Variational Bayes latent class approach for EHR-based phenotyping

Brian Buckley[1], Adrian O'Hagan[1,2] & Marie Galligan[1]

[1] School of Mathematics & Statistics, University College Dublin.
[2] The Insight Centre for Data Analytics, University College Dublin.

brian.buckley.1@ucdconnect.ie

INTRODUCTION

With growing acceptance by clinical regulators of the value of real-world evidence to supplement clinical trials, there is increasing interest in the use of Bayesian analysis for both experimental and observational clinical studies[1]. Bayesian statistics provides a formal mathematical method for combining prior information with current information at the design stage, during the conduct of a study and at the analysis stage. This interest has been limited by the computational challenges of applying the Markov-chain Monte-carlo (MCMC) approach to large real-world clinical data. The variational approach minimises the distance between a postulated family of standard distributions to approximate the posterior distribution rather than directly sampling from it as done by MCMC. This optimisation approach is often significantly more computationally efficient than MCMC.

We investigate the performance and characteristics of currently available R and Python Variational Bayes (VB) software for Bayes Latent Class Analysis (LCA) on an Electronic Health Records (EHR) dataset with mixed continuous and binary data. This work extends the MCMC Bayesian model described in Hubbard et al.[2]. The implementations include several algorithms for VB: coordinate ascent mean-field, stochastic, automatic differentiation and two application-specific R packages. For the baseline comparison we found implementations with predefined analytically derived objective functions are computationally efficient with best predictive performance and low programming complexity. However there are currently no closed-form solutions for real-world Bayes LCA using VB so we are focusing on improving posterior accuracy and computational run time from default settings for automatic VB methods.

Background

The LCA model is based on Hubbard et al.[2] and follows a general specification shown in table 1.

Table 1: Model specification for Bayesian latent variable model for EHR-derived phenotypes for the *i*th patient.

$g(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$

	Latent Phenotype	Availability of Biomarkers	Biomarkers	Clinical Codes	Prescription Medications
Example	Type 2 Diabetes	Availability of glucose or HbA1c data	Glucose or HbA1c values	Diabetes ICD-9 code; Endocrinologist visits	Diabetes medication
Variable	D_i	$R_{ij}, j = 1, \dots, J$	$Y_{ij}, j = 1, \dots, J$	$W_{ik}, k = 1, \dots, K$	$P_{il}, l = 1, \dots, L$
Model	$D_i \sim \text{Bern}(g(\boldsymbol{X}_i\boldsymbol{\beta}^D + \eta_i))$	$R_{ij} \sim \text{Bern}(g((1, \boldsymbol{X}_i, D_i)\boldsymbol{\beta}_j^R))$	$Y_{ij} \sim \text{N}((1, \boldsymbol{X}_i, D_i)\boldsymbol{\beta}_j^Y, \tau_j^2)$	$W_{ik} \sim \text{Bern}(g((1, \boldsymbol{X}_i, D_i)\boldsymbol{\beta}_k^W)$	$P_{il} \sim \text{Bern}(g((1, \boldsymbol{X}_i, D_i)\boldsymbol{\beta}_l^P))$
Priors	$\boldsymbol{\beta}^D \sim \text{MVN}(0, \Sigma_D)$ $\eta_i \sim \text{Unif}(a, b)$	$\boldsymbol{\beta}_j^R \sim \text{MVN}(\boldsymbol{\mu}_R, \Sigma_R)$	$\boldsymbol{\beta}_j^Y \sim \text{MVN}(\boldsymbol{\mu}_Y, \Sigma_Y)$ $\tau_j^2 \sim \text{InvGamma}(c, d)$	$\boldsymbol{\beta}_k^W \sim \text{MVN}(\boldsymbol{\mu}_W, \Sigma_W)$	$\boldsymbol{\beta}_l^P \sim \text{MVN}(\boldsymbol{\mu}_P, \Sigma_P)$

Abbreviations: N, normal; Bern, Bernoulli; MVN, multivariate normal; Unif, uniform; InvGamma, inverse gamma, HbA1c, Hemoglobin A1c.

MAIN OBJECTIVE

Benchmark current VB software against MCMC in estimating the MCMC posterior model, predictive performance and computational performance and complexity.

MATERIALS & METHODS

Baseline Comparison (Pima Indian data)

We analysed Pima Indian Type 2 Diabetes data [3]. The response, Y_i, is the variable *Outcome* (diagnosed type 2 diabetes). The predictors, X_i, are all continuous variables (*Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigree, Age*). A logistic regression model was fitted to all predictors. Two VB packages in R were compared (*sparsevb* and *varbvs*) and implementations of CAVI and SVI from the Github of Durante and Rigon [4]. We included *Stan MC* to compare more efficient Hamiltonian monte-carlo to Gibbs/Metropolis Hastings JAGS samping methods and the *Stan VB* R packages. The Python package *PyMC3* supports four VB methods that were included in this study (Table 1). We applied 5-fold cross validation to the dataset to investigate the stability of the models and for algorithms with hyperparameters we performed a grid search over a range for each hyperparameter.

Table 2: Brief description of VB algorithms studied. Automatic methods do not require analytical derivation of the ELBO objective function.

Algorithm	Description	Type	Automatic	Programming
CAVI	Coordinate Ascent Variational Inference	mean-field	No	R
Own VI	Own implementation of CAVI based on [5]	mean-field	No	R
SVI	Stochastic Variational Inference	mean-field	No	R
varbvs	Fast Variable Selection for Large-scale Regression	mean-field	No	R
sparsevb	Spike-and-Slab VB for Linear and Logistic Regression	mean-field	No	R
Stan MC	MCMC using No U-turn Hamiltonian monte-carlo	MCMC (to compare with JAGS)	No	R
Stan VB	Automatic Differentiation Variational Inference	mean-field	Yes	R
ADVI	Automatic Differentiation Variational Inference	mean-field	Yes	Python
FRADVI	Full-rank Automatic Differentiation Variational Inference	full-rank	Yes	Python
NFVI	Normalizing Flow Variational Inference	mean-field or full-rank	Yes	Python
ASVGD	Amortized Stein Variational Gradient Descent	operator	Yes	Python

Bayes LCA Model (Optum$^{\text{TM}}$ data)

We applied the Hubbard et al. model shown in Table 1 to JAGS MCMC, Stan MCMC and Stan ADVI VB. We could not use the same data as Hubbard et al. and instead used similar data from Optum$^{\text{TM}}$. We carefully selected Optum$^{\text{TM}}$ data that aligns closely with the characteristics of the Hubbard et al. data. The Optum$^{\text{TM}}$ data includes two continuous biomarker laboratory measures; random plasma glucose test and hemoglobin A1c (HbA1c) test. There are five indicator variables; visit to an endocrinologist, two diabetes medications (metformin and insulin) and diagnosis codes for type 1 and type 2 diabetes mellitus. In addition, there are three demographic variables; age at study baseline, BMI z-score and whether the patient is from a high risk ethnicity for diabetes mellitus. To compare with maximum likelihood LCA we compared with the R package *clustMD*. JAGS MCMC samples the posterior distribution of all variables. Stan MC integrates out the binary latent variables and Stan VB employs automatic differentiation variational inference. We translated the JAGS BUGS model to Stan to be as as close as possible. This could have resulted in a less than optimum model as Stan has several model reparamerterizations that could improve efficiency at the cost of very different model definitions. For details please see the **Reparamerterization** section in part 2 of the Stan Users Guide.

RESULTS

Baseline Comparison

Eleven VB methods were compared against MCMC. The mean and coefficient of variation of the model coefficients calculated over the 5 folds were compared with the MCMC baseline (Figure 1). The heatmap indicates some variables are more challenging across multiple VB methods. For example, *SkinThickness*, which is dominated by zero values.

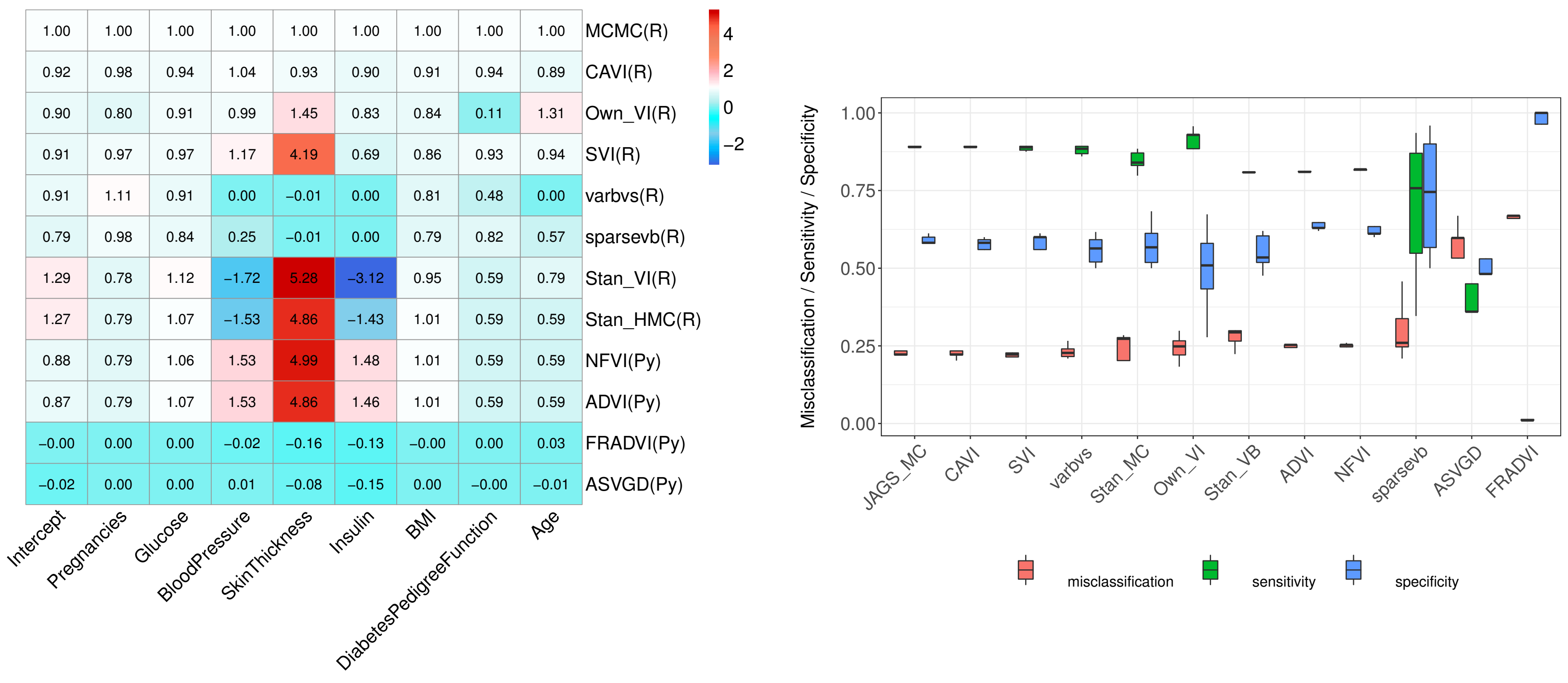

Figure 1: Coefficient mean as a proportion of MCMC model, empirical predictive performance. The programming environment is indicated by R for R programming and Py for Python programming

The predictive performances closest to MCMC were CAVI, SVI and varbvs. All three are mean-field methods that require analytical derivation of the optimization ELBO for logistic regression (and any other application) in contrast to automatic methods e.g. ADVI. Stan MC, and perhaps also Stan VB, might benefit from model reparameterization as described in part 2 of the Stan users guide.

Bayes LCA Model

Table 3 shows the comparative results for the LCA models using JAGS MCMC, Stan MC and Stan VB. Given the use of different real-world data sets for Hubbard et al. and our models the results are quite similar. This LCA model has very good general applicability. The Stan VB model has generally high accuracy versus MCMC methods apart from mean shift in glucose. Our LCA models agree with the Hubbard et al. conclusions that the latent phenotype approach may substantially improve on the standard clinical rule-based phenotyping approaches.

Table 3: Comparison of LCA model results with Hubbard et al.

	Posterior Mean (95% CI)			
	(a) Hubbard et al.	(b) JAGS MCMC	(c) Stan MC	(d) Stan VB
	N = 68,265	N = 16,580	N = 16,580	N = 16,580
Mean shift in glucose	90.62 (90.25, 91.00)	89.30 (89.10, 90.01)	88.59 (88.48, 88.71)	22.8 (21.06, 24.92)
Mean shift in HbA1c	3.15 (3.06, 3.24)	4.80 (4.72, 4.81)	4.77 (4.76, 4.78)	4.77 (4.75, 4.78)
T2DM code sensitivity	0.17 (0.15, 0.20)	0.15 (0.12, 0.18)	0.10 (0.09, 0.11)	0.12 (0.1, 0.12)
T2DM code specificity	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (0.99, 1.00)	0.99 (0.99, 0.99)
Endocrinologist visit code sensitivity	0.94 (0.92, 0.95)	0.18 (0.15, 0.21)	0.20 (0.18, 0.21)	0.22 (0.19, 0.22)
Endocrinologist visit code specificity	0.93 (0.93, 0.94)	0.99 (0.98, 0.99)	0.98 (0.97, 0.99)	0.97 (0.97, 0.99)
Metformin code sensitivity	0.31 (0.28, 0.35)	0.40 (0.36, 0.44)	0.21 (0.20, 0.21)	0.19 (0.19, 0.20)
Metformin code specificity	0.99 (0.99, 0.99)	0.98 (0.98, 0.99)	0.93 (0.92, 0.93)	0.93 (0.92, 0.94)
Insulin code sensitivity	0.66 (0.61, 0.70)	0.55 (0.51, 0.59)	0.35 (0.31, 0.35)	0.20 (0.19, 0.20)
Insulin code specificity	1.00 (1.00, 1.00)	1.00 (1.00, 1.00)	1.00 (0.99, 1.00)	1.00(0.99, 1.00)

To complete the LCA study, we used the R package *clustMD* to compare the Bayesian approach to a maximum likelihood (MLE) approach. The latent cluster means for glucose and HbA1c are close to the Bayesian LCA results. The MLE approach cannot be compared with the sensitivity and specificity as those require priors.

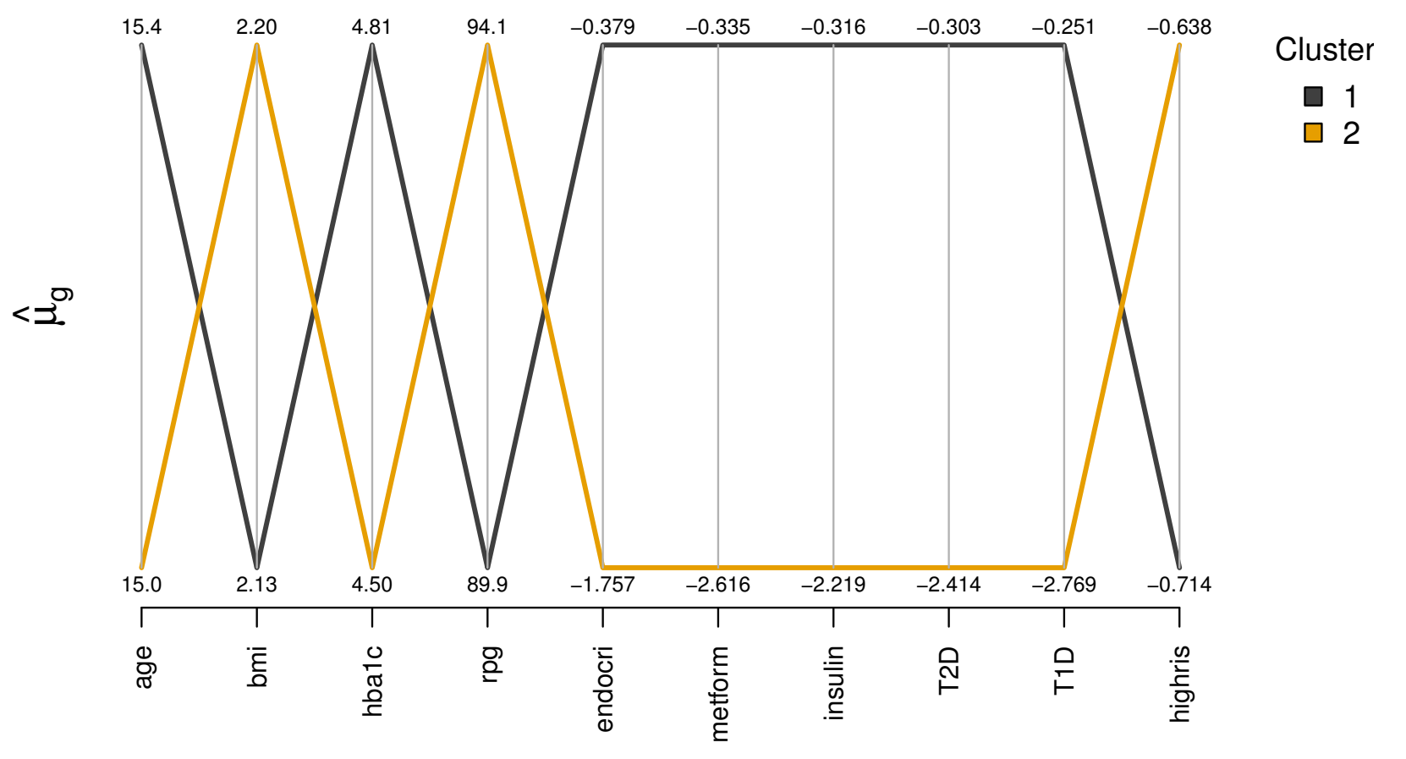

Figure 2: Cluster mean centres from the R package clustMD. The T2D latent variable is cluster 2

CONCLUSIONS

- In the baseline Pima Indian data, mean-field methods with analytical derivations for the ELBO have best predictive power and computational performance without complex hyperparameterization.
- More general methods such as ADVI are very sensitive to hyperparameter settings and might require many iterations to achieve ELBO optimization.
- A significant advantage of applying variational Bayes to LCA is overcoming the label switching problem that multi-chain MCMC are subject to.

FORTHCOMING RESEARCH

Comparison with maximum likelihood suggests some variables are not important to the LCA model. It might be useful to create a solution for variable selection in Bayesian LCA especially in the clinical observational context where there can be many variables as well as very large numbers of observations. Variable selection in Bayesian LCA is a potentially novel avenue for further research. The pros and cons of integrating out categorical latent variables rather than estimating their posterior distributions, as is the mechanism in Stan, might benefit from a detailed investigation for accuracy and flexibility given the prevalence of discrete variables in clinical data.

References

[1] Bruno Boulanger and Bradley P Carlin. How and why Bayesian Statistics are Revolutionizing Pharmaceutical Decision Making. *Clinical Researcher*, 703:20, 2021.

[2] Rebecca A Hubbard, Jing Huang, Joanna Harton, Arman Oganisian, Grace Choi, Levon Utidjian, Ihuoma Eneli, L Charles Bailey, and Yong Chen. A bayesian latent class approach for ehr-based phenotyping. *Statistics in medicine*, 38(1):74–87, 2019.

[3] Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association, 1988.

[4] Daniele Durante and Tommaso Rigon. Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science*, 34(3):472–485, 2019.

[5] Kevin P Murphy. *Machine learning: A Probabilistic Perspective*. MIT press, 2012.