PIMixture Manual

Noorie Hyun and Li Cheung

09/18/2018

1. Introduction

The webtool, PIMixture estimates the absolute risk of asymptomatic disease or disease precursors. Because asymptomatic disease/disease precursors are often discovered through screening, collected data may present challenges for absolute and relative risk estimation. First, the time of disease onset for an individual is unobserved and falls between screening visits (interval-censoring). Second, there may be prevalent disease present at the initial screening visit. Third, prevalent disease is not always immediately diagnosed (e.g. in those with negative or equivocal screening results), and thus some disease found at later screening visits are missed prevalent disease. (See Figure 1 for an example of such data.)

Typically, disease prevalence (a person's likelihood of currently having a disease) and disease incidence (a person's likelihood of acquiring future disease) are analyzed separately, but this is not always possible because it is not always known whether detected disease is prevalent or incident. PIMixture approaches this problem by simultaneously analyzing prevalent and incident disease. In PIMixture, logistic regression is used to model disease prevalence and proportional hazards models are used for disease incidence subject to interval-censoring; model parameters are then jointly estimated to account for disease where it is uncertain whether it is prevalent or incident.

The PIMixture web tool provides multiple options for defining the baseline hazard of the proportional hazard model -- a fully parametric model that assumes a Weibull baseline hazard; a weakly-parametric model that uses integrated B-splines to model the baseline hazard; and a semi-parametric model. Both the weakly-parametric model and the semiparametric model also supports stratified random sample.

Both absolute risk as a function of time (beginning with the initial screen) and relative risk, when covariates are included in the model, are estimated. Relative risks are presented as odds ratios for prevalent disease and hazards ratios for incident disease.

The PIMixture webtool is based on R-package PIMixture (Cheung et al., 2017; Hyun et al., 2017). A thorough study on application of PIMixture models to electronic health records data as compared to existing methods has been done by Landy et al. 2018.

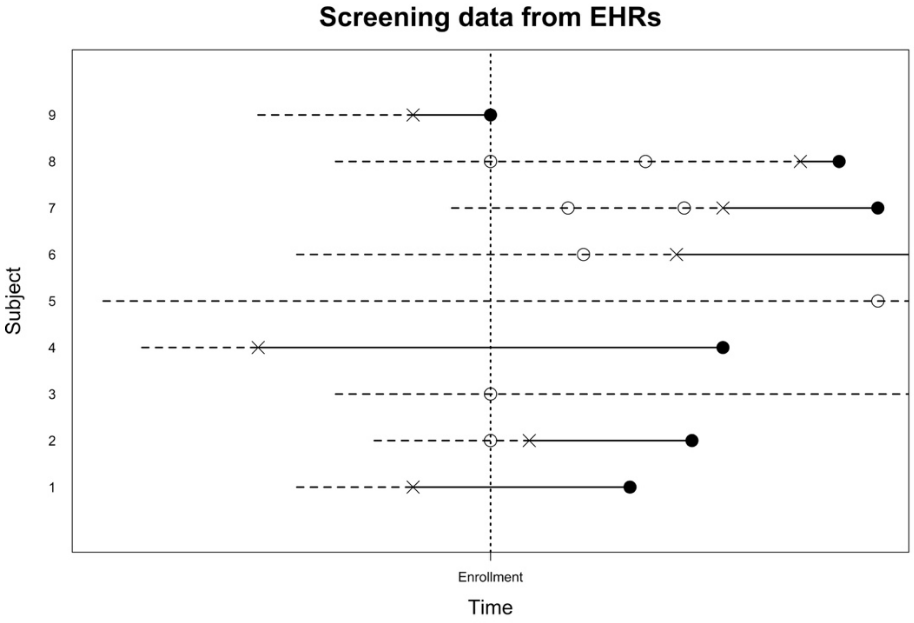


Figure 1: Screening data (obtained from electronic health records) for 9 subjects. Subjects become susceptible to disease (start of dashed lines) at some point before enrollment (vertical dotted line), may acquire clinically-detectable disease (denoted with x) and then may be subsequently diagnosed (solid circles). Disease status is known only at specific times (unfilled circles represent known disease-free status; solid circles represent known diseased status). EHR=electronic health records.

2. Models

PIMixture employs logistic regression models for prevalence and proportional hazards models for incidence and allows including multiple predictors or no predictor. For example, we consider a logistic regression model and a proportional hazards model including one predictor, x.

Logistic regression model: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x,$ (1)

where p=probability of disease presence at baseline and 0< p <1; β_0 is the intercept, e^{β_1} is the odds ratio for disease presence when x=0 to when x=1.

Proportional hazards model: $\lambda(t \mid x) = \lambda_0(t) e^{\gamma_1 x}$, $\mathrm{t} > 0,$ (2)

where $\lambda_0(t)$ is a baseline hazard function for disease occurrence; $e^{\gamma_1}$is the hazards ratio for disease incidence when x=0 to when x=1; The baseline hazard function $\lambda_0(t)$ is the instantaneous rate at which a disease occurs to a subject, given that the subject has been disease-free up to time t.

For incidence estimation, we need to estimate $\gamma_1$as well as $\lambda_0(t)$ (or a cumulative baseline hazard function $\Lambda_0(\mathrm{t}) = \int_0^{\mathrm{t}} \lambda_0(s) ds$), and there are three options in PIMixture to estimate $\lambda_0(t)$, "parametric", "weakly-parametric" and "semiparametric" methods. Once we estimate $\lambda_0(t)$ or $\Lambda_0(t)$ and γ_1, we can estimate survival function, $S(t \mid x) = \Pr(T > t|x) = e^{-\Lambda_0(t) e^{\gamma_1 x}}$.

1. Parametric model: we assume the baseline hazard function is generated from a Weibull distribution, and accordingly, $\lambda_0(t) = \frac{\alpha_1}{\alpha_2}\left(\frac{t}{\alpha_2}\right)^{\alpha_1 - 1}$, for $t \geq 0$ and $\alpha_1, \alpha_2 > 0$. Weibull distributions are determined by two parameters α_1 and α_2. Hence, if the distribution assumption is correct, parametric models are the most efficient.
2. Weakly-parametric model: Like piecewise constant hazard functions, we approximate the cumulative hazard function $\Lambda_0(\mathrm{t}) \approx \sum_{\mathrm{k}=1}^{\mathrm{K}} \alpha_k B_k(t)$, using integrated cubic B-splines $\mathrm{B_k}(t)$'s and the piecewise intervals are determined by quantiles of observation times. The weakly-parametric method of PIMixture needs to estimate total 7 parameters $\mathrm{\alpha_k}$'s by default.
3. Semiparametric model: we assume $\Lambda_0(\mathrm{t})$ is a step function with jumps at unique observation times at most. Hence as the number of unique observation times increases, the number of parameters for $\Lambda_0(\mathrm{t})$ increases. The iterative convex algorithm is employed to nonparametrically estimate $\Lambda_0(\mathrm{t})$.

The assumptions are strong in the order of "Parametric" > "Weakly-parametric" > "Semiparametric", whereas if data satisfies the assumptions, variance of estimates is in the order of "Parametric" < "Weakly-parametric" < "Semiparametric". Computation burden increases according to the number of parameters to estimate.

3. Sampling Design

PIMixture provides two options for unweighted and weighted data. Specifically, unweighted data represents a simple random sample or an entire cohort; everyone of a simple random sample has an equal selection probability, so we don't have to add sampling weight. Weighted data in PIMixture represents a stratified random sample, of which selection probabilities vary across strata and are the same within a stratum, and the selection probabilities are known. For weighted data analysis, users additionally specify two variables for strata and sampling weights (>=1).

Table 1 Available options of PIMixture

	Parametric models*	Weakly-parametric models**	Semiparametric models**
Unweighted data (a simple random sample or a cohort)	Available	Available	Available (standard error and confidence interval are not available)
Weighted data (stratified random sample with known sample weights)	Not available	Available	Available (standard error and confidence interval are not available)

* Cheung et al. 2017; ** Hyun et al. 2017.

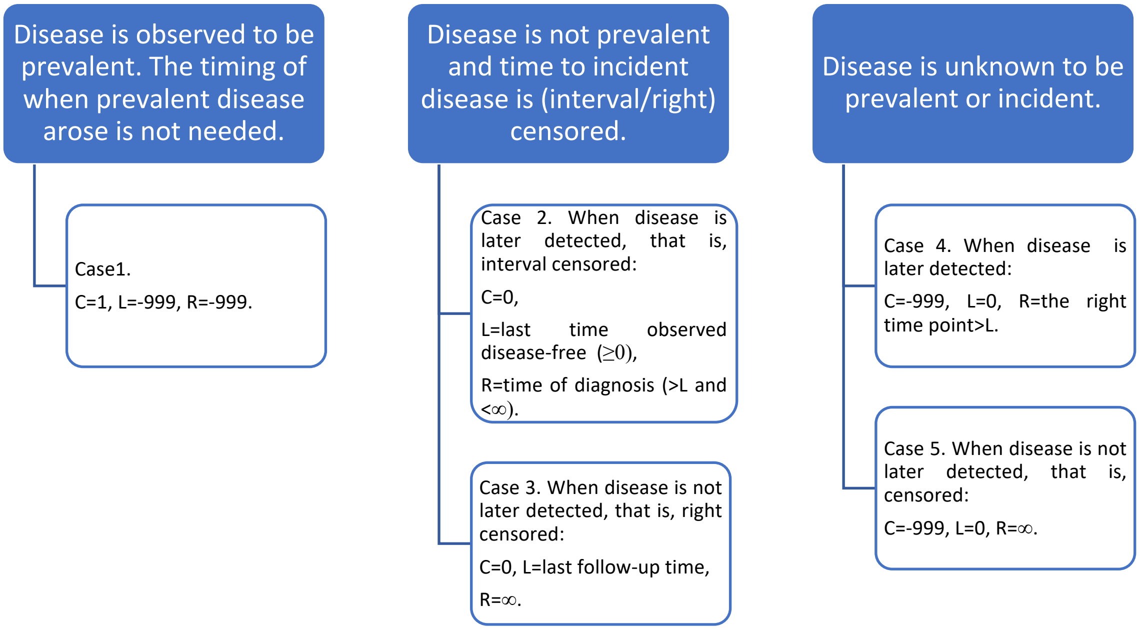
4. Data

This section explains what types of data should be included in an input dataset and coding rules. The file format of Comma Separate Value (CSV) is allowed.

4.1 Outcome (necessary information)

The outcome of interest is the time of clinically-detectable disease onset, and three variables for the outcome should be included in the input data: for simplicity, we define C=prevalence indicator, L=left time point, i.e. the latest time at which a subject is disease-free, R=right time point, i.e., the earliest time at which a subject is diagnosed with a disease. These variable names can be changed. In the webtool, users can choose which variables correspond to “C”, “L” and “R”. General coding rules are as following:

(1) C=1 if prevalent disease, C=0 if no prevalent disease, C=-999 if unknown status. Note that even if disease status is not ascertained at the initial screen, a later screen that ascertains the absence of disease means we know there was no prevalent disease.
(2) L and R have values equal to or greater than 0 (any unit, such as day, month and year can be used); however, when C=1, L=R=-999.
(3) For right/interval censoring, L is smaller than R.
(4) For right censoring, R=Inf, where Inf means infinity, ∞.
(5) L should not be equal to R except when C=1 because PIMixture does not handle exact event time. However, if data includes exact event times, users can use a trick, adding a very small interval to the exact event times to define “L” and “R”.



Disease is observed to be prevalent. The timing of when prevalent disease arose is not needed.

Case1.

C=1, L=-999, R=-999.

Disease is not prevalent and time to incident disease is (interval/right) censored.

Case 2. When disease is later detected, that is, interval censored:

C=0,

L=last time observed disease-free (≥0),

R=time of diagnosis (>L and <∞).

Case 3. When disease is not later detected, that is, right censored:

C=0, L=last follow-up time, R=∞.

Disease is unknown to be prevalent or incident.

Case 4. When disease is later detected:

C=-999, L=0, R=the right time point>L.

Case 5. When disease is not later detected, that is, censored:

C=-999, L=0, R=∞.

Figure 1 All possible cases we may observe

Table 2 Example of input data (description column is not necessary).

C	L	R	Description
1	-999	-999	case 1: the subject has disease present at baseline
0	180	365	case 2: the subject had been disease -free at visit 1 (180 days since baseline) and was diagnosed with a disease at visit 2 (365 days)

0	1825	Inf	case 3: the last follow-up is 1825 days since baseline and no incidence is observed.
-999	0	730	case 4: baseline status is not observed and at visit 1 (730 days) disease is observed.
-999	0	Inf	case 5: the subject is uninformative for risk estimation

4.2 Predictors (optional information)

In PIMixture, it is optional to include predictors in logistic and proportional hazards models. When included, relative risks are given in terms of odd ratios for prevalent disease and hazard ratios for incident disease. There are two options for predictors, continuous or categorical variables. For example, let us consider a logistic and proportional hazards models for assessing the association between the probability of being osteoporosis and body mass index (BMI). BMI can be coded either continuous or categorical (1: BMI≤18.5, 2: 18.5<BMI≤25, 3: 25<BMI≤30, 4: BMI>30 or low: BMI≤18.5, normal: 18.5<BMI≤25, overweight: 25<BMI≤30, obese: BMI>30).

4.2.1 Continuous Predictor

Continuous predictor values should be numerical:

Table 3 Example of data including continuous BMI as a predictor

C	L	R	BMI
1	-999	-999	17
0	180	365	19
0	1825	Inf	27
-999	0	730	24
-999	0	Inf	31

The logistic and proportional hazards models are $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 BMI \quad , \lambda(\, t \mid x\,) = \lambda_0(t) e^{\gamma_1 BMI}$.

A reference level (or group) for continuous predictors is 0 by default; however, users can set the reference level to be another level rather than 0; when users set a reference level for BMI to be "20". Then it shifts BMI levels to be centered at 20.

C	L	R	BMI
1	-999	-999	-3
0	180	365	-1
0	1825	Inf	7
-999	0	730	4
-999	0	Inf	11

4.2.2 Categorical Predictor

Categorical predictors allow two types of data, numerical and characteristic levels:

C	L	R	BMI
1	-999	-999	1
0	180	365	2
0	1825	Inf	3
-999	0	730	2
-999	0	Inf	4

Or

C	L	R	BMI
1	-999	-999	low
0	180	365	normal
0	1825	Inf	overweight
-999	0	730	normal
-999	0	Inf	obese

Users should specify a reference level for categorical predictors. For example, when we set a reference level for BMI to be 2 (or “normal’), the logistic and proportional hazards models are

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 I(BMI = "low") + \beta_2 I(BMI = "overweight") + \beta_3 I(BMI = "obese"),$$

$$\lambda(t \mid x) = \lambda_0(t) e^{\gamma_1 I(BMI=\text{low})+\gamma_2 I(BMI=\text{overweight})+\gamma_3 I(BMI=obese)},$$

where I(BMI=a)=1 if BMI=a; 0 otherwise.

Please note that the number of parameters for continuous and categorical variables are 1 and the total levels-1 in logistic and proportional hazards models.

4.2.3 Interaction Effects

Let us consider two factors, sex (binary) and BMI (categorical) and interaction effects between two factors. Then the logistic and proportional hazards models are as following:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 I(BMI = "low") + \beta_2 I(BMI = "overweight") + \beta_3 I(BMI = "obese") + \beta_4 I(sex = \text{"Male"}) + \beta_5 I(sex = \text{"Male" and BMI="low"}) + \beta_6 I(sex = \text{"Male" and BMI="overweight"}) + \beta_7 I(sex = \text{"Male" and BMI="obese"}),$$

$$\lambda(t \mid x) = \lambda_0(t) e^{\gamma_1 I(BMI=\text{low})+\gamma_2 I(BMI=\text{overweight})+\gamma_3 I(BMI=obese)+\gamma_3 I(sex=\text{"Male"})} \text{ x } e^{\gamma_4 I(sex=\text{"Male" and BMI="low"})+\gamma_5 I(sex=\text{"Male" and BMI="overweight"})+\gamma_6 I(sex=\text{"Male" and BMI="obese"})}.$$

5. Prediction

Based on the regression parameters and/or cumulative hazard function, we can predict prevalence and incidence using an independent data (called test data) including the predictors used for estimating the parameters and cumulative hazard function. For prediction, users need to upload fitted model, test data and input time points. When time point=0, predicted probabilities mean the prevalences of subgroups characterized by specific predictors; when time point=t>0, predicted cumulative risk up to time t includes prevalence too. Test data should include the same variables used for fitting the model.

6. Tips for handling errors

- The general coding rules described in Section 4.1 are one of primary screening filters. For checking if coding for the outcome variables is correct, users can try a simple model with no predictor.

- When models include too many predictors, it can cause convergence issue because the number of events in a subgroup may be too small.

7. References

1. Cheung LC, Pan Q, Hyun N, Schiffman M, Fetterman B, Castle PE, Katki HA. Mixture models for left-censored and irregularly interval-censored data: Application to a cancer screening cohort assembled from electronic health records. *Stat Med*. 2017 Sep 30; 36(22):3583-95.

2. Hyun N, Cheung LC, Pan Q, Schiffman M, Katki HA. Flexible risk prediction models for left or interval-censored data from electronic health records. *Ann Appl Stat*. 2017 Jun;11(2): 1063-84.

3. Landy R, Cheung LC, Schiffman M, Gage JC, Hyun N, Wentzensen N, Sasieni PD, Katki HA. Challenges in risk estimation using routinely collected clinical data: The example of estimating cervical cancer risks from electronic health-records. *Prev Med.* 2018 June; 111:429-435.