{1+e^{X_i^T\boldsymbol{\beta}}}\right)\Bigg\}.$$

The following is the EM algorithm:

- (Initial step) Obtain initial values for $\boldsymbol{\theta}$. The initial value for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be obtained by fitting a nested logistic regression of $1\{M_i=0\}$, $1\{0<M_i<n_i\}$ and $1\{M_i=n_i\}$ on $\boldsymbol{X}_i$. The initial values for $\boldsymbol{\gamma}$ can be obtained by fitting a binomial regression of M_i on $\boldsymbol{X}_i$ for the subset of data with $0<M_i<n_i$.
- (E-step) Given the current value of $\boldsymbol{\theta}^{(m)}$, calculate the conditional expectation of the indicators in the complete data log-likelihood given the observed data M_i and conditional on $\boldsymbol{X}_i$.

$$Pr(C_i=c|M_i,\boldsymbol{X}_i)=\frac{Pr(M_i|C_i=c,\boldsymbol{X}_i)Pr(C_i=c|\boldsymbol{X}_i)}{\sum_{c'\in\{0,1/2,1\}}Pr(M_i|C_i=c',\boldsymbol{X}_i)Pr(C_i=c'|\boldsymbol{X}_i)},\quad (c=0,1/2,1).$$

Denote

$$\begin{aligned}p_{0i}^{(m)}&=Pr(C_i=0|M_i,\boldsymbol{X}_i;\boldsymbol{\theta}^{(m)})\\ p_{1/2,i}^{(m)}&=Pr(C_i=1/2|M_i,\boldsymbol{X}_i;\boldsymbol{\theta}^{(m)})\\ p_{1i}^{(m)}&=Pr(C_i=2|M_i,\boldsymbol{X}_i;\boldsymbol{\theta}^{(m)}).\end{aligned}\tag{6}$$

- (M-step) Given $p_{0i}^{(m)}$, $p_{1/2,i}^{(m)}$, and $p_{1i}^{(m)}$, we maximize the following expected log complete data likelihood with respect to $\boldsymbol{\theta}$:

$$\sum_{i=1}^{n}\Bigg\{p_{0i}^{(m)}(-1)\log\left(1+e^{X_i^T\boldsymbol{\alpha}}\right)+p_{1/2,i}^{(m)}\Bigg[\log\left(\frac{e^{X_i^T\boldsymbol{\alpha}}}{1+e^{X_i^T\boldsymbol{\alpha}}}\frac{1}{1+e^{X_i^T\boldsymbol{\beta}}}\right)+\log\binom{n_i}{M_i}+$$

$$+M_i\boldsymbol{X}_i^T\boldsymbol{\gamma}-n_i\log\left(1+e^{X_i^T\boldsymbol{\gamma}}\right)\Bigg]+p_{1i}^{(m)}\log\left(\frac{e^{X_i^T\boldsymbol{\alpha}}}{1+e^{X_i^T\boldsymbol{\alpha}}}\frac{e^{X_i^T\boldsymbol{\beta}}}{1+e^{X_i^T\boldsymbol{\beta}}}\right)\Bigg\}.$$

Since the terms involving $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are linearly additive, we can maximize with respect to $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ separately in three M-steps. Each step involves maximization of a concave function with closed form first and second derivatives, and can be completed using the Newton–Raphson method.
- Iterate between E-Step and M-Step until the following convergence criteria is satisfied:

$$\arg\max|\boldsymbol{\theta}^{(m)}-\boldsymbol{\theta}^{(m-1)}|<\epsilon_0.$$

That is, we iterate till all the elements in the vector $|\boldsymbol{\theta}^{(m)} - \boldsymbol{\theta}^{(m-1)}|$ be less than a prespecified small positive number ϵ_0. We set $\epsilon_0 = 10^{-6}$ for the numerical study in this paper. If it converges at step K, we denote the final point estimator as $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(K)}$.

Once the nuisance parameters $\hat{\boldsymbol{\theta}}$ is obtained, we can estimate the marginal prevalence probabilities (1) according to (4):

$$\hat{p}_0 = \frac{1}{n}\sum_{i=1}^{n} p_{0i}^{(K)}, \quad \hat{p}_{1/2} = \frac{1}{n}\sum_{i=1}^{n} p_{1/2,i}^{(K)}, \quad \hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} p_{1i}^{(K)},$$

where the individual conditional probabilities $p_{0i}^{(K)}$, $p_{1i}^{(K)}$, and $p_{2i}^{(K)}$ can be obtained directly from the last (the K-th) iteration of the EM algorithm.

The asymptotic variance matrix of $\hat{\boldsymbol{\theta}}$ can be estimated by the inverse of the negative Hessian matrix, obtained by numerically differentiating the log-likelihood (3). However, for the purpose of this research, $\boldsymbol{\theta}$ is nuisance parameter, and the estimands of interest are p_0, $p_{1/2}$, and p_1. We use the following partial M-estimation procedure (Stefanski and Boos, 2002) to estimate the variances of $\hat{p}_0$, $\hat{p}_{1/2}$, and $\hat{p}_1$. The estimator $\hat{p}_0$ and $\hat{\boldsymbol{\theta}}$ can be viewed as solutions to the following estimating equation for $\boldsymbol{\kappa} = (p_0, \boldsymbol{\theta}^T)^T$:

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{\phi}_i(\boldsymbol{\kappa}) = \sum_{i=1}^{n} \begin{bmatrix} Pr(C_i = 0|M_i, \boldsymbol{X}_i; \boldsymbol{\theta}) - p_0 \\ \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\sum_{c\in\{0,1/2,1\}} Pr(M_i|C_i = c, \boldsymbol{X}_i) Pr(C_i = c|\boldsymbol{X}_i)\right) \end{bmatrix}.$$

By the sandwich method, $\widehat{\text{var}}(\hat{p}_0)$ is the first diagonal entry of the matrix $\widehat{\text{var}}(\hat{\boldsymbol{\kappa}}) = n^{-1}\hat{\boldsymbol{A}}_n^{-1}\hat{\boldsymbol{B}}_n\hat{\boldsymbol{A}}_n^{-T}$, with $\hat{\boldsymbol{A}}_n = n^{-1}\sum_{i=1}^{n} \partial\boldsymbol{\phi}_i(\boldsymbol{\kappa})/\partial\boldsymbol{\kappa}^T|_{\boldsymbol{\kappa}=\hat{\boldsymbol{\kappa}}}$ and $\hat{\boldsymbol{B}}_n = n^{-1}\sum_{i=1}^{n} \boldsymbol{\phi}_i(\boldsymbol{\kappa})\boldsymbol{\phi}_i(\boldsymbol{\kappa})^T|_{\boldsymbol{\kappa}=\hat{\boldsymbol{\kappa}}}$. The derivatives can be calculated numerically. Variances of $\hat{p}_{1/2}$ and $\hat{p}_1$ are estimated in similar ways.

3 Simulation

We conducted simulations to study the performance of the proposed model and method. At each simulation, we generated two standard Gaussian baseline covariates with a correlation coefficient of 0.2. We consider a setting in which the three marginal prevalence probabilities, $p_0 = Pr(C_i = 0)$, $p_{1/2} = Pr(C_i = 1/2)$, and $p_1 = Pr(C_i = 1)$, are close to each other and a setting in which p_0 and p_1 are much smaller than $p_{1/2}$. Under the first setting, $\boldsymbol{\alpha} = (1.2, -1.0, 1.0)^T$, $\boldsymbol{\beta} = (-0.2, 0.6, -0.8)^T$, $\boldsymbol{\gamma} = (0.3, 1.2, 1)^T$; under the second setting, $\boldsymbol{\alpha} = (2.5, -1.0, 1.0)^T$, $\boldsymbol{\beta} = (-2.0, 0.6, -0.8)^T$, $\boldsymbol{\gamma} = (0.3, 1.2, 1)^T$. We varied the sample size between $n = 250$ and 1000, and varied the median number of repeated measures (n_i) between 5 (range 1–9 in a uniform distribution) and 10 (range 5–15). Therefore, we had a total of eight simulation scenarios. One thousand simulations were run in each setting. We compared the naive method, given by (2) and the proposed estimator (4).

The simulation results are summarized in Table 1. The naive method generally leads to biased estimator of marginal prevalence probabilities and sometimes the bias is quite large, while the proposed method is unbiased in all settings. The bias of the naive method decreases when there are more repeated measures per subject, and increasing the sample size does not help reducing bias. When there is a large number of diagnoses per subject, the chance that all these diagnosis fail to capture any AF episodes on a paroxysmal AF patient is greatly reduced, and so is the chance that these diagnoses capture at least one AF episode on the paroxysmal AF patient. The 95% confidence intervals are correct coverage probabilities, though slight under coverage can be seen in small data sets with small sample size and number of repeated measures per subject. The simulation also shows that the proposed estimators of the nuisance parameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$) are nearly unbiased under all the simulation scenarios (result omitted).

Table 1 Simulation results comparing the naive and the proposed methods in estimating marginal prevalence probabilities p_0, $p_{1/2}$, and p_1.

Setting	n	n_i	Method	p_0				$p_{1/2}$				p_1			
				Est	EmpSD	AvgSE	CI	Est	EmpSD	AvgSE	CI	Est	EmpSD	AvgSE	CI
1			True value	0.2838				0.4276				0.2886			
	250	Small	Naive	0.3587	0.0302			0.2434	0.0264			0.3980	0.0310		
	250	Small	Proposed	0.2825	0.0340	0.0332	0.936	0.4321	0.0444	0.0450	0.945	0.2854	0.0410	0.0408	0.932
	250	Large	Naive	0.3144	0.0301			0.3428	0.0307			0.3428	0.0303		
	250	Large	Proposed	0.2829	0.0311	0.0300	0.937	0.4293	0.0373	0.0364	0.934	0.2878	0.0335	0.0331	0.941
	1000	Small	Naive	0.3594	0.0150			0.2416	0.0135			0.3990	0.0153		
	1000	Small	Proposed	0.2838	0.0164	0.0166	0.944	0.4279	0.0224	0.0224	0.941	0.2883	0.0204	0.0204	0.951
	1000	Large	Naive	0.3153	0.0146			0.3410	0.0150			0.3437	0.0149		
	1000	Large	Proposed	0.2840	0.0151	0.0150	0.944	0.4272	0.0185	0.0182	0.945	0.2888	0.0166	0.0166	0.941
2			True value	0.1211				0.7627				0.1162			
	250	Small	Naive	0.2574	0.0282			0.4314	0.0315			0.3112	0.0296		
	250	Small	Proposed	0.1176	0.0273	0.0279	0.937	0.7606	0.0444	0.0430	0.939	0.1176	0.0361	0.0352	0.921
	250	Large	Naive	0.1777	0.0246			0.6100	0.0305			0.2123	0.0259		
	250	Large	Proposed	0.1212	0.0230	0.0230	0.944	0.7615	0.0332	0.0332	0.951	0.1173	0.0263	0.0263	0.944
	1000	Small	Naive	0.2578	0.0138			0.4317	0.0159			0.3105	0.0153		
	1000	Small	Proposed	0.1218	0.0141	0.0137	0.944	0.7611	0.0211	0.0210	0.944	0.1170	0.0173	0.0173	0.949
	1000	Large	Naive	0.1785	0.0122			0.6090	0.0159			0.2124	0.0129		
	1000	Large	Proposed	0.1218	0.0115	0.0115	0.944	0.7614	0.0162	0.0164	0.948	0.1168	0.0126	0.130	0.951

The results are reported as the average (Est) and empirical standard deviation (empSD) of the point estimators, the average estimated standard errors (AvgSE), and the coverage probability of 95% confidence intervals (CI).

The proposed mixture regression model is a working model to help identify the estimands in (1), which are otherwise unidentified from the data. Hence, we conducted additional simulation to study the proposed estimators when the working model is misspecified. The simulation was designed in a similar setup as in Setting 1 of Table 1. The sample size is fixed at 500. The median number of repeated measures is 5 (range 1–9), 10 (range 5–15), or 20 (range 10–30). We fit the correctly specified three-part mixture model as well as an incorrectly specified model omitting X_2. The result is visualized in Fig. 1. The result reconfirms the finding in Table 1 that the correctly specified mixture model produces unbiased estimators. When the model is misspecified, there is bias in the proposed estimator, but the bias is much smaller than the naive estimator in all scenarios. When the number of repeated measures per subject increases, the bias of the naive method decreases, and so does the bias of the proposed method under misspecified model. A heuristical explanation is that when the monitoring frequency increases, it is difficult for a paroxysmal AF patient to have $\boldsymbol{Y}_i \equiv 0$. In other words, there is less evidence in the data for a paroxysmal AF patient to have a high probability of being in the other two groups, even when there is some model misspecification. Hence the proposed estimator converges with the naive estimator, whose bias decreases with increasing monitoring frequency. Therefore, based on this result, we recommend the use of the proposed methodology in all scenarios in replacement of the naive method, regardless of the monitoring frequency.

The simulations in this section were performed using functions written in `R` (R Core Team, 2014). The R code is available for download at the publisher's website. The initial values of the EM algorithm were chosen using the simple method described in Section 2, and convergence was achieved for all the simulation runs. Analyzing a typical simulated data set with $n = 1000$ took about 80 seconds on a PC with 3.30 GHz CPU and 16 GB memory. The good convergence performance may be partly attributed to the concavity of the target functions in the M-step of the EM algorithm. Previous literature suggests that it is desirable to run the EM algorithm using different initial values to protect against convergence to local, instead of global maximum (Biernachi et al., 2003). We intentionally let the initial values deviate from the recommended values in Section 2. The algorithm always converged to the same estimators, unless excessively large deviations were used that led to unreasonable initial values, in which case the algorithm did not converge.

4 Applications

We illustrate the proposed methodology with two real data applications. The first application is from a retrospective observational study on the effectiveness of three different surgical ablation procedures in reducing AF: the Cox-Maze ablation procedure, which is a cut-and-sew procedure, considered as the gold standard, but is more complicated and time consuming; the pulmonary vein isolation, which is the simplest; pulmonary vein isolation alone with added lesions, which is between the other two procedures in terms of complexity. Data on heart rhythm were collected from postoperative ECGs and extracted from the electronic health records (Gillinov et al., 2006). The data set includes 413 patients with at least one ECG diagnosis between six and 24 months. The ECG data within the first six months after surgery have been excluded as patients' heart rhythms may be affected by many other perioperative risk factors and are not stable shortly after the open heart surgery. Of these patients, 169 had one ECG, 105 has two, 49 had 3, and the maximum number of ECGs is 14. The naive method estimated that $\hat{p}_0 = 0.564$, $\hat{p}_{1/2} = 0.196$, $\hat{p}_1 = 0.240$. The proposed method estimated that $\hat{p}_0 = 0.379(0.0171)$, $\hat{p}_{1/2} = 0.481(0.0208)$, $\hat{p}_1 = 0.140(0.00682)$ (the numbers in brackets are standard errors). The proportion of permanent AF patients appears to be overestimated under the naive method. Table 2 shows the estimated regression model parameters. These are nuisance for the purpose of estimating the marginal prevalence probabilities, but can be informative for exploring the association between risk factors and outcomes.

The second data application is from a prospective randomized study to determine if adding surgical ablation to mitral valve surgery is more effective than surgery alone in reducing AF at six and

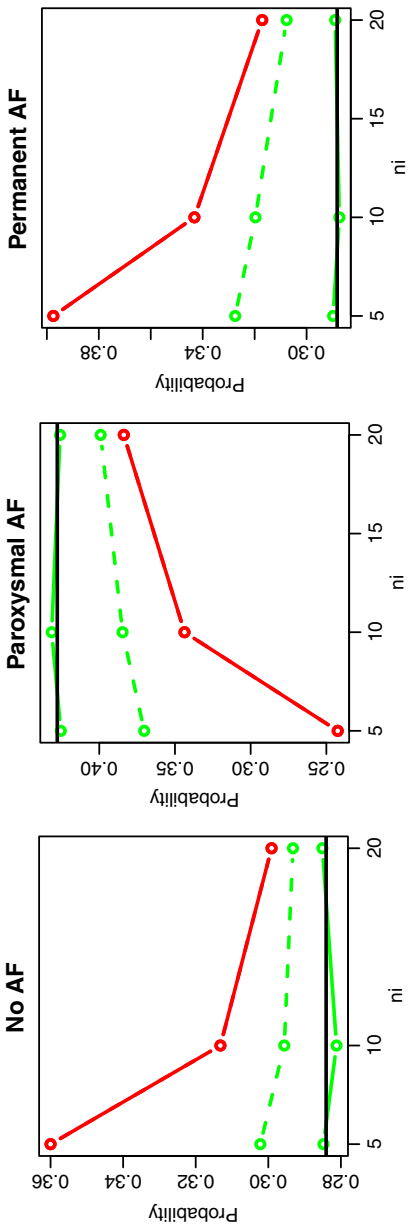

Figure 1 The average estimated marginal prevalence probabilities from correctly specified three-part mixture model (green, solid line), misspecified three-part mixture model (green, dashed line), and the naive method (red), when the median number of repeated measures (n_i, horizontal axis) is 5, 10, or 20. The black horizontal line indicates the true marginal prevalence probabilities.

Table 2 The estimated model parameters from the three-part mixture model applied to the first data application.

Parameter	Estimator	Std Err	p-Value
α_0	−4.65	2.13	0.029
α_1	−0.00329	0.0206	0.87
α_2	0.207	0.176	0.24
α_3	0.685	0.307	0.026
α_4	0.546	0.543	0.31
α_5	1.62	0.443	< 0.001
β_0	−8.40	2.83	0.003
β_1	0.0410	0.0279	0.14
β_2	0.495	0.260	0.057
β_3	0.312	0.289	0.28
β_4	1.71	1.04	0.10
β_5	0.521	0.706	0.460
γ_0	−2.63	1.47	0.074
γ_1	0.0299	0.0150	0.046
γ_2	0.111	0.122	0.36
γ_3	−0.00248	0.205	0.99
γ_4	0.174	0.697	0.80
γ_5	−0.324	0.308	0.29

Results are presented as the point estimator, standard error estimator, and p-value from a Wald-type test against the null hypothesis of zero parameter. The subscripts of each parameter represent: 0: intercept, 1: patient age, 2: log of the average duration of AF, 3: left atrial diameter (mm), 4: pulmonary vein isolation alone, 5: pulmonary vein isolation with added lesions.

12 months. In this study, patients underwent weekly transtelephonic monitoring (TTM) of their heart rhythm after cardiac surgery. The TTM system functions in a similar way as the ECG in the physician's office, but it is smaller and more portable so that patients can use it to monitor their heart rhythm at home. The device can be connected to the home telephone line to transmit the heart rhythm data to the data center of the study. The patients were instructed to make at least one transmission each week, though more frequent transmissions were permitted. There are $n = 150$ patients with weekly TTM data between six and 12 months. The median number of repeated measures is 23, minimum 1, and maximum 46. The naive method estimated that $\hat{p}_0 = 0.38$, $\hat{p}_{1/2} = 0.28$, $\hat{p}_1 = 0.34$. The proposed method estimated that $\hat{p}_0 = 0.375(0.0168)$, $\hat{p}_{1/2} = 0.287(0.0128)$, $\hat{p}_1 = 0.339(0.0152)$. Unlike the first example, the results from the proposed method and the naive method are very similar. As shown in the simulations, when the number of repeated measures per subject is large, the discrepancy between the proposed and the naive methods diminishes. In other words, the naive method produces nearly unbiased result with large number of repeated measures. From a clinical perspective, this finding highlights the importance and usefulness of the modern technology such as the TTM system in the diagnosis and monitoring of patients at risk of AF. Table 3 shows the estimated model parameters. As an additional sensitivity analysis, we artificially reduced the number of diagnoses per subject in this data set by randomly sampling a fraction of the repeated measures with a probability of retention φ, and plotted the change in estimated marginal prevalence probabilities from the naive and the proposed methods in Fig. 2. The estimated probabilities from the proposed method are less sensitive to the monitoring frequency than those from the naive method.

5 Discussion

In this paper, we proposed a new statistical methodology to address one aspect of the widely recognized diagnosis problem in AF research, that the estimated prevalence of AF increases with the monitoring intensity (Senatore et al., 2006; Arya et al., 2007). This phenomenon arises because AF episodes occur intermittently and may be easily missed when the monitoring frequency is not adequate. Estimating the

Table 3 The estimated model parameters from the three-part mixture model applied to the second data application.

Parameter	Estimator	Std Err	p-Value
$\boldsymbol{\alpha}_0$	−2.68	1.34	0.045
$\boldsymbol{\alpha}_1$	0.0346	0.0196	0.077
$\boldsymbol{\alpha}_2$	1.46	0.406	< 0.001
$\boldsymbol{\alpha}_3$	1.14	0.626	0.070
$\boldsymbol{\alpha}_4$	0.273	0.447	0.54
$\boldsymbol{\beta}_0$	0.0899	1.83	0.96
$\boldsymbol{\beta}_1$	−0.00474	0.0260	0.86
$\boldsymbol{\beta}_2$	0.0850	0.422	0.84
$\boldsymbol{\beta}_3$	0.0380	0.593	0.95
$\boldsymbol{\beta}_4$	0.436	0.552	0.43
$\boldsymbol{\gamma}_0$	−1.57	0.587	0.0076
$\boldsymbol{\gamma}_1$	0.0260	0.00883	0.0032
$\boldsymbol{\gamma}_2$	−0.801	0.153	< 0.001
$\boldsymbol{\gamma}_3$	1.07	0.217	< 0.001
$\boldsymbol{\gamma}_4$	−0.0465	0.193	0.81

Results are presented as the point estimator, standard error estimator, and p-value from a Wald-type test against the null hypothesis of zero parameter. The subscripts of each parameter represent: 0: intercept, 1: patient age, 2: congestive heart failure, 3: history of cardiovascular disease, 4: hypertension.

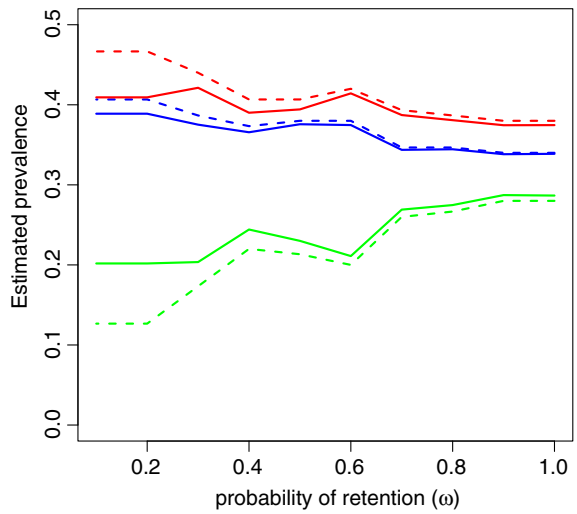

Figure 2 The estimated prevalence probabilities by the naive (dashed lines) and proposed (solid lines) methods, as the probability of retention ω decreases, causing a reduction in the number of repeated measures per subject from the original TTM data set ($\omega = 1$). The estimated probabilities from the proposed method are less sensitive to ω than those from the naive method. Red: probability of no AF (p_0); green: probability of paroxysmal AF ($p_{1/2}$); blue: probability of permanent AF (p_1).

prevalence of people with no AF, paroxysmal AF and permanent AF in a given population is of scientific interest in many contexts including the evaluation of the efficacy of AF ablation surgeries. If these probabilities are estimated based on the observed proportions such as in the naive method, the results will be biased unless a large number of repeated diagnoses are made, such as in the TTM data example. Hence, the results of this article highlight the importance of using modern, high frequency monitoring technology such as the TTM system. However, high frequent monitoring is not always available in clinical practice. When the monitoring frequency is not large, we propose to use a model-based approach. Our simulation shows that the proposed approach outperforms the naive approach and produces unbiased result regardless of the monitoring frequency. Even when the model is moderately misspecified, the proposed method can still achieve substantial bias reduction and outperforms the naive method. Our real data applications show that when the monitoring frequency is high as in the second example, the proposed method and the naive method give similar results; when the monitoring frequency is low as in the first data example, there are considerable differences between the two methods but the proposed method is better justified than the naive method. In addition, the result of the proposed method is less sensitive to the monitoring frequency than the naive method, as demonstrated in the real data experiment with the second data example (Fig. 2). Based on all the results above, we recommend the proposed methodology in replacement of the naive method in all situations, regardless of the monitoring frequency.

The marginal prevalence probabilities are not identifiable with only the AF outcome $\boldsymbol{Y}$. They are identifiable in our approach because of the use of the three-part mixture model. This is analogous in a perspective to the cure model in survival analysis (Farewell, 1982), where the cured patients are not expected to have the clinical event of interest and are indistinguishable from the censored patients without additional modeling assumptions. The covariates $\boldsymbol{X}$ can be thought as auxiliary variables, and we make use of them to recover the lost information in the missing data, that is, the group membership C of each patients in the no AF, paroxysmal AF, and permanent AF groups. Among all the patients with $\boldsymbol{Y} \equiv 0$, some are expected to have higher risks than others according to their baseline variables $\boldsymbol{X}$. This information is exploited by the three-part mixture model and is reflected in its calculation of the conditional probability of C_i given the observed data via Eq. (4), which leads to a correction of the bias from the naive method. There are other examples in the statistical literature that use auxiliary variables to recover information lost due to missing data. For example, Faucett et al. (2002) used time-dependent covariates as auxiliary variables to impute the time to event data missing due to censoring. There is some literature on zero-inflated binomial regression, which applies to counts data with a binomial distribution with excessive zero counts (Hall, 2000; Hall and Zhang, 2004). While the proposed methodology in this paper uses a similar EM algorithm to find the estimators, it differs from the literature in the novel motivating application, a focus on the estimation of the marginal prevalence instead of regression coefficients, and the use of three-class mixture model instead of two-part mixture model.

6 Supporting Information

"code.zip" in the Supporting Information contains the code required to generate the simulation results in this paper together with instructions for their use. Available at the publisher's website.

Acknowledgments This research is funded by NIH grants R01HL103552 and P30CA016672. We want to thank the two referees and the Associate Editor for constructive comments that improved this paper.

Conflict of interest
The authors declare no conflict of interest.

References

Arya, A., Piorkowski, C., Sommer, P., Kottkamp, H. and Hindricks, G. (2007). Clinical implications of various follow up strategies after catheter ablation of atrial fibrillation. *Pacing and Clinical Electrophysiology* **30**, 458–462.

Biernachi, C., Celeux, G. and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis* **41**, 561–575.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041–1046.

Faucett, C. L., Schenker, N. and Taylor, J. M. (2002). Survival analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data. *Biometrics* **58**, 37–47.

Fuster, V. (2006). ACC/AHA/ESC 2006 guidelines for the management of patients with atrial fibrillation. *Circulation* **114**, e257–354.

Gaillard, N., Deltour, S., Vilotijevic, B., Hornych, A., Crozier, S., Leger, A., Frank, R. and Samson, Y. (2010). Detection of paroxysmal atrial fibrillation with transtelephonic EKG in TIA or stroke patients. *Neurology* **74**, 1666–1670.

Gillinov, A. M., Bhavani, S., Blackstone, E. H., Rajeswaran, J., Svensson, L. G., Navia, J. L., Pettersson, B. G., Sabik, J. F. 3rd, Smedira, N. G., Mihaljevic, T., McCarthy, P. M., Shewchik, J. and Natale, A. (2006). Surgery for permanent atrial fibrillation: impact of patient factors and lesion set. *Annals of Thoracic Surgery* **82**, 502–513.

Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* **56**, 1030–1039.

Hall, D. B. and Zhang, Z. (2004). Marginal models for zero inflated clustered data. *Statistical Modeling* **4**, 161–180.

R Core Team (2014). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Senatore, G., Stabile, G., Bertaglia, E., Donnici, G., De Simone, A., Zoppo, F., Turco, P., Pascotto, P. and Fazzari, M. (2006). Role of transtelephonic electrocardiographic monitoring in detecting short-term arrhythmia recurrences after radiofrequency ablation in patients with atrial fibrillation. *Journal of the American College of Cardiology* **45**, 873–876.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *American Statistician* **56**, 29–38.

