Published in final edited form as:
Ann Appl Stat. 2023 June ; 17(2): 1220–1238. doi:10.1214/22-AOAS1666.

BAYESIAN ANALYSIS FOR IMBALANCED POSITIVE-UNLABELLED DIAGNOSIS CODES IN ELECTRONIC HEALTH RECORDS

Ru Wang[1,*], **Ye Liang**[1,†], **Zhuqi Miao**[2], **Tieming Liu**[3]

[1]Department of Statistics, Oklahoma State University

[2]School of Business, State University of New York at New Paltz

[3]School of Industrial Engineering and Management, Oklahoma State University

Abstract

With the increasing availability of electronic health records (EHR), significant progress has been made on developing predictive inference and algorithms by health data analysts and researchers. However, the EHR data are notoriously noisy due to missing and inaccurate inputs despite the information is abundant. One serious problem is that only a small portion of patients in the database has confirmatory diagnoses while many other patients remain undiagnosed because they did not comply with the recommended examinations. The phenomenon leads to a so-called positive-unlabelled situation and the labels are extremely imbalanced. In this paper, we propose a model-based approach to classify the unlabelled patients by using a Bayesian finite mixture model. We also discuss the label switching issue for the imbalanced data and propose a consensus Monte Carlo approach to address the imbalance issue and improve computational efficiency simultaneously. Simulation studies show that our proposed model-based approach outperforms existing positive-unlabelled learning algorithms. The proposed method is applied on the Cerner EHR for detecting diabetic retinopathy (DR) patients using laboratory measurements. With only 3% confirmatory diagnoses in the EHR database, we estimate the actual DR prevalence to be 25% which coincides with reported findings in the medical literature.

Keywords and phrases:

classification; multivariate *t* mixture; consensus Monte Carlo; diabetic retinopathy

1. Introduction.

The health care industry has become more computerized and digital in the past decade. The electronic health records (EHR), which are digitally archived data from hospitals and clinics, contain a tremendous amount of information on patients' medical history. The rich information in EHR is vital for research in personalized medicine and clinic decision support systems, which are potentially revolutionary to many traditional medical fields.

[*] ru.wang@okstate.edu . [†] ye.liang@okstate.edu .

While offering potential benefits to the clinical decision making process, EHR are highly subject to missing and inaccurate inputs (Hripcsak and Albers, 2012), which often leads to poor decisions. For instance, in many EHR-based predictive modelling, the diagnosis codes (e.g., the ICD-9 codes) are often used to determine the patient cohorts. Patients with certain codes of interest are identified as cases, and patients without such codes are typically assumed of no disease and thus identified as controls (Ng et al., 2016; Piri et al., 2017). However, such practices can be problematic with the assumption that missing diagnosis code equals to no disease. In fact, an underlying patient without the ICD-9 codes of interest may still have the disease but no examination was performed or the medical care was from a different health system. Ignorance of the uncertainty in the binary case-control labelling could severely dampen the performance of a predictive model (Zawistowski et al., 2017).

It is necessary to re-consider the data quality for any secondary analysis of EHR (Dziadkowiec et al., 2016; Botsis et al., 2010). In fact, the EHR data that we encounter are known as presence-only data in which a patient is either positively labelled or unlabelled. The problem of (re)labelling undiagnosed patients is referred to as the positive-unlabelled (PU) learning (Mordelet and Vert, 2014). The problem of PU learning arises from various situations, including disease-related gene identification (Yang et al., 2012; Mordelet and Vert, 2011), medical diagnosis (Zuluaga et al., 2011) and text classification (Li and Liu, 2003; Liu et al., 2003). Algorithm-based approaches are popular in the literature, for example, the two-step approach (Li and Liu, 2003; Liu et al., 2002; Li et al., 2009), biased support vector machine (SVM) (Liu et al., 2003) and ensemble algorithms based on bootstrap aggregating (Mordelet and Vert, 2014; Claesen et al., 2015).

The existing algorithm-based PU learning methods often perform poorly on heavy-tailed distributions (Xu, Crammer and Schuurmans, 2006) that are common for medical lab measurements. We also find that the existing algorithms can be sensitive to the positive proportion in the unlabelled group in simulation studies. Furthermore, parameter estimation and uncertainty quantification are in general unavailable for algorithm-based methods. There is a lack of model-based approaches for PU learning in the literature and a sophisticated statistical model can potentially address the aforementioned issues.

In this paper, we propose a Bayesian finite mixture model and use Markov Chain Monte Carlo (MCMC) for the Bayesian computation. It is noteworthy that although the finite mixture modelling has been extensively used in clustering and classification problems (Huang et al., 2005; Dean, Murphy and Downey, 2006; Martella et al., 2011; McNicholas, 2016), we are not aware of any direct application on the PU learning problem. There are two unique challenges when applying the Bayesian finite mixture model to our EHR data whose details are described in Section 6. First, the proposed finite mixture model should not suffer from the so-called label-switching issue with the presence of positively labelled observations. However, our real EHR data are extremely imbalanced with only 3% positively labelled cases, and we show that the label-switching occur and consequently estimates can be incorrect. Second, the Bayesian computation is intensive for the large EHR data so that an efficient computation is desired. In this paper, we propose to address the two challenges simultaneously by using the consensus Monte Carlo approach which splits data and then combines posterior subsamples.

Our contribution in this paper is twofold. First, on the methodological side, we develop a model-based approach for extremely imbalanced positive-unlabelled data, and show superior performance comparing with existing algorithm-based methods. Moreover, comparing with deterministic algorithms, the Bayesian finite fixture model offers better interpretability of the unlabelled records from a probabilistic perspective. The distributional assumptions and parameter specifications enable statistical inference for researchers and clinical physicians, allowing them to better understand the data generating process and population characteristics. Second, it is of practical significance to tackle the emerging problem of secondary analysis of EHR data. In the application of detecting patients with diabetic retinopathy (DR), with only 3% confirmatory diagnoses in the EHR database, we estimate the actual prevalence to be 25% which is consistent with reported findings from the medical community. In an external validation, we show that machine learning using classified results gain improvements. Finally, the proposed solution for PU-learning is generalizable for other types of applications that are beyond the scope of EHR data analysis.

The reminder of this paper is organized as follows. In section 2, we propose a Bayesian finite mixture model with multivariate t distributions conditional on categories. Section 3 is on the Bayesian computation, specifically the details of Gibbs sampling. In section 4, we illustrate the issue of label-switching for imbalanced data and propose a consensus Monte Carlo approach. We perform simulation studies in section 5 and show that our proposed approach outperforms existing approaches under various settings. The real Cerner EHR dataset for DR diagnosis and detection is extensively analyzed in section 6. Finally, section 7 contains concluding remarks for this paper.

2. Bayesian Finite Mixture Modelling.

2.1. Notations.

First we define useful notations. Let $\mathscr{P}$ denote the set of positively labelled observations and $\mathscr{U}$ denote the set of unlabelled observations, which is a mixture of positively and negatively labelled observations but the underlying label is unobserved. Let n_p and n_u denote the sample size of $\mathscr{P}$ and $\mathscr{U}$, respectively. Each observation in the data is a vector of measurements, which contains values of both categorical and numerical variables. Let $\boldsymbol{x}_i = \left(x_i^d, \boldsymbol{x}_i^c\right)$ be the feature vector, where x_i^d is a one-dimensional categorical variable and $\boldsymbol{x}_i^c$ is a vector of continuous variables. Note that multiple categorical variables can be collapsed into one categorical variable and thus it is sufficient that x_i^d is one-dimensional.

2.2. The finite mixture of multivariate t distributions.

Let our samples be $\mathscr{P} = \{\boldsymbol{x}_{p,i}\}$, $i = 1, \ldots, n_p$ and $\mathscr{U} = \{\boldsymbol{x}_{u,i}\}$, $i = 1, \ldots, n_u$. Assume that $\boldsymbol{x}_{p,i} \sim f_+(\boldsymbol{x})$ independently and $\boldsymbol{x}_{u,i} \sim f_u(\boldsymbol{x})$ independently, where

$$f_u(\boldsymbol{x}) = \pi f_+(\boldsymbol{x}) + (1 - \pi) f_-(\boldsymbol{x}),$$

where $f_-(\boldsymbol{x})$ is the underlying distribution of the negatively labelled samples, and the mixing probability $\boldsymbol{\pi}$ is called the contamination rate, which is the proportion of positive instances in the unlabelled set. To model $f_+(\boldsymbol{x})$ and $f_-(\boldsymbol{x})$, recall that $\boldsymbol{x}$ has two components, of which

x^d is categorical and $\boldsymbol{x}^c$ is continuous. It is natural that $f_+(\boldsymbol{x})$ is modelled conditionally, $f_+(\boldsymbol{x}) = f_+(\boldsymbol{x}^c \mid x^d = j)f_+(x^d = j)$, where $f_+(x^d = j)$, $j = 1, \ldots, J$ is a categorical distribution with category probabilities $\boldsymbol{p}_1$. Consider a multivariate t distribution for the conditional density $f_+(\boldsymbol{x}^c \mid x^d = j) = t(\boldsymbol{x}^c \mid \boldsymbol{\mu}_{1,j}, \Sigma_{1,j}, \nu_{1,j})$, where $\boldsymbol{\mu}_{1,j}$ is the location parameter, $\Sigma_{1,j}$ is the scale matrix and $\nu_{1,j}$ is the degrees of freedom, and the density function for an h-dimensional t is

$$t(\boldsymbol{x}^c \mid \boldsymbol{\mu}_{1,j}, \boldsymbol{\Sigma}_{1,j}, \nu_{1,j}) = \frac{\Gamma\left(\frac{h+\nu_{1,j}}{2}\right)\left|\boldsymbol{\Sigma}_{1,j}\right|^{-1/2}}{(\pi\nu_{1,j})^{h/2}\Gamma\left(\frac{\nu_{1,j}}{2}\right)\left[1+(\boldsymbol{x}^c-\boldsymbol{\mu}_{1,j})'\boldsymbol{\Sigma}_{1,j}^{-1}(\boldsymbol{x}^c-\boldsymbol{\mu}_{1,j})/\nu_{1,j}\right]^{\frac{h+\nu_{1,j}}{2}}}.$$

The degrees of freedom $\nu_{1,j}$ controls the thickness of distribution tails, and the multivariate t distribution approaches the multivariate normal distribution $N(\boldsymbol{x}^c \mid \boldsymbol{\mu}_{1,j}, \Sigma_{1,j})$ as $\nu_{1,j} \to \infty$. Similarly, let $f_-(\boldsymbol{x}) = f_-(\boldsymbol{x}^c \mid x^d = j)f_-(x^d = j)$, where $f_-(x^d = j)$ is a categorical distribution with parameters $\boldsymbol{p}_2$ and $f_-(\boldsymbol{x}^c \mid x^d = j) = t(\boldsymbol{x}^c \mid \boldsymbol{\mu}_{2,j}, \Sigma_{2,j}, \nu_{2,j})$.

In a mixture model, it is convenient to introduce a latent indicator $z_i \in \{1, 2\}$ for observations in $\mathscr{U}$, for which $z_i = 1$ denotes a positively labelled observation and $z_i = 2$ denotes a negatively labelled observation. Hence, $p(z_i = 1) = \boldsymbol{\pi}$ and $p(z_i = 2) = 1 - \boldsymbol{\pi}$. Given indicators $\boldsymbol{z} = \{z_i, i = 1, \ldots, n_u\}$, the likelihood function for the complete dataset $\{\mathscr{P}, \mathscr{U}\}$ can be written as

$$L(\{\boldsymbol{\mu}_{k,j}, \boldsymbol{\Sigma}_{k,j}, \nu_{k,j}, \boldsymbol{p}_k\}, \pi; \mathscr{P}, \mathscr{U}, \boldsymbol{z}) = \prod_{\boldsymbol{x}_{p,i} \in \mathscr{P}} f\left(x_{p,i}^d = j \mid \boldsymbol{p}_1\right)t\left(\boldsymbol{x}_{p,i}^c \mid \boldsymbol{\mu}_{1,j}, \boldsymbol{\Sigma}_{1,j}, \nu_{1,j}\right) \prod_{\boldsymbol{x}_{u,i} \in \mathscr{U}} p(z_i = k \mid \pi)f\left(x_{u,i}^d = j \mid \boldsymbol{p}_k\right)t\left(\boldsymbol{x}_{u,i}^c \mid \boldsymbol{\mu}_{k,j}, \boldsymbol{\Sigma}_{k,j}, \nu_{k,j}\right)$$

for $k = 1, 2$ and $j = 1, \ldots, J$. Note that a multivariate t distribution $\boldsymbol{X} \sim t(\boldsymbol{\mu}, \Sigma, \nu)$ can be constructed hierarchically by a multivariate normal distribution and a gamma distribution

$$\boldsymbol{X} \mid \tau \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/\tau), \quad \tau \sim \mathrm{Gamma}(\nu/2, \nu/2),$$

which is more convenient for statistical inference and computations. This model specification using multivariate t distributions has been considered in Andrews, McNicholas and Subedi (2011) for model-based classification problems, where the authors proposed an expectation-maximization algorithm for the inference. In this paper, we consider a Bayesian inference for the proposed model.

2.3. The prior distributions.

We now need to fully specify the Bayesian model by specifying the prior distributions for $\boldsymbol{\pi}$ and $\{\boldsymbol{\mu}_{k,j}, \Sigma_{k,j}, \nu_{k,j}, \boldsymbol{p}_k\}$, $k = 1, 2$, $j = 1, \ldots, J$. With the normal-gamma mixture representation for the multivariate t distribution, we can choose semi-conjugate priors for most parameters in the model. Let $\boldsymbol{\pi}$ have a beta prior $\mathrm{Beta}(\alpha_0, \beta_0)$ with $\alpha_0 = \beta_0 = 1$ if no extra information is known for the mixing probability. Let $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ both be $\mathrm{Dirichlet}(\lambda_1, \ldots, \lambda_J)$. Let $\boldsymbol{\mu}_{k,j} \sim \mathrm{N}(\boldsymbol{\mu}_0, \Sigma_0)$, $\Sigma_{k,j} \sim \text{Inv-Wishart}(\Phi_0, \xi_0)$ and $\nu_{k,j} \sim \mathrm{Gamma}(a_0, b_0)$ for any k and j. We assign $\lambda_1 = \cdots = \lambda_J = 1$, $\boldsymbol{\mu}_0 = \boldsymbol{0}$, $\Sigma_0 = 10^5\boldsymbol{I}$, $\Phi_0 = \boldsymbol{I}$ and $\xi_0 = 3$ to make these priors only weakly informative. In our experience, the degrees of freedom parameter $\nu_{k,j}$ sometimes can be difficult to identify, depending on the data in practice. Therefore we suggest to use an informative prior to avoid a potential non-identifiability situation. In our applications, we choose $a_0 = 5$ and $b_0 = 1$ whose mean is 5 and variance is 5. In the literature, a fixed value of $\nu_{k,j} = 4$ has been suggested to provide good protection against outliers (Lange, Little and Taylor, 1989; Stephens, 2000).

3. Bayesian Computations.

With priors specified in Section 2.3, we can use the Gibbs sampler to sample from the joint posterior distribution. Recall that n_p denotes the sample size of $\mathscr{P}$ and n_u denotes the sample size for $\mathscr{U}$. Define the following counts for convenience. Let $n_{p,1}, \ldots, n_{p,J}$ denote the counts for categories 1 to J in $\mathscr{P}$, i.e. $n_{p,j} = \sum_{\boldsymbol{x}_{p,i} \in \mathscr{P}} 1\left(x_{p,i}^d = j\right)$, where $1(\cdot)$ is an indicator function. Let $n_{u,p}$ and $n_{u,n}$ denote the counts of observations in $\mathscr{U}$ that are classified into the positive group and the negative group, respectively, i.e. $n_{u,p} = \sum_{\boldsymbol{x}_{u,i} \in \mathscr{U}} 1(z_i = 1)$ and $n_{u,n} = \sum_{\boldsymbol{x}_{u,i} \in \mathscr{U}} 1(z_i = 2)$. Dividing $n_{u,p}$ into J categories leads to the counts $n_{u,p,j}$, $j = 1, \ldots, J$, i.e. $n_{u,p,j} = \sum_{\boldsymbol{x}_{u,i} \in \mathscr{U}} 1\left(z_i = 1 \cap x_{u,i}^d = j\right)$. We define $n_{u,n,j}$, $j = 1, \ldots, J$ similarly. The full conditional distributions for Gibbs sampling are given as follows.

1. Sample the mixing probability $\boldsymbol{\pi}$ from

$$p(\pi \mid \text{others}) \sim \text{Beta}(\alpha_0 + n_{u,p}, \beta_0 + n_{u,n})$$

2. Sample the latent indicators z_i for $\boldsymbol{x}_{u,i} \in \mathscr{U}$ from Bernoulli with probability $p(z_i = 1 \mid \text{others})$ being

$$\frac{\pi f\left(x_{u,i}^d = j \mid \boldsymbol{p}_1\right) t\left(\boldsymbol{x}_{u,i}^c \mid \boldsymbol{\mu}_{1,j}, \boldsymbol{\Sigma}_{1,j}, \nu_{1,j}\right)}{\pi f\left(x_{u,i}^d = j \mid \boldsymbol{p}_1\right) t\left(\boldsymbol{x}_{u,i}^c \mid \boldsymbol{\mu}_{1,j}, \boldsymbol{\Sigma}_{1,j}, \nu_{1,j}\right) + (1 - \pi) f\left(x_{u,i}^d = j \mid \boldsymbol{p}_2\right) t\left(\boldsymbol{x}_{u,i}^c \mid \boldsymbol{\mu}_{2,j}, \boldsymbol{\Sigma}_{2,j}, \nu_{2,j}\right)}$$

   and $p(z_i = 2 \mid \text{others}) = 1 - p(z_i = 1 \mid \text{others})$

3. Sample $\boldsymbol{p}_1$ from Dirichlet$(\lambda_1 + n_{p,1} + n_{u,p,1}, \ldots, \lambda_J + n_{p,J} + n_{u,p,J})$.

4. Sample $\boldsymbol{p}_2$ from Dirichlet$(\lambda_1 + n_{u,n,1}, \ldots, \lambda_J + n_{u,n,J})$.

5. Sample $\boldsymbol{\mu}_{1,j}$ from $p(\boldsymbol{\mu}_{1,j} \mid \text{others}) \propto \mathbf{N}\left(\widetilde{\boldsymbol{\mu}}_{1,j}, \widetilde{\boldsymbol{\Sigma}}_{1,j}\right)$, where $\widetilde{\boldsymbol{\Sigma}}_{1,j}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_{1,j}^{-1} \sum_{\boldsymbol{x}_i \in \mathscr{D}_1} \tau_i$, $\widetilde{\boldsymbol{\mu}}_{1,j} = \widetilde{\boldsymbol{\Sigma}}_{1,j}\left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{1,j}^{-1} \sum_{\boldsymbol{x}_i \in \mathscr{D}_1} \tau_i \boldsymbol{x}_i^c\right)$, and define $\mathscr{D}_1 = \left\{\boldsymbol{x}_{p,i} : x_{p,i}^d = j\right\} \cup \left\{\boldsymbol{x}_{u,i} : z_i = 1 \cap x_{u,i}^d = j\right\}$.

6. Sample $\boldsymbol{\mu}_{2,j}$ from $p(\boldsymbol{\mu}_{2,j} \mid \text{others}) \propto N\left(\widetilde{\boldsymbol{\mu}}_{2,j}, \widetilde{\boldsymbol{\Sigma}}_{2,j}\right)$, where $\widetilde{\boldsymbol{\Sigma}}_{2,j}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Sigma}_{2,j}^{-1} \sum_{\boldsymbol{x}_i \in \mathscr{D}_2} \tau_i$, $\widetilde{\boldsymbol{\mu}}_{2,j} = \widetilde{\boldsymbol{\Sigma}}_{2,j}\left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{2,j}^{-1} \sum_{\boldsymbol{x}_i \in \mathscr{D}_2} \tau_i \boldsymbol{x}_i^c\right)$ and define $\mathscr{D}_2 = \left\{\boldsymbol{x}_{u,i} : z_i = 2 \cap x_{u,i}^d = j\right\}$.

7. Sample $\Sigma_{1,j}$ from $p(\Sigma_{1,j} \mid \text{others}) \propto$ Inv-Wishart$(\boldsymbol{A}_1 + \boldsymbol{\Phi}_0, n_{p,j} + n_{u,p,j} + \xi_0)$, where $\boldsymbol{A}_1 = \sum_{\boldsymbol{x}_i \in \mathscr{D}_1} \tau_i (\boldsymbol{x}_i^c - \boldsymbol{\mu}_{1,j})(\boldsymbol{x}_i^c - \boldsymbol{\mu}_{1,j})'$.

8. Sample $\Sigma_{2,j}$ from $p(\Sigma_{2,j} \mid \text{others}) \propto$ Inv-Wishart$(\boldsymbol{A}_2 + \boldsymbol{\Phi}_0, n_{u,n,j} + \xi_0)$, where $\boldsymbol{A}_2 = \sum_{\boldsymbol{x}_i \in \mathscr{D}_2} \tau_i (\boldsymbol{x}_i^c - \boldsymbol{\mu}_{2,j})(\boldsymbol{x}_i^c - \boldsymbol{\mu}_{2,j})'$.

9. Sample $\boldsymbol{\tau}_{p,i}$ for $\boldsymbol{x}_{p,i} \in \mathscr{P}$ from

$$p(\tau_{p,i} \mid \text{others}) \propto \text{Gamma}\left(\frac{\nu_{1,j} + \dim(\boldsymbol{x}^{C})}{2}, \frac{\nu_{1,j} + (\boldsymbol{x}_{p,i}^{c} - \boldsymbol{\mu}_{1,j})' \boldsymbol{\Sigma}_{1,j}^{-1} (\boldsymbol{x}_{p,i}^{c} - \boldsymbol{\mu}_{1,j})}{2}\right),$$

where $\dim(\boldsymbol{x}^{C})$ is the dimension of the continuous random vector $\boldsymbol{x}^{C}$.

10. Sample $\boldsymbol{\tau}_{u,i}$ for $\boldsymbol{x}_{u,i} \in \mathscr{U}$ from

$$p(\tau_{u,i} \mid \text{others}) \propto \text{Gamma}\left(\frac{\nu_{z_i,j} + \dim(\boldsymbol{x}^{C})}{2}, \frac{\nu_{z_i,j} + (\boldsymbol{x}_{u,i}^{c} - \boldsymbol{\mu}_{z_i,j})' \boldsymbol{\Sigma}_{z_i,j}^{-1} (\boldsymbol{x}_{u,i}^{c} - \boldsymbol{\mu}_{z_i,j})}{2}\right).$$

11. Sample $\nu_{k,j}$ from

$$p(\nu_{k,j} \mid \text{others}) \propto \nu_{k,j}^{a_0 - 1} e^{-b_0 \nu_{k,j}} \prod_{\boldsymbol{x}_i \in \mathscr{D}_k} \left\{ \frac{(\nu_{k,j}/2)^{\nu_{k,j}/2}}{\Gamma(\nu_{k,j}/2)} \tau_i^{\nu_{k,j}/2} e^{-\tau_i \nu_{k,j}/2} \right\}.$$

We use the Metropolis-Hastings algorithm to sample from a transformed version of this density $u_{k,j} = \log(\nu_{k,j})$ so that the support is $(-\infty, \infty)$. The proposal distribution is then chosen to be a normal distribution.

4. Label Switching for Imbalanced Data.

4.1. Illustration of label switching.

The label switching is a well-known issue for finite mixture models due to that components are unidentifiable in the likelihood. However, it need not be an issue when the likelihood contains a component for the positively labelled data, theoretically. Despite components are identifiable in the finite mixture model for positive-unlabelled data, the identifiability can be weak if the data are extremely imbalanced, i.e. data only contain a small portion of positively labelled cases. It can be illustrated that label switching occurs in the computation due to the weak identifiability and consequently leads to incorrect estimation and inference.

For the purpose of illustration, consider a finite mixture model as follows,

$$\boldsymbol{x}_{p,i} \sim f(x_{p,i}^{d} = j \mid \boldsymbol{p}_{+}) t(\boldsymbol{x}_{p,i}^{c} \mid \boldsymbol{\mu}_{+,j}, \boldsymbol{\Sigma}_{+,j}, \nu_{+,j}),$$

$$\boldsymbol{x}_{u,i} \sim \pi_{+} f(x_{u,i}^{d} = j \mid \boldsymbol{p}_{+}) t(\boldsymbol{x}_{u,i}^{c} \mid \boldsymbol{\mu}_{+,j}, \boldsymbol{\Sigma}_{+,j}, \nu_{+,j}) + \pi_{-} f(x_{u,i}^{d} = j \mid \boldsymbol{p}\_) t(\boldsymbol{x}_{u,i}^{c} \mid \boldsymbol{\mu}_{-,j}, \boldsymbol{\Sigma}_{-,j}, \nu_{-,j}),$$

for $\boldsymbol{x}_{p,i} \in \mathscr{P}$ and $\boldsymbol{x}_{u,i} \in \mathscr{U}$, where $\boldsymbol{\pi}_{-} = 1 - \boldsymbol{\pi}_{+}$. The log-likelihood function for this model can be written as follows,

$$
\begin{aligned}
l(\{\pi_{\pm}, \boldsymbol{\mu}_{\pm,j}, \boldsymbol{\Sigma}_{\pm,j}, \nu_{\pm,j}, \boldsymbol{p}_{\pm}\}; \mathscr{P}, \mathscr{U}) = {} & \sum_{j=1}^{J} \sum_{\substack{\boldsymbol{x}_{p,i} \in \mathscr{P} \\ x_{p,i}^{d} = j}} \log\left[p_{+,j} t\left(\boldsymbol{x}_{p,i}^{c} \mid \mu_{+,j}, \boldsymbol{\Sigma}_{+,j}, \nu_{+,j}\right)\right] \\
& + \sum_{j=1}^{J} \sum_{\substack{\boldsymbol{x}_{u,i} \in \mathscr{U} \\ x_{u,i}^{d} = j}} \log\left[\pi_{+} p_{+,j} t\left(\boldsymbol{x}_{u,i}^{c} \mid \boldsymbol{\mu}_{+,j}, \boldsymbol{\Sigma}_{+,j}, \nu_{+,j}\right) + \pi_{-} p_{-,j} t\left(\boldsymbol{x}_{u,i}^{c} \mid \boldsymbol{\mu}_{-,j}, \boldsymbol{\Sigma}_{-,j}, \nu_{-,j}\right)\right],
\end{aligned}
$$

If we switch the positive label and the negative label, the log-likelihood will differ as the first term changes. This log-likelihood function is then unimodal and hence components are identifiable as long as there is at least one positively labelled case. However, when the data are extremely imbalanced, that is, the sample size of $\mathscr{U}$ is much greater than that of $\mathscr{P}$, the first term in the log-likelihood will be negligible. In that case, the log-likelihood will be practically multimodal, although the global maximum is unique. In this mixture model, each labelled group is further a mixture of J categories, and each category j is prone to labeling switching as well. Therefore there are practically 2^J modes with close likelihood values in the scenario of imbalanced labels.

To show the non-convergence of Markov chains under the multimodal scenario, we simulate two synthetic data sets. For both data sets, we assume $n_p = 100$, $\boldsymbol{\pi}_+ = 0.3$, $\boldsymbol{\pi}_- = 0.7$, $p_{+,1} = p_{+,2} = p_{-,1} = p_{-,2} = 0.5$, $\boldsymbol{\mu}_{+,1} = (1, 1, 1)'$, $\boldsymbol{\mu}_{+,2} = (2, 2, 2)'$, $\boldsymbol{\mu}_{-,1} = (3, 3, 3)'$, $\boldsymbol{\mu}_{-,2} = (4, 4, 4)'$, $\Sigma_{+,1} = \Sigma_{+,2} = \Sigma_{-,1} = \Sigma_{-,2} = \boldsymbol{I}$, $\nu_{+,1} = \nu_{+,2} = 2$ and $\nu_{-,1} = \nu_{-,2} = 5$. Then we assign $n_u = 200$ for data set 1 and $n_u = 4,000$ for data set 2 to mimic the balanced scenario and the imbalanced scenario. Let $\{\boldsymbol{\pi}_+, \boldsymbol{\mu}_{+,1}, \boldsymbol{\mu}_{+,2}, \boldsymbol{\mu}_{-,1}, \boldsymbol{\mu}_{-,2}, \nu_{+,1}, \nu_{+,2}, \nu_{-,1}, \nu_{-,2}, \boldsymbol{p}_{\pm}, \boldsymbol{\Sigma}_{\pm,j}\}$ be the parameters of interest. It is obvious that the log-likelihood has the following $2^2 = 4$ local maximums.

$$\{0.3, (1, 1, 1)', (2, 2, 2)', (3, 3, 3)', (4, 4, 4)', 2, 2, 5, 5, 0.5, \boldsymbol{I}\},$$

$$\{0.7, (3, 3, 3)', (4, 4, 4)', (1, 1, 1)', (2, 2, 2)', 5, 5, 2, 2, 0.5, \boldsymbol{I}\},$$

$$\{0.5, (3, 3, 3)', (2, 2, 2)', (1, 1, 1)', (4, 4, 4)', 5, 2, 2, 5, 0.5, \boldsymbol{I}\},$$

$$\{0.5, (1, 1, 1)', (4, 4, 4)', (3, 3, 3)', (2, 2, 2)', 2, 5, 5, 2, 0.5, \boldsymbol{I}\},$$

for which only the first one is the global maximum. For the sake of convenience, refer them as Modes 1 to 4, respectively. For the relatively balanced data set 1, we expect that Mode 1 be easily identified as the global maximum and label switching does not occur. On the other hand, for the extremely imbalanced data set 2, we expect that all four modes have close log-likelihood values and label switching occurs. Figure 1 subplots (a) and (b) overlap the log-likelihood as a function of $\boldsymbol{\pi}_+$ with other parameters fixed at the four modes, respectively. Subplot (a) is for data set 1. It is clear that $\boldsymbol{\pi}_+$ takes values 0.3, 0.5, 0.5, 0.7 at the four modes and Mode 1 with $\boldsymbol{\pi}_{+} = 0.3$ is the global maximum. Subplot (b) is for data set 2. Mode 1 with $\boldsymbol{\pi}_{+} = 0.3$ is still the global maximum but it is hardly distinguishable from other modes, indicating a weak identifiability. Figure 1 subplots (c) and (d) show MCMC trace plots for $\boldsymbol{\pi}_{+}$ for data set 1 and 2, respectively. We purposely choose dispersed initial values for four Markov chains. In subplot (c), all chains converge to the correct mode of $\boldsymbol{\pi}_{+}$. In subplot (d), the chains do not converge to the correct mode but stuck with local modes depending on the initial values. Thus, analysts should be careful about the label switching problem when their data appear to be imbalanced.

4.2. Consensus Monte Carlo.

It is obvious that the weak identifiability is not because of the model but is due to the imbalanced dataset. An immediate solution is to make the data relatively balanced. Let us equally split the unlabelled group $\mathscr{U}$ into $\mathscr{S}$ subgroups $\{\mathscr{U}^{(1)}, \cdots, \mathscr{U}^{(S)}\}$ such that the sample size of a subgroup $\mathscr{U}^{(s)}$ is comparable to that of $\mathscr{P}$. Consider S parallel sub-datasets: $\{\mathscr{P}, \mathscr{U}^{(1)}\}, \cdots, \{\mathscr{P}, \mathscr{U}^{(S)}\}$, and each of them is relatively balanced. We apply the finite mixture model and Bayesian computations on each sub-dataset, and write the sub-likelihood function as

$$L^{(s)}\left(\theta_1^{(s)}, \theta_2^{(s)}, \pi^{(s)}; \mathscr{P}, \mathscr{U}^{(s)}\right) = \prod_{\boldsymbol{x}_i \in \mathscr{P}} f_{+}\left(\boldsymbol{x}_i; \theta_1^{(s)}\right) \prod_{\boldsymbol{x}_i \in \mathscr{U}^{(s)}} \left[\pi^{(s)} f_{+}\left(\boldsymbol{x}_i; \theta_1^{(s)}\right) + \left(1 - \pi^{(s)}\right) f_{-}\left(\boldsymbol{x}_i; \theta_2^{(s)}\right)\right].$$

To combine posterior samples obtained from the sub-likelihood functions, we utilize the consensus Monte Carlo (Scott et al., 2016), which was introduced for parallelizing MCMC. A more efficient computation is a by-product but also desired since EHR data are typically large. Note that our implementation is slightly different from the original consensus Monte Carlo in that $\mathscr{P}$ is repeatedly used instead of being split. The proposed split-and-combine procedure is outlined as follows.

1. Split $\mathscr{U}$ into S subgroups $\{\mathscr{U}^{(1)}, \cdots, \mathscr{U}^{(S)}\}$ with equal size $n_u^{(s)}$.

2. For $s = 1, \cdots, S$, in parallel, apply the Bayesian finite mixture model. The MCMC gives posterior samples from

$$p\left(\boldsymbol{\Theta} \mid \mathscr{P}, \mathscr{U}^{(s)}\right) \propto p(\boldsymbol{\Theta}) p\left(\mathscr{P}, \mathscr{U}^{(s)} \mid \boldsymbol{\Theta}\right),$$

   where $\boldsymbol{\Theta} = \{\{\boldsymbol{\mu}_{k,j}, \boldsymbol{\Sigma}_{k,j}, \nu_{k,j}, \boldsymbol{p}_k\}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\tau}\}$, $k = 1, 2$, $j = 1, \ldots, J$.

3. Suppose for a parameter θ, we obtain sub-posterior draws from the S subgroups, $\theta_t^{(1)}, \ldots, \theta_t^{(S)}$ at iteration t. To combine them to an aggregated posterior draw θ_t, according to Scott et al. (2016), compute a weighted average

$$\theta_t = \left(\sum_{s=1}^{S} \omega^{(s)}\right)^{-1} \sum_{s=1}^{S} \omega^{(s)} \theta_t^{(s)},$$

   where $\omega^{(s)}$ is the weight for subgroup s.

   a. For parameters $\{\boldsymbol{\mu}_{k,j}, \Sigma_{k,j}, \nu_{k,j}\}$, choose $\omega^{(s)} = \mathrm{Var}(\theta \mid \mathscr{P}, \mathscr{U}^{(s)})^{-1}$, where $\mathrm{Var}(\theta \mid \mathscr{P}, \mathscr{U}^{(s)})$ is estimated by the sample variance of $\{\theta_1^{(s)}, \cdots, \theta_T^{(s)}\}$.

   b. For parameters $\{\boldsymbol{\pi}, \boldsymbol{p}_k\}$, which are [0, 1]-bounded, the arithmetic average tends to perform better in practice, that is to choose $\omega^{(s)} =$ 1.

4. After combining posterior samples, compute the posterior classification probabilities $p(z_i = 1 \mid \mathscr{P}, \mathscr{U})$, for $\mathbf{x}_{u,i} \in \mathscr{U}$ with $x_{u,i}^d = j$, using the following expression

$$\frac{1}{T-B} \sum_{t=B+1}^{T} \frac{\pi^{(t)} f_+\left(\boldsymbol{x}_{u,i} \mid \boldsymbol{p}_1^{(t)}, \boldsymbol{\mu}_{1,j}^{(t)}, \boldsymbol{\Sigma}_{1,j}^{(t)}, \nu_{1,j}^{(t)}\right)}{\pi^{(t)} f_+\left(\boldsymbol{x}_{u,i} \mid \boldsymbol{p}_1^{(t)}, \boldsymbol{\mu}_{1,j}^{(t)}, \boldsymbol{\Sigma}_{1,j}^{(t)}, \nu_{1,j}^{(t)}\right) + \left(1-\pi^{(t)}\right) f_-\left(\boldsymbol{x}_{u,i} \mid \boldsymbol{p}_2^{(t)}, \boldsymbol{\mu}_{2,j}^{(t)}, \boldsymbol{\Sigma}_{2,j}^{(t)}, \nu_{2,j}^{(t)}\right)},$$

   where B is the burn-in sample size.

5. Classify an unlabelled individual i as a positive case if $p(z_i = 1 \mid \mathscr{P}, \mathscr{U}) \geq 0.5$, otherwise as a negative case.

A related alternative to address the imbalance issue is oversampling the minority group. Note that our proposed procedure is equivalent to using the following full likelihood function as we repeatedly use $\mathscr{P}$ for S times,

$$L(\theta_1, \theta_2, \pi; \mathscr{P}, \mathscr{U}) = \prod_{\boldsymbol{x}_i \in \mathscr{P}} [f_+(\boldsymbol{x}_i; \theta_1)]^S \prod_{\boldsymbol{x}_i \in \mathscr{U}} [\pi f_+(\boldsymbol{x}_i; \theta_1) + (1-\pi) f_-(\boldsymbol{x}_i; \theta_2)],$$

which is indeed similar as oversampling $\mathscr{P}$ with S multiples. However, directly using an oversampled large dataset is computationally challenging, and thus a parallel computation has its advantages. It is theoretically difficult to justify an optimal choice of the splitting size S. The bottom line is that the sample sizes of $\mathscr{U}^{(s)}$ and $\mathscr{P}$ are comparable so that label-switching is avoided.

5. Simulation Studies.

5.1. Comparing performance with algorithm-based methods.

In this section, we conduct a simulation study to compare our proposed mixture modelling approach with the state-of-the-art PU learning techniques. To mimic the real EHR data described in Section 6, we now let $\mathscr{P}$ be the group of DR patients and $\mathscr{N}$ be the group of non-DR patients. We let x_i^d be a categorical variable gender ($x_i^d = 1$ denotes male and $x_i^d = 2$ denotes female) and let $\boldsymbol{x}_i^c$ be a vector of three continuous laboratory variables. We consider three distribution settings for the data generation as follows.

1. We let $\boldsymbol{x}_i^c$ for the male DR patients be generated from a multivariate t distribution with $\boldsymbol{\mu}_{1,1} = (1, 1, 1)'$, $\Sigma_{1,1} = \boldsymbol{I}$ and $\nu_{1,1} = 2$, and let $\boldsymbol{x}_i^c$ for the female DR patients be from a multivariate t distribution with $\boldsymbol{\mu}_{1,2} = (2, 2, 2)'$, $\Sigma_{1,2} = \boldsymbol{I}$ and $\nu_{1,2} = 2$. For the non-DR group, we let $\boldsymbol{x}_i^c$ for the male non-DR patients be from a multivariate t distribution with $\boldsymbol{\mu}_{2,1} = (3, 3, 3)'$, $\Sigma_{2,1} = \boldsymbol{I}$ and $\nu_{2,1} = 5$, and let $\boldsymbol{x}_i^c$

for the female non-DR patients be from a multivariate t distribution with $\boldsymbol{\mu}_{2,2} = (4, 4, 4)'$, $\Sigma_{2,2} = \boldsymbol{I}$ and $\nu_{2,2} = 5$. Let the category probabilities be $p_{11} = p_{12} = p_{21} = p_{22} = 0.5$.

2. We consider a skewed distribution model. We let each variable of $\boldsymbol{x}_i^c$ for the male DR patients be from Gamma(2, 1) and for the female DR patients be from Gamma(5, 2). We then shift both gamma distributions to the right by 2 units for the non-DR patients. Let the category probabilities be $p_{11} = p_{12} = p_{21} = p_{22} = 0.5$.

3. We consider Cauchy distributions for which outliers can be extreme. We let the male and female DR groups be Cauchy(1, 0.5) and Cauchy(2, 0.5) for $\boldsymbol{x}_i^c$, respectively, where the first parameter is the location and the second parameter is the scale. We choose Cauchy(3, 0.5) and Cauchy(4, 0.5) for the male and female non-DR groups, respectively. Let the category probabilities be $p_{11} = p_{12} = p_{21} = p_{22} = 0.5$.

Note that in the above three cases, Cases 2 and 3 present a situation where the multivariate t mixture model is misspecified, and hence the robustness of each method is examined under skewed and heavy-tailed scenarios. For all three cases, we fix the sample size for the labelled DR group as $n_p = 100$ and that for the unlabelled group as $n_u = 4{,}000$ to represent imbalanced data in real applications. For each of the three cases, we randomly mix a portion of DR patients with non-DR patients with a mixing probability $\boldsymbol{\pi}$ ranging from 0.3 to 0.7. We randomly generate 30 datesets under each setting and implement the following four methods for each dataset: the two-step method, the biased-SVM, the bagging-SVM and our proposed Bayesian finite mixture model. For the two-step method, we adopt the version in Liu et al. (2003) for which a Naïve Bayesian classifier is built in Step 1 and the SVM is used in Step 2. For the biased-SVM described in Liu et al. (2003), we let $C_+ n_p = C_- n_u$, where C_+ and C_- are the penalties of misclassifying a positive and a negative, respectively. For the bagging-SVM described in Mordelet and Vert (2014), we set the bootstrap sample size $K = n_p$ and the number of bootstrap samples $T = 10$. For both the biased-SVM and the bagging-SVM, we use the grid search and cross-validation to select the optimal cost parameter in a defined set $\{10^{-12}, 10^{-11}, \ldots, 10^1, 10^2\}$. For our proposed mixture modelling approach, we set the splitting size $S = 20$. We adopt the Box-Cox power transformations for skewed samples. Lo and Gottardo (2012) argue that the t-mixture model with Box-Cox transformation performs favorably in terms of accuracy and robustness compared with the skewed-t approach (Azzalini and Capitanio, 2003).

To evaluate the performance of each method, we adopt the conventionally used metrics in the PU learning literature. Let TP denote the number of true positives, TN denote the number of true negatives, FP denote the number of false positives and FN denote the number of false negatives. Define

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN} + \text{FP} + \text{FN}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP+FP}},$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}, \quad \text{Specificity} = \frac{\text{TN}}{\text{TN+FP}},$$

and also define the F-score as follows,

$$F = \frac{2 \times \text{Precision} \times \text{Sensitivity}}{\text{Precision+Sensitivity}}.$$

The accuracy and the F-score are the most popular metrics for comparing PU learning methods in the literature.

The results are shown in Tables 1, 2 and 3 for the three distribution settings. A plot of accuracy versus π is shown in Figure 2 and a plot of F-score versus π is shown in Figure 3. As we can see from the tables and plots, in terms of both accuracy and F-score, the three algorithm-based methods appear to be sensitive to heavy-tailed distributions. We also notice that the performance of the three algorithm-based methods can be compromised by an increased mixing probability π. It should be pointed out that F-score does not exist for the two-step method because the two-step algorithm classifies all positive cases in the unlabelled group as negative. In all settings with various π values, our proposed mixture modelling approach seems to be quite stable and outperforms the other methods in general.

5.2. Parameter estimation in the mixture model.

One advantage of using a model-based method is that parameter estimation and statistical inference are possible, which are of interest for medical practitioners as the interpretability is an important issue despite the goal of classification. Suppose that in our simulation scenarios, besides the mixing probability π, a medical practitioner would also like to know the proportion of male patients in the DR group p_{11} and that in the non-DR group p_{21}. The practitioner also wants to estimate the difference between the DR and non-DR groups in the lab variables for both the male and female populations, that is, to estimate $\mu_{21l} - \mu_{11l}$ and $\mu_{22l} - \mu_{12l}$, for $l = 1, 2, 3$. Table 4 assesses the posterior mean estimates for those parameters of interest under the correct t-distribution model. Both the bias and the mean squared error (MSE) seem to be reasonably small for all parameters. Note that we cannot offer comparisons with the algorithm-based methods since they simply do not produce such parameter estimates.

6. Classifying Unspecified Diagnosis Codes in EHR.

6.1. Data description and preliminary analyses.

Diabetic retinopathy is a vision-threatening microvascular complication of diabetes and a leading cause of blindness among working-aged adults globally (Kobrin and Barbara, 2007; Yau et al., 2012). While diabetes is a common chronic disease worldwide, almost all patients with type 1 diabetes and more than 60% of patients with type 2 diabetes develop retinopathy during their first twenty years of diabetes (Fong et al., 2004). Retinopathy is often asymptomatic in early stages and the vision loss can only be prevented but not restored. Therefore it is imperative to detect and treat patients in their early stages.

However, early detection and screening of DR face challenges from poor adherence to annual examination guideline and lack of resources to deploy comprehensive screening programs (Ting, Cheung and Wong, 2016; Ciulla, Amador and Zinman, 2003). It is crucial to develop cost-effective early detection techniques for medical communities, especially for those rural communities. It is conjectured that routine laboratory measurements can be useful in detecting DR without needing comprehensive eye exams. With abundant laboratory information in EHR data, analysts and researchers have been developing predictive models or algorithms for the DR research in recent years, for instance, Piri et al. (2017), Saleh et al. (2016), Sun and Zhang (2019) and Skevofilakas et al. (2010).

The Cerner Health Facts EHR database (Cerner Corporation, Kansas City, MO) is one of the largest de-identified and HIPAA complied databases in the United States. The database contains time-stamped encounter, diagnosis, procedure, medication, and laboratory test data contributed voluntarily from hospitals running a Cerner EHR system across the U.S. from 2000 to 2018. In the step of data pre-processing, the ICD-9 diagnosis codes were used to specify the study cohort. Patients with one or more 250.xx ICD-9 codes were defined as diabetic patients. Within the set of diabetic patients, those who have 362.0x ICD-9 codes were identified as DR patients (case group), and those without any 362.0x code were considered non-DR patients (control group). Note that labelling patients without any 362.0x code as non-DR is questionable since these codes are in fact unspecified. It is, however, a common practice when pre-processing EHR data for secondary analyses.

Based on the existing literature on clinical predictions of DR, we initially included 26 laboratory variables and 3 demographic variables. For the labelled DR patients, we extracted feature values at their first DR diagnosis encounter, and for the labelled non-DR patients, we extracted feature values from their most recent hospital visit. In total 1, 207 labelled DR patients and 39, 767 labelled non-DR patients were finally included in the study cohort. Figure 4 shows the data extraction process. The whole dataset was then divided into a training set (70%) and a testing set (30%). To handle the imbalanced dataset at this initial stage, we implemented the synthetic minority over-sampling technique (SMOTE) which comprises over-sampling the synthetic minority class and under-sampling the majority class (Chawla et al., 2002). We then applied the random forest (Breiman, 2001) on the training dataset and used a ten-fold cross-validation to select the optimal model in terms of the area under the receiver operating characteristic curve (ROC AUC). The model was evaluated on the testing set and the AUC was 0.8709. The model performance was satisfactory on the testing set but in a later section we will see that the prediction can be disappointing in a completely external validating.

6.2. Applying the proposed PU learning method.

It is a questionable practice to treat patients without any 362.0x code as non-DR patients. As discussed in Section 1, it often happens when a medical examination is not performed on the underlying patient or the record of examination is missing. The apparent consequence is that any secondary analysis based on these data may be misleading as the key label is unreliable. In this section, we consider such a problem as a PU learning task and by applying the proposed Bayesian mixture model, we are able to re-label patients with unspecified codes.

Under the PU learning framework, the case group is denoted by $\mathscr{P}$ and the control group is denoted by $\mathscr{U}$.

The preliminary analysis using machine learning algorithms helped us to identify a few important predictors. We also carefully surveyed the relevant medical literature and determined three laboratory variables that would be used in the PU learning: HbA1c, hemoglobin and BUN. There is also a general interest to compare male patients with female patients so that the categorical variable gender is also included. It is noteworthy that distributions of laboratory measurements are often skewed and heavy-tailed (MIT-Critical-Data, 2016). Therefore, the proposed multivariate t-mixture after appropriate Box-Cox transformations is suitable for such data. The prior distributions are chosen based on Section 2.3. We divided $\mathscr{U}$ into $S = 20$ subgroups. For each subgroup, we ran three Markov chains with distinct initial values and each chain had 10, 000 iterations with burn-in size 4, 000. The convergence for the key parameter $\boldsymbol{\pi}$ is shown in Figure 5 for an illustration. The Gelman-Rubin diagnostic (Gelman and Rubin, 1992) is performed on all continuous parameters $\{\boldsymbol{\pi}, \boldsymbol{\mu}_{k,j}, \Sigma_{k,j}, \nu_{k,j}, \boldsymbol{p}_k\}$ and the worst-case potential scale reduction factor is 1.02 (97.5% upper bound 1.06), well below the commonly used threshold of 1.1.

From the posterior samples, we obtained the posterior mean and standard deviation for the mixing probability $\boldsymbol{\pi}$ as 0.23 and 0.0043, respectively, indicating roughly one fourth of unspecified cases may be classified into the DR patients group based on the extracted laboratory measurements. From the posterior mean of $\boldsymbol{\pi}$, we can estimate the DR prevalence rate for the Cerner Health Facts EHR database, which is $\frac{n_p + \pi n_u}{n_p + n_u} = 0.25$. This estimate is in line with the reported incidence of DR among US adults with diabetes, which is 28.5% (95% CI : 24.9% – 32.5%) (Zhang et al., 2010).

For inferential interest in parameters, the posterior estimates for $\boldsymbol{p}_1$, $\boldsymbol{p}_2$ and $\boldsymbol{\mu}_{k,j}$ are given in Table 5. Notice that while some categories are not distinguishable (HbA1c DR female vs. non-DR female), some others are clearly distinct (Hemoglobin DR male vs. non-DR male). A patient was classified into the DR group if the posterior probability $p(z_i = 1 \mid \text{Data}) \geq 0.5$. Eventually, among 39, 767 unspecified diabetic patients, 7, 010 were classified into the DR patients group. Figure 6 shows, for all three continuous laboratory variables, the distributions of the positive group $\mathscr{P}$, the unlabelled group $\mathscr{U}$, re-labelled positives from the unlabelled group $\mathscr{U}_p$ and re-labelled negatives from the unlabelled group $\mathscr{U}_n$.

To assess the classification accuracy in the environment of real data analysis, we conduct a further simulation study. We acknowledge that the misclassification rate cannot be properly evaluated due to that true labels are unknown in the unlabelled group. Instead, a simulation that mimics the real data situation is used. Consider now all 1207 positive cases are known positives, and we resample 1207 negative cases from patients who are labelled “negative” after mixture model analysis. These samples are treated as known positives and negatives in the simulation setting. Then 30% randomly selected positive cases are masked and mixed with all negative cases, called unlabelled patients. We fit the Bayesian finite mixture model using the synthetic dataset and record classification accuracy and the area under the receiver operating characteristics curve (AUC). The simulation is repeated for 30 times. The mean accuracy is 0.8982 with standard deviation 0.0115, and the mean AUC is 0.8996 with standard deviation 0.0092.

6.3. Sensitivity Analysis.

We conduct a sensitivity analysis for the mixing probability π with both a formal approach and an informal approach. We adopt a formal sensitivity analysis in Roos et al. (2015) for the Bayesian ϵ-local sensitivity. Given a scalar parameter θ, we denote the base prior and the corresponding marginal posterior by $p_{\gamma 0}(\theta)$ and $p_{\gamma 0}(\theta \mid y)$, where γ_0 denotes fixed hyperparameters in the base prior. The ϵ-local circular sensitivity $S^c_{\gamma_0}(\epsilon)$ is defined as

$$S^c_{\gamma_0}(\epsilon) = \left\{ \frac{d\left(p_\gamma(\theta \mid y), p_{\gamma_0}(\theta \mid y)\right)}{\epsilon}, \text{ for } \gamma \in G_{\gamma_0}(\epsilon) \right\},$$

where the grid $G_{\gamma 0}(\epsilon)$ is a contour line around γ_0 and $G_{\gamma 0}(\epsilon) = \{\gamma : d(p_\gamma(\theta), p_{\gamma 0}(\theta)) = \epsilon\}$. In Roos et al. (2015), the distance $d(\cdot, \cdot)$ is recommended to be the Hellinger distance, and $\epsilon = 0.00354$ as it is calibrated to measure the distance between $N(0, 1)$ and $N(0.01, 1)$, a local perturbation. In our case, the base prior chosen for the mixing probability is a beta distribution with $\gamma_0 = (a_0, \beta_0) = (1, 1)$. Such a contour line $G_{\gamma 0}(\epsilon)$ is depicted in Figure 7 (a). It is recommended to check the worst-case sensitivity $\max\{S^c_{\gamma_0}(\epsilon)\}$ in the ϵ-local grid. A sensitivity value larger than 1 leads to the scenario of super-sensitivity as the marginal posterior changes more than the prior change. Figure 7 (a) shows that our worst-case sensitivity is 0.021, indicating that the marginal posterior is robust against a prior change.

An informal sensitivity analysis is then performed by modifying prior hyperparameters in an ad hoc way. The following five scenarios are considered: (1) base-case with no modification; (2) Beta(0.5, 0.5) for π; (3) Beta(5, 12) for π, which centers at about 0.3 with 95% interval (0.1, 0.5); (4) Beta(5, 12) for π and Gamma(3, 1) for the degrees of freedom $\nu_{k,j}$; (5) Beta(5, 12) for π and Gamma(10, 1) for $\nu_{k,j}$. For all scenarios, models are re-fitted and resulting marginal posterior distributions for π against each prior choice are shown in Figure 7 (b). The marginal posterior distribution is in general robust against prior modifications.

6.4. Secondary analyses and external validation.

After we re-labelled the EHR dataset based on the PU learning, an analyst now can use the “new” dataset for a secondary analysis. In this section, we show that a machine learning algorithm using the “new” re-labelled data will enjoy a great improvement in terms of predicting a completely external dataset, comparing with using the original dataset.

Suppose now an analyst wants to build a DR predictive model using our Cerner EHR data. We provide him/her two datasets: the original dataset where unspecified patients are labelled non-DR and the re-labelled dataset where unspecified patients are classified based on our PU learning results. The analyst has an external EHR dataset from the University of Kansas Medical Center, which contains 1, 060 confirmed DR patients and can be used as a validation dataset. The validation dataset includes 19 laboratory variables so the analyst builds a random forest model using 19 laboratory variables and 3 demographic variables from the Cerner EHR data. The SMOTE technique has been used for both the original EHR dataset and the re-labelled EHR dataset before building the random forest model. The analyst then uses the fitted model to predict DR cases from the 1, 060 confirmed cases in the validation dataset. Using the original EHR dataset, the analyst is able to detect only 338 or 32% DR patients from the validation dataset, while using the re-labelled EHR dataset, he/she is able to detect 587 or 55% DR patients, which is a remarkable improvement. It is noteworthy that in this validation, the specificity cannot be computed because the number of true negatives is unknown.

7. Concluding Remarks.

In this paper, we proposed to use a Bayesian finite mixture model with multivariate t distributions to solve a PU learning problem that arises from labelling diagnosis codes in EHR data. A split-and-combine strategy known as consensus Monte Carlo is used to address the issue of imbalanced data and improve computational efficiency. We demonstrated its performance in the simulation study compared with existing algorithm-based PU learning methods. We applied the proposed approach to a real application where diabetic patients' DR codes need to be re-labelled and showed an improved predictive performance in an external validation. It has been warned in the literature that the quality of EHR data is often compromised due to missing or erroneous inputs. In this paper, we showed that a large EHR database only contains about 3% diagnosed DR cases, which is inconsistent with the medical consensus. We estimate about one fourth of diabetic patients potentially labelled as DR patients, in line with the literature. The proposed method can be used as a pre-processing tool for secondary analyses on positive unlabelled EHR data.

Acknowledgments.

The authors gratefully acknowledge the constructive and insightful comments and suggestions made by an editor and two anonymous reviewers on an earlier version of this manuscript. This work was conducted with data from the Cerner Corporation's Health Facts database of electronic medical records provided by the Oklahoma State University Center for Health Systems Innovation (CHSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Cerner Corporation.

Funding.

Research reported in this publication was partially supported by the Oklahoma Center for the Advancement of Science & Technology under Grant Number HR 18-087 and the National Eye Institute of the National Institutes of Health under Award Number R01EY033861. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Andrews JL, Mcnicholas PD and Subedi S (2011). Model-based classification via mixtures of multivariate t-distributions. Computational Statistics & Data Analysis 55 520–529.

Azzalini A and Capitanio A (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 367–389.

Botsis T, Hartvigsen G, Chen F and Weng C (2010). Secondary use of EHR: data quality issues and informatics opportunities. Summit on Translational Bioinformatics 2010 1.

Breiman L (2001). Random forests. Machine Learning 45 5–32.

Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16 321–357.

Ciulla TA, Amador AG and Zinman B (2003). Diabetic retinopathy and diabetic macular edema: pathophysiology, screening, and novel therapies. Diabetes Care 26 2653–2664. [PubMed: 12941734]

Claesen M, De Smet F, Suykens JA and De Moor B (2015). A robust ensemble approach to learn from positive and unlabeled data using SVM base models. Neurocomputing 160 73–84.

Dean N, Murphy TB and Downey G (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. Journal of the Royal Statistical Society: Series C (Applied Statistics) 55 1–14.

Dziadkowiec O, Callahan T, Ozkaynak M, Reeder B and Welton J (2016). Using a data quality framework to clean data extracted from the electronic health record: A case study. eGEMs 4.

Fong DS, Aiello L, Gardner TW, King GL, Blankenship G, Cavallerano JD, Ferris FL and Klein R (2004). Retinopathy in diabetes. Diabetes Care 27 s84–s87. [PubMed: 14693935]

Gelman A and Rubin DB (1992). Inference from iterative simulation using multiple sequences. Statistical science 7 457–472.

Hripcsak G and Albers DJ (2012). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association 20 117–121. [PubMed: 22955496]

Huang Y, Englehart KB, Hudgins B and Chan AD (2005). A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. IEEE Transactions on Biomedical Engineering 52 1801–1811. [PubMed: 16285383]

Kobrin K and Barbara E (2007). Overview of epidemiologic studies of diabetic retinopathy. Ophthalmic Epidemiology 14 179–183. [PubMed: 17896294]

Lange KL, Little RJ and Taylor JM (1989). Robust statistical modeling using the *t* distribution. Journal of the American Statistical Association 84 881–896.

Li X and Liu B (2003). Learning to classify texts using positive and unlabeled data. In IJCAI 3 587–592.

Li X-L, Yu PS, Liu B and Ng S-K (2009). Positive unlabeled learning for data stream classification. In Proceedings of the 2009 SIAM International Conference on Data Mining 259–270. SIAM.

Liu B, Lee WS, Yu PS and Li X (2002). Partially supervised classification of text documents. In ICML 2 387–394. Citeseer.

Liu B, Dai Y, Li X, Lee WS and Philip SY (2003). Building Text Classifiers Using Positive and Unlabeled Examples. In ICDM 3 179–188. Citeseer.

Lo K and Gottardo R (2012). Flexible mixture modeling via the multivariate *t* distribution with the Box-Cox transformation: an alternative to the skew-*t* distribution. Statistics and Computing 22 33–52. [PubMed: 22125375]

Martella F, Vermunt J, Beekman M, Westendorp R, Slagboom P and Houwing-Duistermaat J (2011). A mixture model with random-effects components for classifying sibling pairs. Statistics in Medicine 30 3252–3264. [PubMed: 21905068]

McNicholas PD (2016). Mixture model-based classification. Chapman and Hall/CRC.

MIT-Critical-Data (2016). Secondary analysis of electronic health records. Springer International Publishing.

Mordelet F and Vert J-P (2011). Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics 12 389. [PubMed: 21977986]

Mordelet F and Vert J-P (2014). A bagging SVM to learn from positive and unlabeled examples. Pattern Recognition Letters 37 201–209.

Ng K, Steinhubl SR, DeFilippi C, Dey S and Stewart WF (2016). Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circulation: Cardiovascular Quality and Outcomes 9 649–658. [PubMed: 28263940]

Piri S, Delen D, Liu T and Zolbanin HM (2017). A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble. Decision Support Systems 101 12–27.

Roos M, Martins TG, Held L and Rue H (2015). Sensitivity analysis for Bayesian hierarchical models. Bayesian Analysis 10 321–349.

Saleh E, Moreno A, Valls A, Romero-Aroca P and de la Riva-Fernandez S (2016). A Fuzzy Random Forest Approach for the Detection of Diabetic Retinopathy on Electronic Health Record Data. In CCIA 169–174.

Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI and McCulloch RE (2016). Bayes and big data: The consensus Monte Carlo algorithm. International Journal of Management Science and Engineering Management 11 78–88.

Skevofilakas M, Zarkogianni K, Karamanos BG and Nikita KS (2010). A hybrid Decision Support System for the risk assessment of retinopathy development as a long term complication of Type 1 Diabetes Mellitus. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology 6713–6716. IEEE.

Stephens M (2000). Bayesian analysis of mixture models with an unknown number of components-an alternative to reversible jump methods. Annals of Statistics 40–74.

Sun Y and Zhang D (2019). Diagnosis and Analysis of Diabetic Retinopathy based on Electronic Health Records. IEEE Access.

Ting DSW, Cheung GCM and Wong TY (2016). Diabetic retinopathy: global prevalence, major risk factors, screening practices and public health challenges: a review. Clinical & Experimental Ophthalmology 44 260–277. [PubMed: 26716602]

Xu L, Crammer K and Schuurmans D (2006). Robust support vector machine training via convex outlier ablation. In AAAI 6 536–542.

Yang P, Li X-L, Mei J-P, Kwoh C-K and Ng S-K (2012). Positive-unlabeled learning for disease gene identification. Bioinformatics 28 2640–2647. [PubMed: 22923290]

Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, Chen S-J, Dekker JM, Fletcher A, Grauslund J et al. (2012). Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care 35 556–564. [PubMed: 22301125]

Zawistowski M, Sussman JB, Hofer TP, Bentley D, Hayward RA and Wiitala WL (2017). Corrected ROC analysis for misclassified binary outcomes. Statistics in Medicine 36 2148–2160. [PubMed: 28245528]

Zhang X, Saaddine JB, Chou C-F, Cotch MF, Cheng YJ, Geiss LS, Gregg EW, Albright AL, Klein BE and Klein R (2010). Prevalence of diabetic retinopathy in the United States, 2005–2008. JAMA 304 649–656. [PubMed: 20699456]

Zuluaga MA, Hush D, Leyton EJD, Hoyos MH and Orkisz M (2011). Learning from only positive and unlabeled data to detect lesions in vascular CT images. In International Conference on Medical Image Computing and Computer-Assisted Intervention 9–16. Springer.

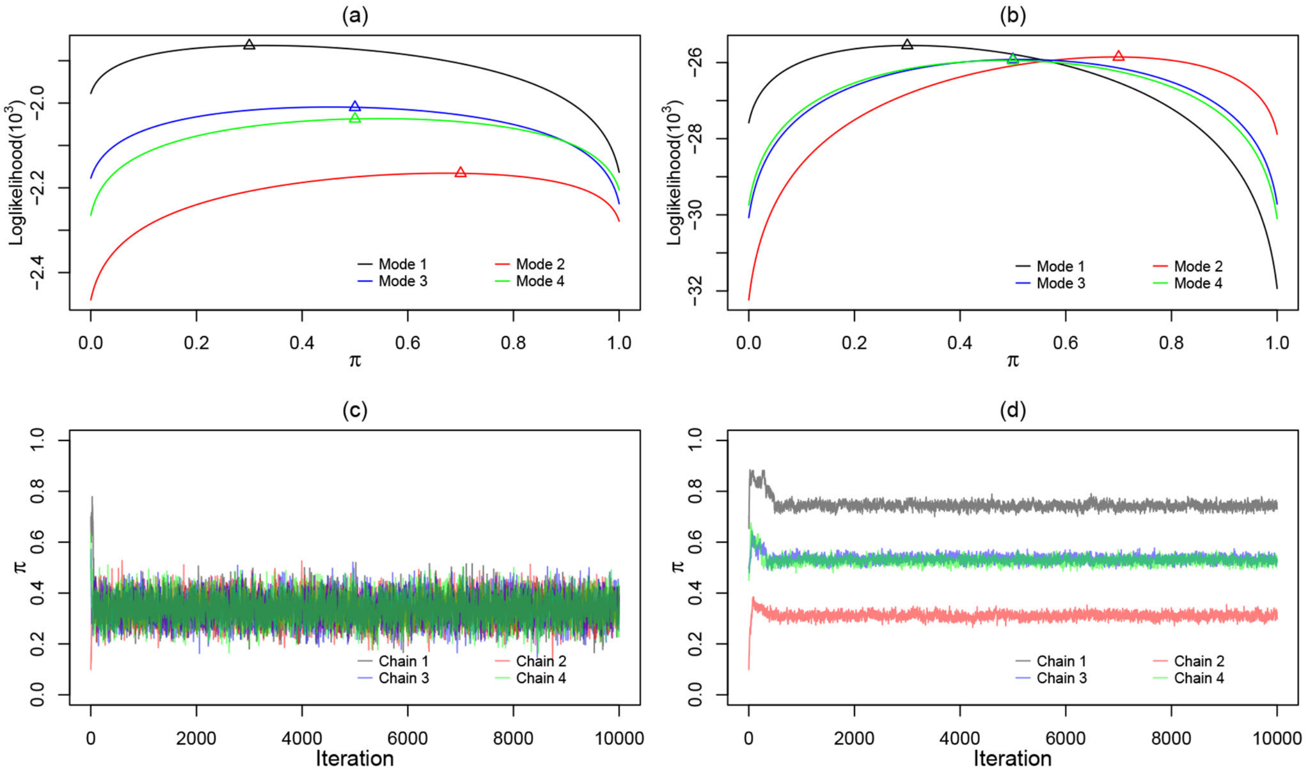

Fig 1.

Plots of log-likelihood function with respect to **π** and MCMC trace plots with distinct initial values for **π**: (a)(c). Synthetic data set 1: n_p = 100, n_u = 200. (b)(d). Synthetic data set 2: n_p = 100, n_u = 4, 000

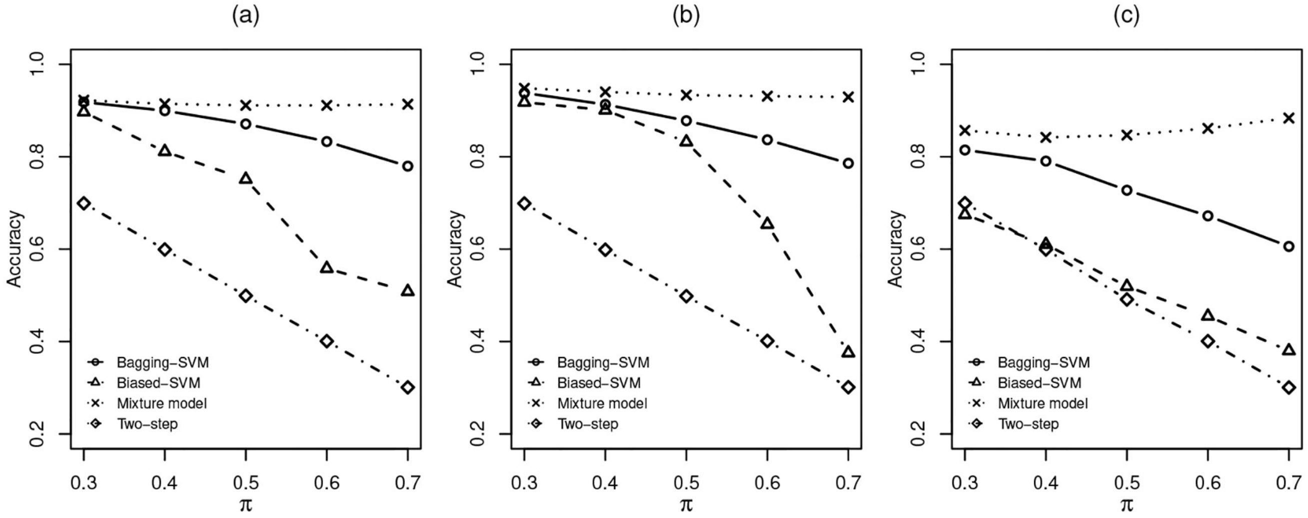

Fig 2.

Accuracy under different settings: (a). Multivariate t. (b). Gamma. (c). Cauchy

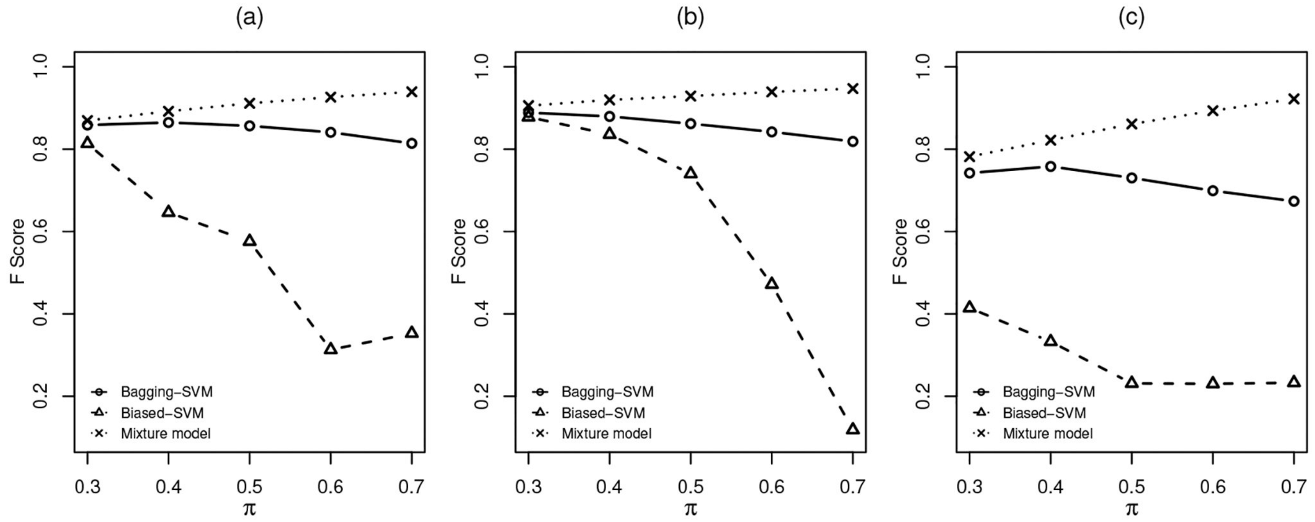

Fig 3.

F-score under different settings: (a). Multivariate t. (b). Gamma. (c). Cauchy

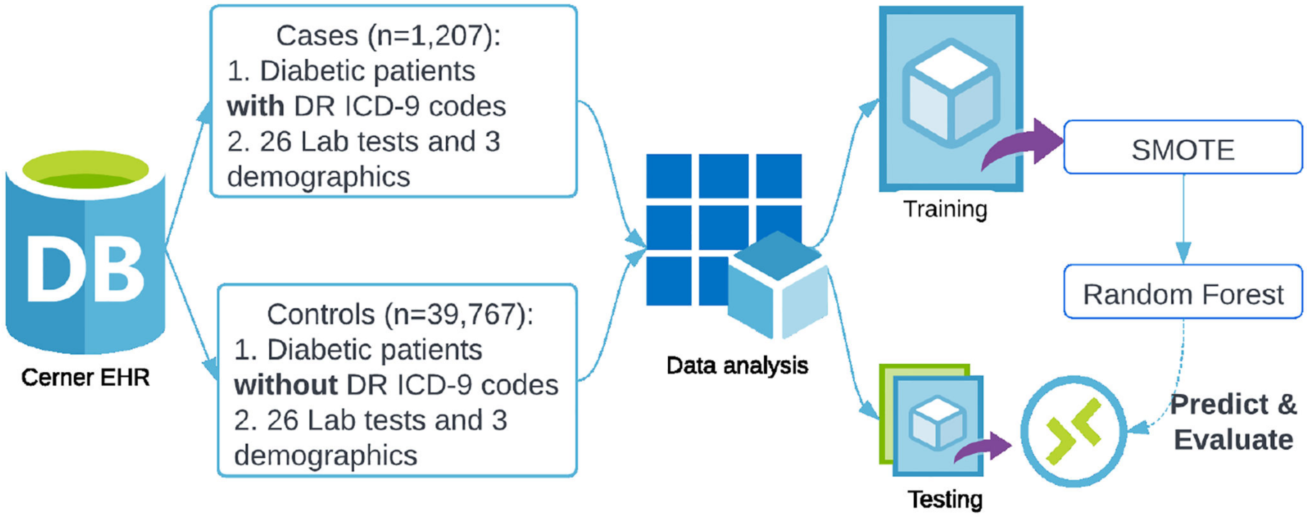
Cases (n=1,207):
1. Diabetic patients **with** DR ICD-9 codes
2. 26 Lab tests and 3 demographics
DB
Cerner EHR
Controls (n=39,767):
1. Diabetic patients **without** DR ICD-9 codes
2. 26 Lab tests and 3 demographics
Data analysis
Training
SMOTE
Random Forest
Testing
Predict & Evaluate

Fig 4.
Data extraction and initial analysis

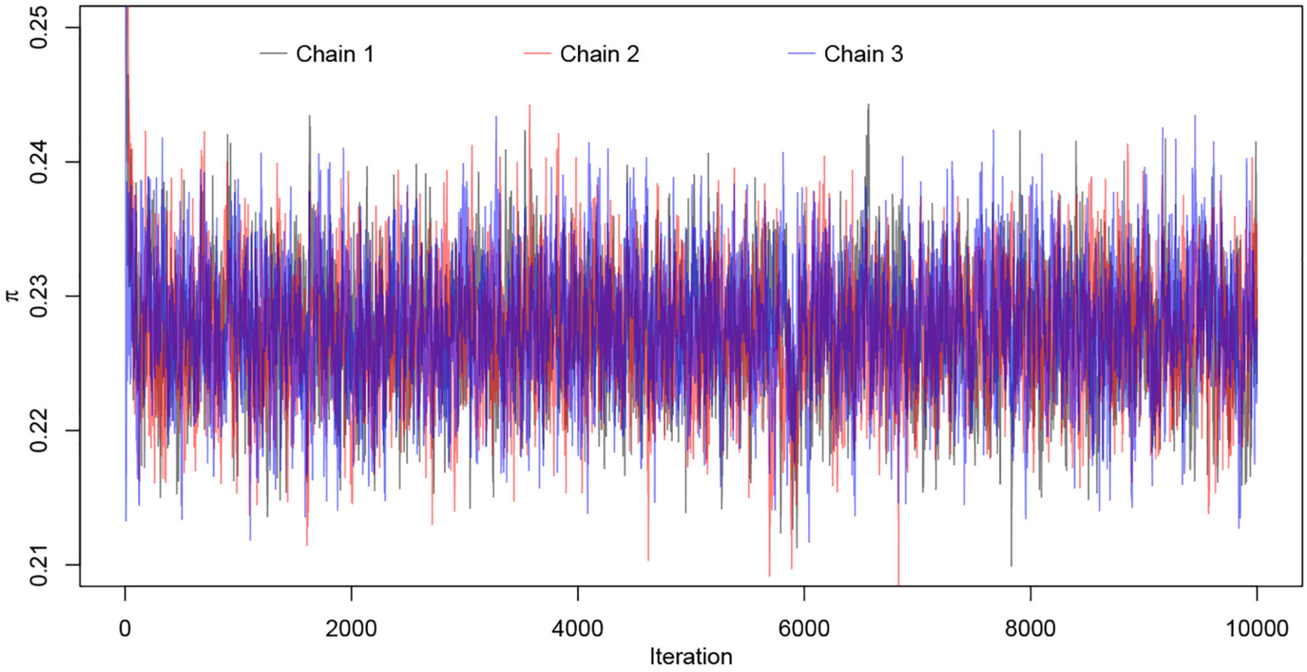

Fig 5.

MCMC convergence for **π**: three chains with distinct initial values.

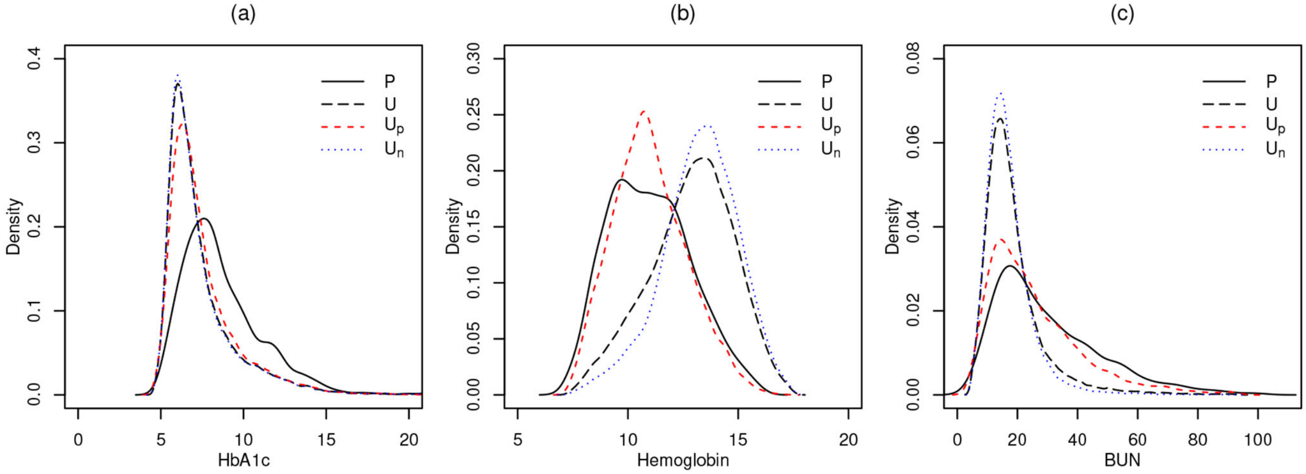

Fig 6.

Distributions of the positive group $\mathscr{P}$ (black solid line), the unlabelled group $\mathscr{U}$ (black dashed line), the re-labelled positives $\mathscr{U}_p$ (red dashed line) and the re-labelled negatives $\mathscr{U}_n$ (blue dotted line). Plot (a): HbA1c; plot (b): Hemoglobin; plot (c): BUN.

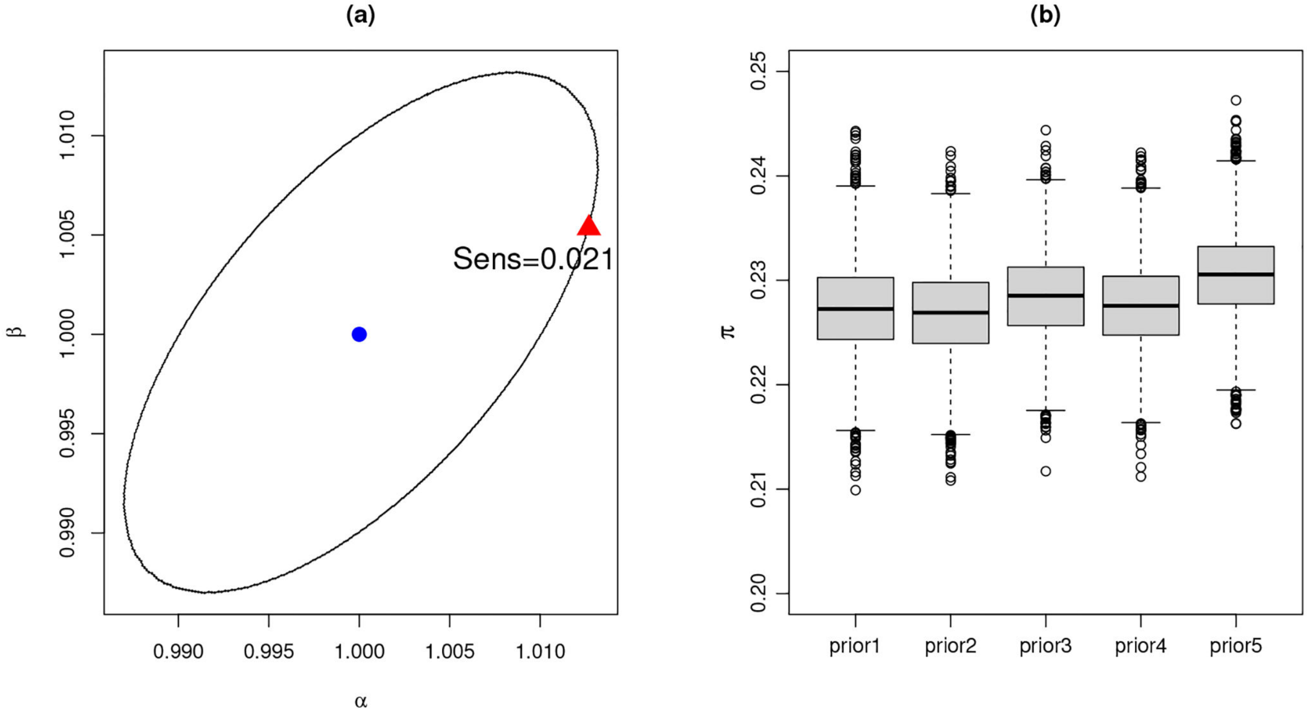

Fig 7.

Sensitivity analysis for π: (a). The ϵ-local circular sensitivity. The ellipse is the contour line around the centroid (blue dot) at $(\alpha, \beta) = (1, 1)$ with Hellinger distance $\epsilon = 0.00354$. The worst-case sensitivity is obtained at the red triangle with sensitivity value 0.021. (b). An ad hoc sensitivity analysis with modified prior hyperparameters. The box plots show posterior samples of π with respect to each prior modification.

Table 1

Performance metrics under Case 1: t-distribution. Table shows mean metrics (standard deviation).

π	Method	Accuracy	*F*-score	Sensitivity	Specificity
0.3	Two-step	0.6991 (0.0067)	-	0 (0)	1 (0)
	Biased-SVM	0.8968 (0.0237)	0.8137 (0.0571)	0.7675 (0.1034)	0.9526 (0.0137)
	Bagging-SVM	0.9182 (0.0055)	0.8587 (0.0103)	0.8269 (0.0179)	0.9574 (0.0061)
	Mixture model	0.9223 (0.0051)	0.8696 (0.0104)	0.8586 (0.0258)	0.9498 (0.0091)
0.4	Two-step	0.5992 (0.0074)	-	0 (0)	1 (0)
	Biased-SVM	0.8111 (0.1159)	0.6462 (0.1440)	0.5785 (0.3160)	0.9679 (0.0230)
	Bagging-SVM	0.8996 (0.0087)	0.8649 (0.0132)	0.8031 (0.0222)	0.9641 (0.0058)
	Mixture model	0.9144 (0.0041)	0.8921 (0.0068)	0.8890 (0.0174)	0.9311 (0.0090)
0.5	Two-step	0.4991 (0.0073)	-	0 (0)	1(0)
	Biased-SVM	0.7511 (0.1882)	0.5760 (0.0792)	0.5348 (0.3939)	0.9703 (0.0233)
	Bagging-SVM	0.8709 (0.0121)	0.8567 (0.0156)	0.7721 (0.0249)	0.9700 (0.0046)
	Mixture model	0.9112 (0.0061)	0.9115 (0.0068)	0.9134 (0.0196)	0.9087 (0.0192)
0.6	Two-step	0.4009 (0.0074)	-	0 (0)	1 (0)
	Biased-SVM	0.5581 (0.2157)	0.3130 (0.0454)	0.2945 (0.3963)	0.9522 (0.1811)
	Bagging-SVM	0.8328 (0.0256)	0.8410 (0.0270)	0.7403 (0.0379)	0.9713 (0.0110)
	Mixture model	0.9111 (0.0078)	0.9264 (0.0065)	0.9381 (0.0167)	0.8713 (0.0259)
0.7	Two-step	0.3009 (0.0067)	-	0 (0)	1 (0)
	Biased-SVM	0.5082 (0.2436)	0.3527 (0.0467)	0.3028 (0.3553)	0.9858 (0.0193)
	Bagging-SVM	0.7796 (0.0469)	0.8142 (0.0472)	0.6983 (0.0659)	0.9690 (0.0179)
	Mixture model	0.9136 (0.0123)	0.9393 (0.0082)	0.9560 (0.0147)	0.8154 (0.0523)

Table 2

Performance metrics under Case 2: gamma distribution. Table shows mean metrics (standard deviation).

π	**Method**	**Accuracy**	***F*-score**	**Sensitivity**	**Specificity**
0.3	Two-step	0.6988 (0.0062)	-	0 (0)	1 (0)
	Biased-SVM	0.9178 (0.1160)	0.8781 (0.0782)	0.8443 (0.0377)	0.9496 (0.1795)
	Bagging-SVM	0.9380(0.0042)	0.8888 (0.0090)	0.8233 (0.0238)	0.9875 (0.0059)
	Mixture model	0.9481 (0.0067)	0.9058 (0.0131)	0.8303 (0.0233)	0.9989 (0.0010)
0.4	Two-step	0.5986 (0.0064)	-	0 (0)	1 (0)
	Biased-SVM	0.9010 (0.0836)	0.8360 (0.0109)	0.7720 (0.2110)	0.9879 (0.0069)
	Bagging-SVM	0.9131 (0.0091)	0.8796 (0.0144)	0.7923 (0.0255)	0.9940 (0.0035)
	Mixture model	0.9402 (0.0073)	0.9197 (0.0106)	0.8542 (0.0198)	0.9979 (0.0017)
0.5	Two-step	0.4982 (0.0062)	-	0 (0)	1 (0)
	Biased-SVM	0.8320 (0.1507)	0.7402 (0.0133)	0.6734 (0.3071)	0.9909 (0.0069)
	Bagging-SVM	0.8780 (0.0136)	0.8620 (0.0174)	0.7608 (0.0276)	0.9961 (0.0034)
	Mixture model	0.9333 (0.0100)	0.9290 (0.0115)	0.8713 (0.0223)	0.9958 (0.0034)
0.6	Two-step	0.4014 (0.0064)	-	0 (0)	1 (0)
	Biased-SVM	0.6536 (0.2411)	0.4715 (0.0150)	0.4255 (0.4052)	0.9948 (0.0064)
	Bagging-SVM	0.8367 (0.0200)	0.8421 (0.0221)	0.7292 (0.0337)	0.9972 (0.0031)
	Mixture model	0.9310 (0.0116)	0.9390 (0.0112)	0.8895 (0.0224)	0.9928 (0.0061)
0.7	Two-step	0.3012 (0.0063)	-	0 (0)	1 (0)
	Biased-SVM	0.3755 (0.1901)	0.1184 (0.0347)	0.1152 (0.3015)	0.9759 (0.0906)
	Bagging-SVM	0.7858 (0.0220)	0.8188 (0.0224)	0.6946 (0.0321)	0.9975 (0.0034)
	Mixture model	0.9292 (0.0136)	0.9472 (0.0103)	0.9109 (0.0249)	0.9720 (0.0563)

Table 3

Performance metrics under Case 3: Cauchy distribution. Table shows mean metrics (standard deviation).

π	**Method**	**Accuracy**	***F*-score**	**Sensitivity**	**Specificity**
0.3	Two-step	0.6994 (0.0062)	-	0 (0)	1 (0)
	Biased-SVM	0.6746 (0.1430)	0.4143 (0.2151)	0.4759 (0.3115)	0.7604 (0.2360)
	Bagging-SVM	0.8146 (0.1036)	0.7422 (0.0705)	0.8372 (0.0468)	0.8047 (0.1614)
	Mixture model	0.8570 (0.0141)	0.7819 (0.0133)	0.8509 (0.0352)	0.8596 (0.0326)
0.4	Two-step	0.5991 (0.0066)	-	0 (0)	1 (0)
	Biased-SVM	0.6101 (0.1161)	0.3326 (0.2237)	0.3550 (0.3290)	0.7819 (0.2557)
	Bagging-SVM	0.7906 (0.0953)	0.7581 (0.0887)	0.7957 (0.0790)	0.7871 (0.1385)
	Mixture model	0.8418 (0.0093)	0.8219 (0.0079)	0.9104 (0.0278)	0.7959 (0.0303)
0.5	Two-step	0.4911 (0.0068)	-	0 (0)	1 (0)
	Biased-SVM	0.5193 (0.0561)	0.2312 (0.2179)	0.2286 (0.2854)	0.8127 (0.2630)
	Bagging-SVM	0.7275 (0.1225)	0.7303 (0.1148)	0.7277 (0.1170)	0.7275 (0.1743)
	Mixture model	0.8468 (0.0077)	0.8612 (0.0056)	0.9484 (0.0191)	0.7448 (0.0291)
0.6	Two-step	0.4009 (0.0066)	-	0 (0)	1 (0)
	Biased-SVM	0.4549 (0.0768)	0.2302 (0.2137)	0.2146 (0.2818)	0.8130 (0.2729)
	Bagging-SVM	0.6719 (0.1523)	0.6992 (0.1646)	0.6588 (0.1735)	0.6908 (0.2009)
	Mixture model	0.8614 (0.0073)	0.8935 (0.0052)	0.9703 (0.0105)	0.6987 (0.0272)
0.7	Two-step	0.3006 (0.0061)	-	0 (0)	1 (0)
	Biased-SVM	0.3800 (0.0997)	0.2329 (0.2107)	0.1995 (0.2483)	0.8010 (0.2483)
	Bagging-SVM	0.6058 (0.1291)	0.6736 (0.1193)	0.5984 (0.1445)	0.6230 (0.2268)
	Mixture model	0.8837 (0.0062)	0.9221 (0.0041)	0.9845 (0.0053)	0.6500 (0.0238)

Table 4

Parameter estimation under the t-distribution for $\pi = 0.3$

		True Value	Bias	MSE
	π	0.3	0.0019	0.0192
	p_{11}	0.5	−0.0164	0.0294
	p_{21}	0.5	0.0063	0.0109
Male	$\mu_{211} - \mu_{111}$	2	0.0043	0.1147
	$\mu_{212} - \mu_{112}$	2	−0.0682	0.1113
	$\mu_{213} - \mu_{113}$	2	−0.0027	0.1027
Female	$\mu_{221} - \mu_{121}$	2	−0.0079	0.0842
	$\mu_{222} - \mu_{122}$	2	−0.0127	0.0867
	$\mu_{223} - \mu_{123}$	2	−0.0257	0.0963

Table 5

Posterior means (standard deviations) for parameters of inferential interest in the mixture model for Cerner EHR data.

	DR		**Non-DR**	
	Male	**Female**	**Male**	**Female**
Proportion	0.5344 (0.0030)	0.4656 (0.0030)	0.4735 (0.0032)	0.5265 (0.0032)
HbA1c	0.9503 (0.0004)	0.8901 (0.0004)	0.9303 (0.0005)	0.8678 (0.0004)
Hemoglobin	4.0014 (0.0037)	2.1320 (0.0012)	4.6537 (0.0034)	2.2783 (0.0010)
BUN	3.7724 (0.0067)	3.2758 (0.0057)	3.1203 (0.0042)	2.7608 (0.0039)