Update of TBM Latent Class Analysis

Trinh Dong Huu Khanh

4/12/2020

Contents

1 Overview

1.1 Objectives

- Give an improved score table based on Statistics replacing the current, partially expertise-based, TBM definition Consensus Score (2010)
- Give a predicted probability of TBM at roughly admission time...
- ...And after confirmation test results (mainly all negative)

1.2 Concept and Terms

- Latent Class Analysis = Finite Bernoulli Mixture Model
- Manifest variable = (Confirmation) Test = Indicator
- Predictor = Co-variate
- Prevalence = Theta = Latent class = Latent variable

2 Current approach

2.1 Core model

I use a similar set of clinical and laboratorial signs and symptoms to the TBM definition score as **predictors**. Weakly t distribution are used.

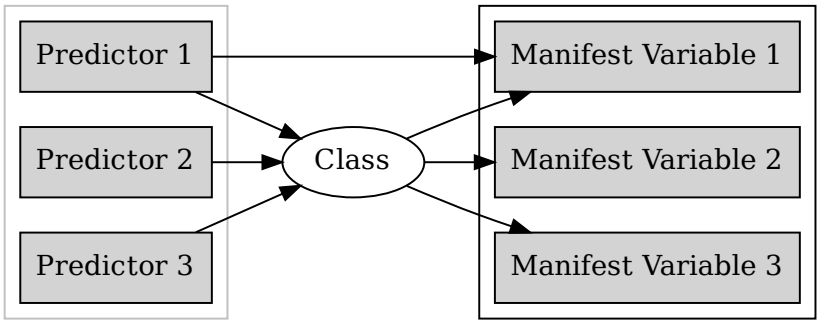

Figure 1: Model Concept

Continuous predictors are transformed to have a symmetric distribution.

$$\begin{aligned} nu &\sim -(2, 0.1) \\ a_0 &\sim \mathrm{t}(\nu, 0, 2.5) \\ a &\sim \mathrm{t}(\nu, 0, 1) \end{aligned}$$

Three confirmation tests are chosen as manifest variables. Their **False positive rate (FPR)** and **True positive rate (TPF)** are assumed to follow a Normal distribution on the logit scale.

Informative priors are imposed on **FPR** according to priors knowledge and experience (thanks to Joe and Julie from TB). For **TPR**, a weakly informative was used with sufficient collapse toward .5 after transformation back to normal scale.

$$TPR \sim \mathcal{N}(0, .6)$$

$$\begin{aligned} FPR_{Smear} &\sim \mathcal{N}(logit(.001), \ \ .82) \\ FPR_{Mgit} &\sim \mathcal{N}(logit(.001), \ \ .82) \\ FPR_{Xpert} &\sim \mathcal{N}(logit(.005), \ 1.59) \end{aligned}$$

Bacillary burden was assumed to be a random variable whose mean depends on **HIV** status.

$$\begin{aligned} b_{RE} &\sim \mathrm{t}(\nu, 0, 1) \\ b_{HIV} &\sim \mathrm{t}(\nu, 0, 1) \\ RE &\sim \mathcal{N}(0, 1) \\ bac\_load &= b_{RE} * RE + b_{HIV} * HIV \end{aligned}$$

2.2 Imputation model

4 cases with missing data in confirmation tests are removed.

Missing data in predictors are assumed MAR and imputed within the model using Stan.

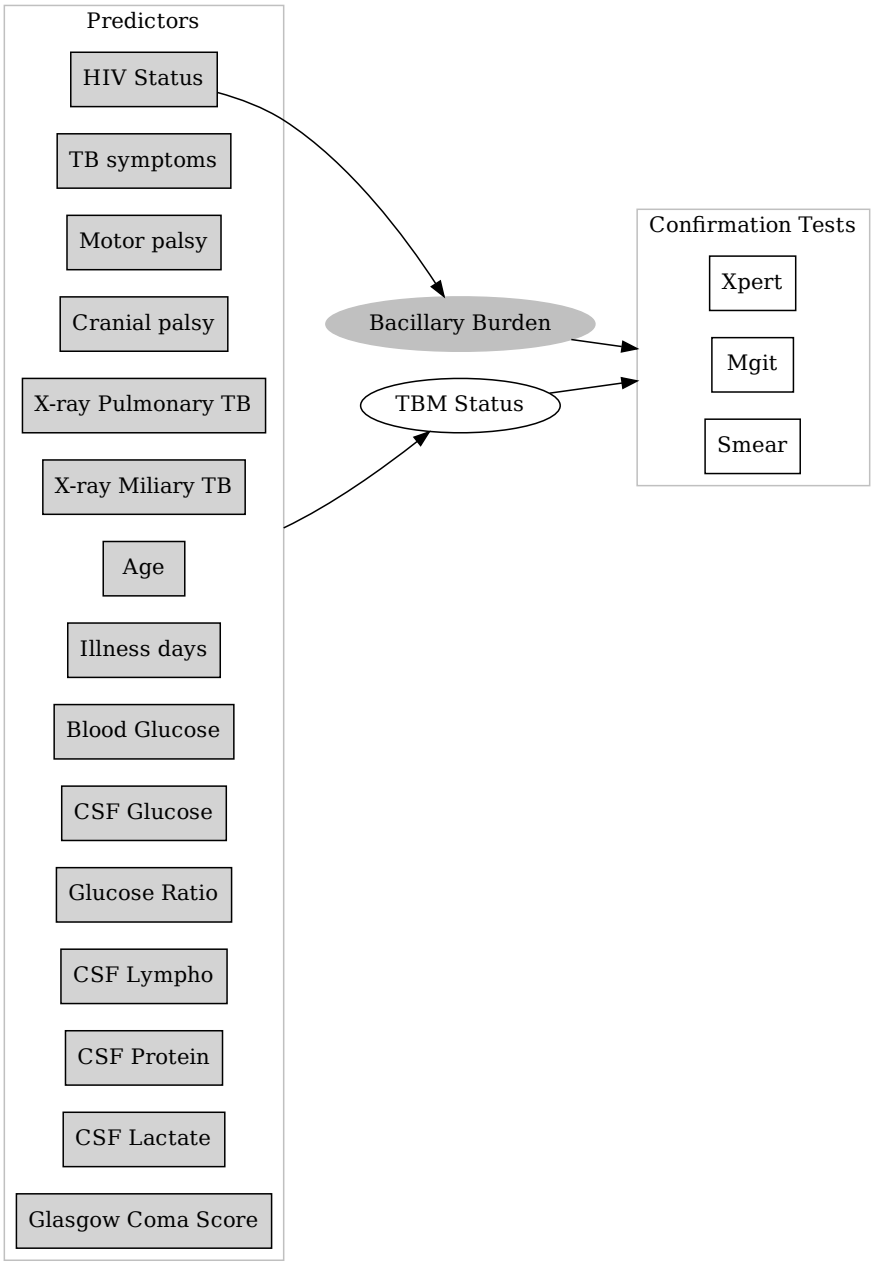

Figure 2: Core model

- HIV is imputed using **logistic regression** and have the probabilities marginalised.

$$HIV_{a_0} \sim \mathrm{t}(\nu, 0, 1)$$

- Those which are compound of several other binary variables, such as TB Symptoms = Weight Loss || Night Sweats || Coughing, have there respective compartments imputed and combined: TB symptoms, Motor Palsy. These compartments are in turn imputed using a **multivariate probit regression**.

$$\begin{aligned} L_{Omega_{cs}} &\sim LKJCorrelationCholesky(4) \\ cs_{a_0} &\sim \mathrm{t}(\nu, 0, 1) \\ cs_a &\sim \mathrm{t}(\nu, 0, 1) \\ cs_z \sim \mathcal{N}(cs_{a_0} + cs_a^{(1)} * HIV + cs_a^{(2)} * TB\_Day, L_{Omega_{cs}}); \end{aligned}$$

- Clinical continuous variables, after transformed, are imputed using **linear regression** corrected for **HIV Status**: Age, TB Days, GCS.

$$\begin{aligned} age_{a_0} &\sim \mathrm{t}(\nu, 0, 1) \\ age_a &\sim \mathrm{t}(\nu, 0, 1) \\ age &\sim \mathcal{N}(age_{a_0} + age_a * HIV, age_\sigma) \end{aligned}$$

- Lab values are imputed toegether using **seemingly unrelated regression**: Blood Glucose, CSF Glucose, CSF Lymphocyte count, CSF Protein, CSF Lactate. Formulation similar to those of **multivariate probit regression**.
- GCS, integer, which is a linear sum of GCSV, GCSM, GCSE, have its only missing value imputed as a continuous variable.

$$\begin{aligned} GCSV_{a_0} &\sim \mathrm{t}(\nu, 0, 1) \\ GCSV_a &\sim \mathrm{t}(\nu, 0, 1) \\ GCSV_\sigma &\sim \mathcal{N}(0, 1) \\ GCSV \sim \mathcal{N}(GCSV_{a_0} * \begin{bmatrix} GCSE \\ GCSM \end{bmatrix}, GCSV_\sigma) \\ GCS &= GCSV + GCSM + GCSE \end{aligned}$$

2.3 Current Results

Below are current fit results for model **m2c** (model with Random Effects and GCS as continuous predictor), **m2cp** (GCS with quadratic effect), and **m2d** (GCS as dichotomous predictor, as used in the original definition score).

The estimation of sensitivities and specificities are similar for all three models.

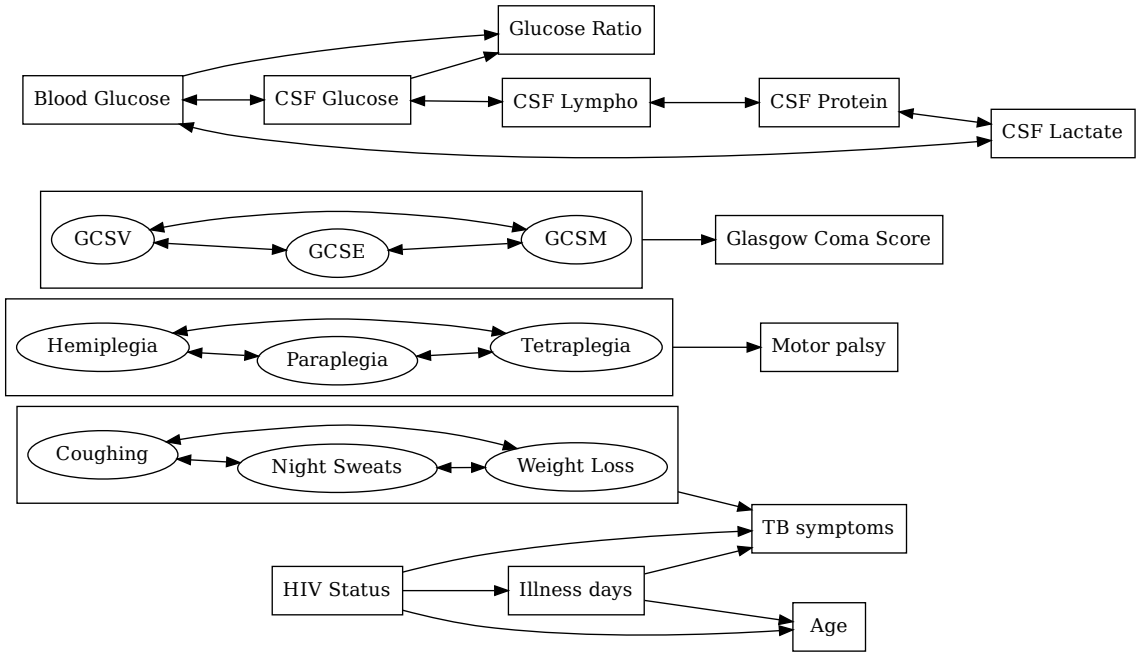

Figure 3: Imputation model

2.4 Problems

- Very complicated imputation model, especially with **discrete variables** ⇒ Prone to errors.
- Does illness days capture the effect from HIV? Should I include both?
- Problems with **Multivariate normal distribution** when doing partial prediction (i.e. imputation of partially missing combinations). **multi_normal_(cholesky)_rng** in Stan yields a whole vector by default ⇒ Has to do manually with conditional distribution. Sometimes, case by case.
- More (missing) problems with new data??
- Switch to another language? Pyro seems to be a good replacement?

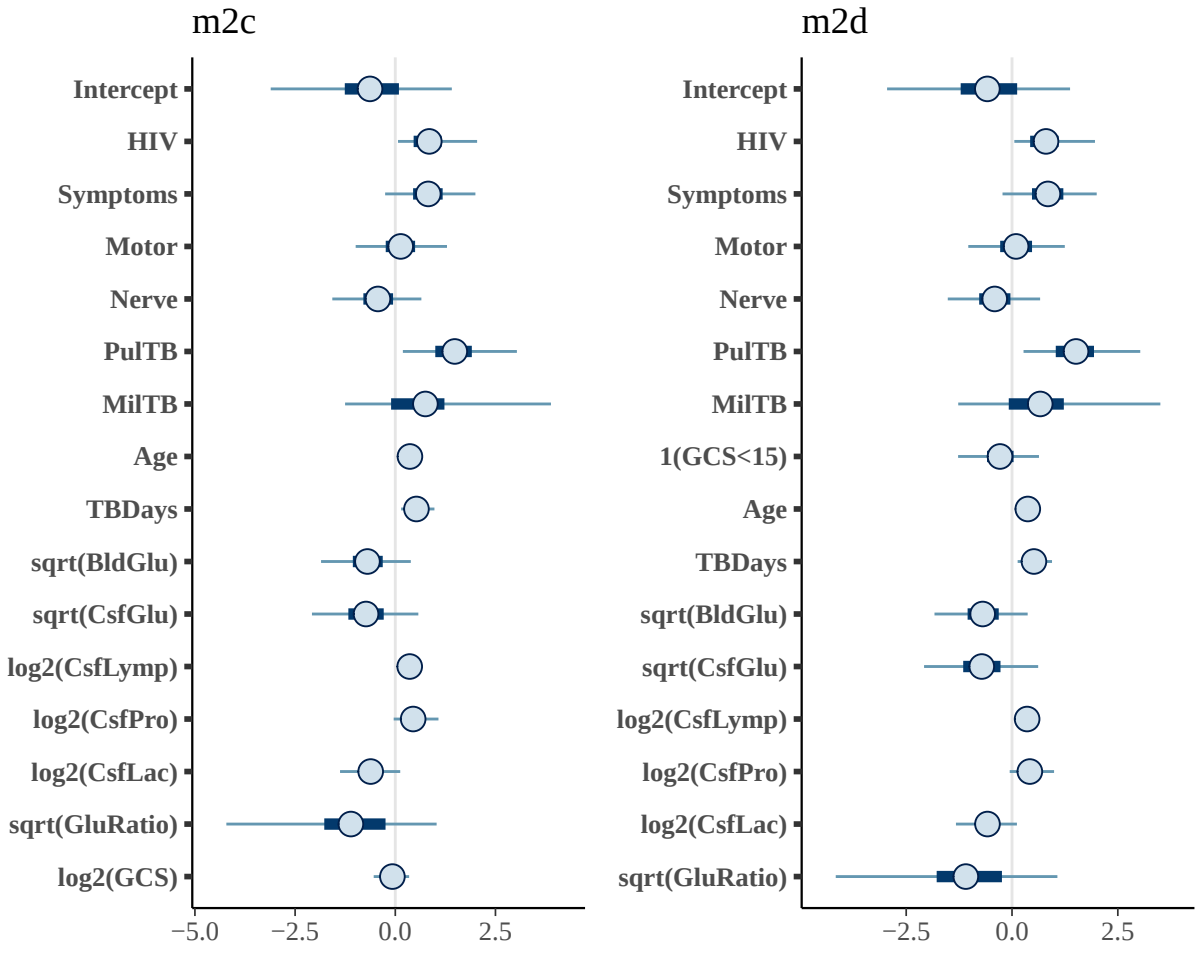

Figure 4: (#fig:plot_result_a)Coefs in m2c and m2d

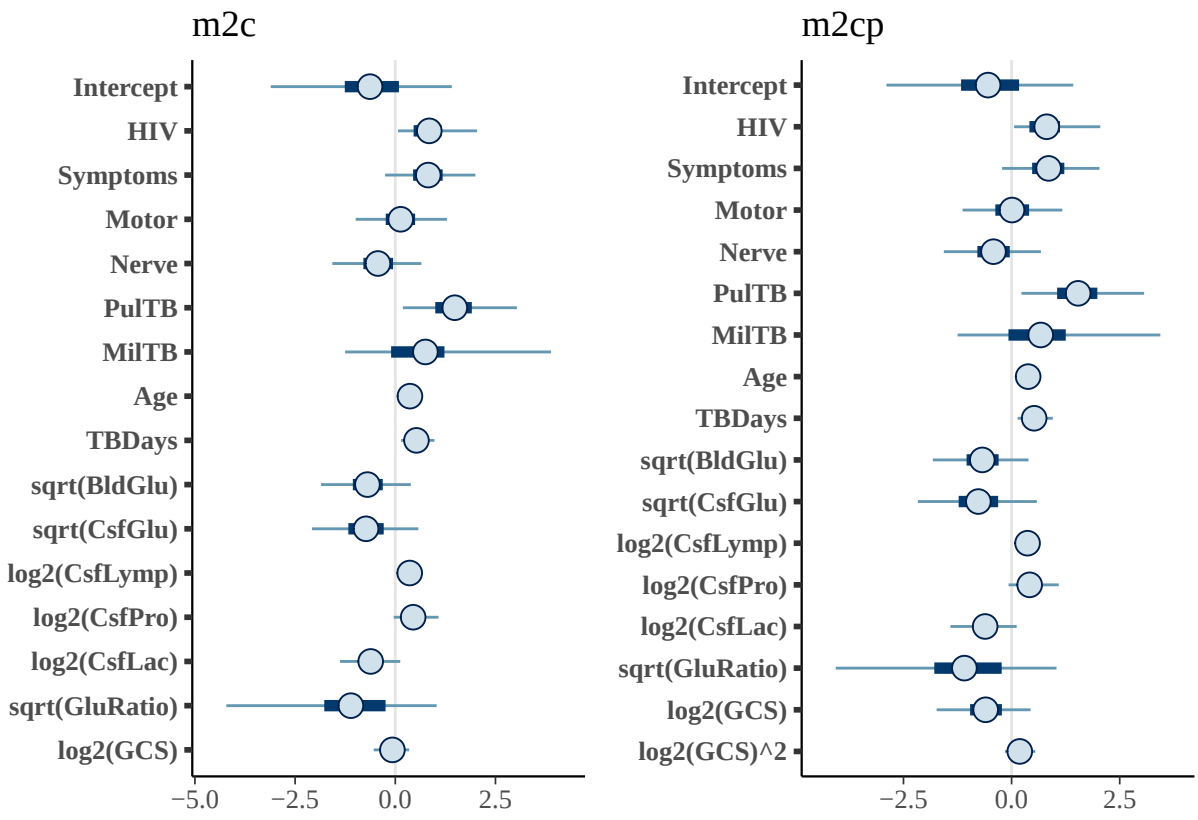

Figure 5: (#fig:plot_result_a2)Coefs in m2c and m2cp

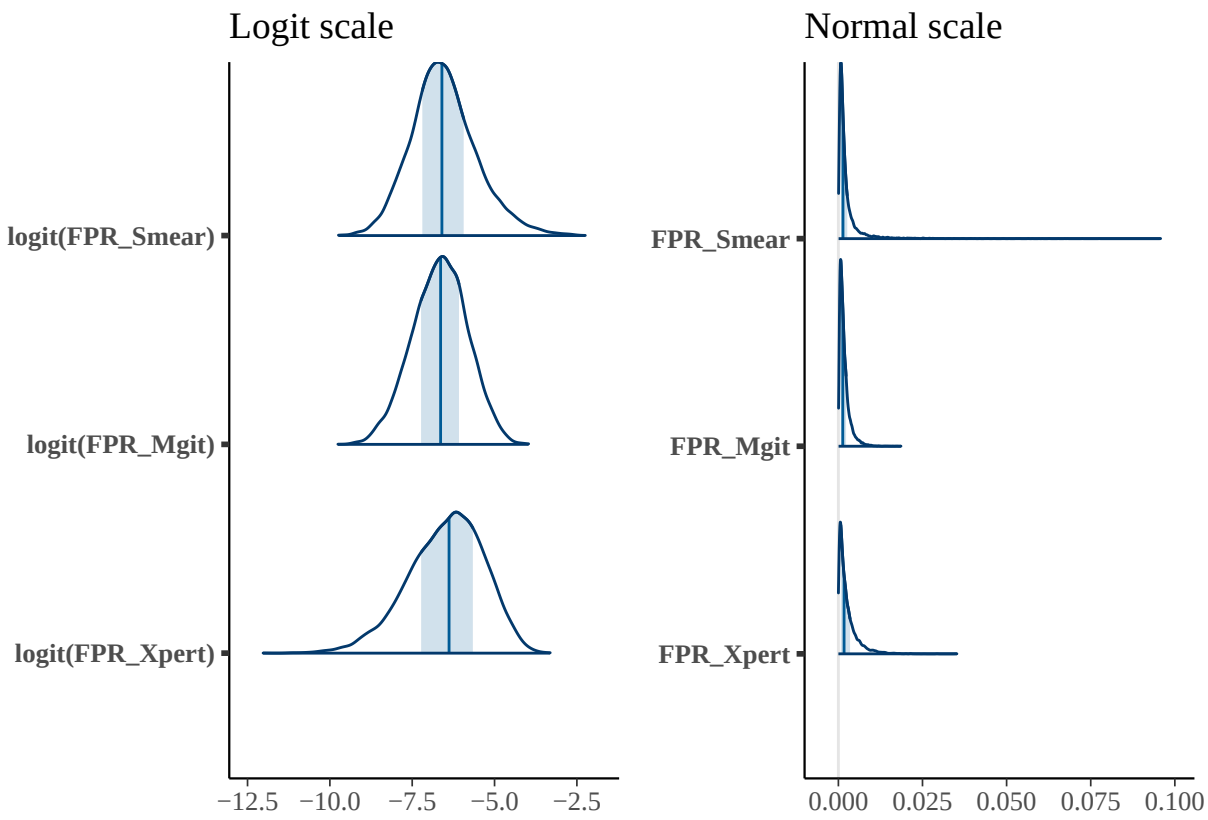

Figure 6: (#fig:plot_result_FPR)False positive rates

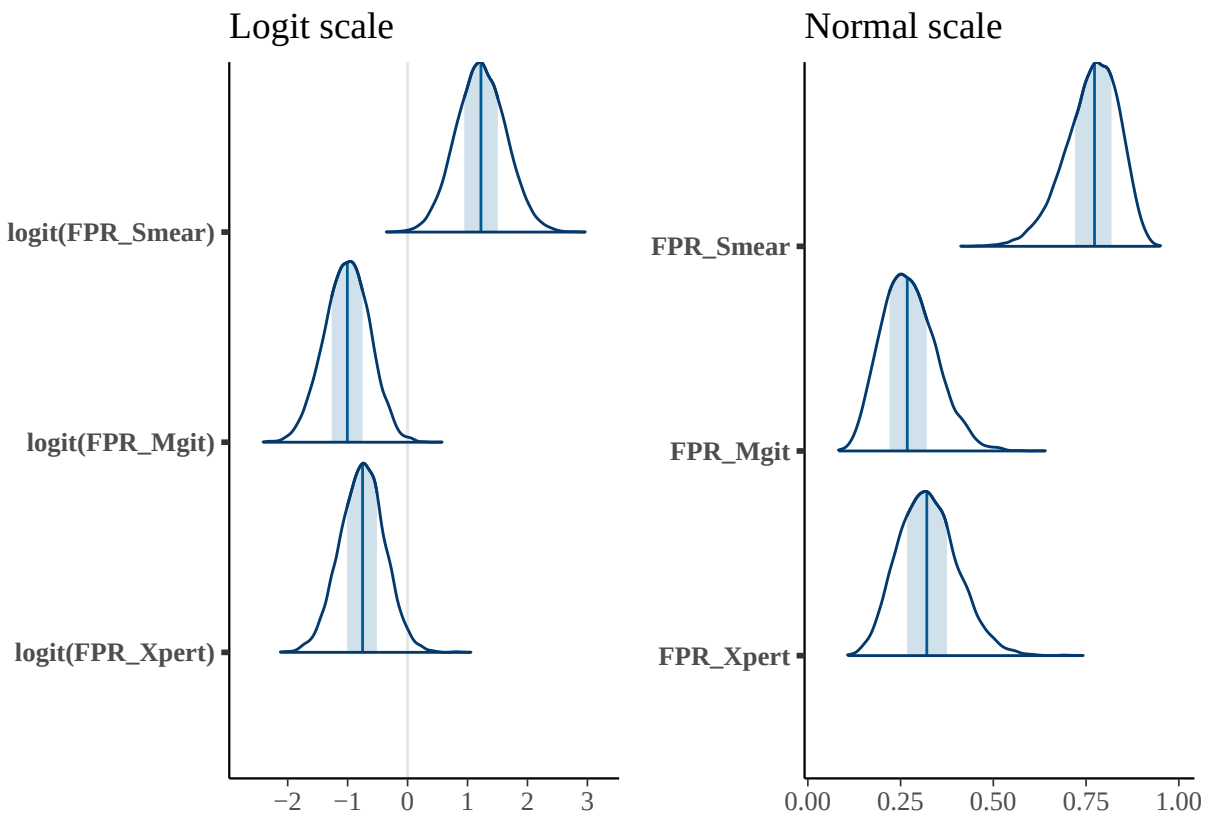

Figure 7: (#fig:plot_result_TPR)True positive rates

USUBJID	bld_glucose	csf_glucose	csf_lympho	csf_protein	csf_lactate
003-048		2.88	1	0.386	1.56
003-055	8.02	2.6		3.86	13.9
003-068	9.25	2.75		3.67	12.3
003-078	10.6	0.12		7.06	12.4
003-102					
003-167	4.04	2.44		2.81	14.6
003-288		5.18	19.9	1.12	2.7
003-311		3.49	1	0.504	2.02

Table 1: Subset of individuals with missing CSF lab results