Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Enhancing early prediction of heart failure: Hidden Markov model with covariates

Luca Brusa[1]
(luca.brusa@unimib.it)

Fulvia Pennoni[1]
Francesco Bartolucci[2]

[1]University of Milano-Bicocca, Department of Statistics and Quantitative Methods (Milan, Italy)

[2]University of Perugia, Department of Economics (Perugia, Italy)

Outline

Hidden Markov model: notation and formulation

- **Univariate categorical response variables** $Y_i^{(t)}$ with c categories, observed for unit i at time occasion t
- **Hidden process** $U_i^{(t)}$, following a first-order Markov chain with state-space $\{1, \ldots, k\}$
- **Covariates** $\boldsymbol{x}_i^{(t)}$, representing the vector of observed individual covariates for unit i at time t
- Composed of two sub-models:
  1. **Measurement model**: conditional distribution of the response variable $Y_i^{(t)}$ given the latent variable $U_i^{(t)}$ and the possible influence of individual time-varying and time-fixed covariates $\boldsymbol{x}_i^{(t)}$
  2. **Latent model**: non-parametric distribution of the latent process; accounts for the **unobserved heterogeneity** between individuals

Hidden Markov model: parameters

- **Initial probabilities**: $\pi_u = p(U_i^{(1)} = u)$
- **Transition probabilities**: $\pi_{u|\bar{u}} = p(U_i^{(t)} = u | U_i^{(t-1)} = \bar{u})$
- **Conditional response probabilities**:

$$\phi_{y|u\boldsymbol{x}}^{(t)} = p\left(Y_i^{(t)} = y | U_i^{(t)} = u, \boldsymbol{X}_i^{(t)} = \boldsymbol{x}\right),$$

based on the following **global logit parameterization**:

$$\log \frac{\phi_{y|u\boldsymbol{x}}^{(t)} + \cdots + \phi_{c-1|u\boldsymbol{x}}^{(t)}}{\phi_{0|u\boldsymbol{x}}^{(t)} + \cdots + \phi_{y-1|u\boldsymbol{x}}^{(t)}} = \mu_y + \alpha_u + \boldsymbol{x}'\boldsymbol{\beta},$$

with:

- $\mu_1, \ldots, \mu_{c-1}$: intercepts specific for each response category
- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$: support points of the latent variables
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$: regression parameters for the covariates

Maximum likelihood estimation

- The **expectation-maximization** (EM) algorithm is employed to perform **maximum likelihood estimation**
- It maximizes the observed-data log-likelihood function $\ell(\boldsymbol{\theta})$ relying on the **complete-data log-likelihood** function $\ell^*(\boldsymbol{\theta})$
- It alternates the following steps until convergence:
  - **E-step**: **compute the conditional expected value** of $\ell^*(\boldsymbol{\theta})$ given the value of the parameters at the previous step and the observed data
  - **M-step**: **update the model parameters** by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$:
    - explicit solutions are available for π_u and $\pi_{u|\bar{u}}$
    - a Newton-Raphson algorithm is used for updating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

Outline

Forecasting

- **Forecasting probability** of category y for unit i at a future time occasion t^*:

$$\hat{p}_{iy}^{(t^*)} = \sum_{u=1}^{k} \hat{q}^{(t^*)}(u|\boldsymbol{x}_i, \boldsymbol{y}_i)\ \hat{\phi}_{y_i|u\boldsymbol{x}_i}^{(t^*)}$$

  - $\hat{\phi}_{y|u\boldsymbol{x}_i}^{(t^*)}$: estimated conditional response probabilities at time t^*
  - $\hat{q}^{(t^*)}(u|\boldsymbol{x}_i, \boldsymbol{y}_i)$: estimated posterior distribution of the latent variable $U_i^{(t^*)}$ given the observed responses $\boldsymbol{y}_i$ and covariates $\boldsymbol{x}_i$

- Event y is then forecast if $\hat{p}_{iy}^{(t^*)} > c$, with $c \in [0, 1)$ a suitable **cutoff** chosen according to:
  - either the **Yuoden's J statistics**
  - or the **F1 score**

Choice of the Cutoff

- Selection of c through a **k-fold cross-validation** procedure:
  - Data are split into training ($k-1$ folds) and validation (1 fold) sets
  - The model is estimated on the training folds and predict data of the test fold
  - Performance is evaluated using both the **F1 score** and the **Youden's J statistics**, and the (fold-specific) cutoffs are selected by maximizing these quantities
  - The final (overall) cutoffs are obtained by averaging the fold-specific values
- This strategy ensures that the cutoff is **data-driven**, accounts for **class imbalance**, and **avoids overfitting**
- The procedure is implemented for the case of a categorical response through suitable R functions

Outline

Heart failure

- **Heart failure** (**HF**) is a complex clinical syndrome in which the heart is unable to pump blood at a rate sufficient to meet the body's needs, or only at the cost of elevated filling pressures
- It is influenced by multiple heterogeneous and interacting **factors**:
  - **Non-modifiable**, like age, sex, genetic predispositions
  - **Modifiable**, like obesity, diabetes, hypertension, smoking, sedentary lifestyle, excessive alcohol consumption
- **Global burden**:
  - Leading cause of morbidity and mortality, affecting $\sim$55.5 million people worldwide (2021)
  - Associated with high hospital readmission rates and substantial healthcare costs (estimated at \$108 billion/year in 2012)
- **Early stages** of HF can be **asymptomatic**; timely identification of cardiac changes is crucial to prevent progression and reduce mortality

The data

- We consider historical data based on a retrospective study of **2008 patients** hospitalized with HF at Zigong Fourth People's Hospital (Sichuan, China), observed between **December 2016 and June 2019**

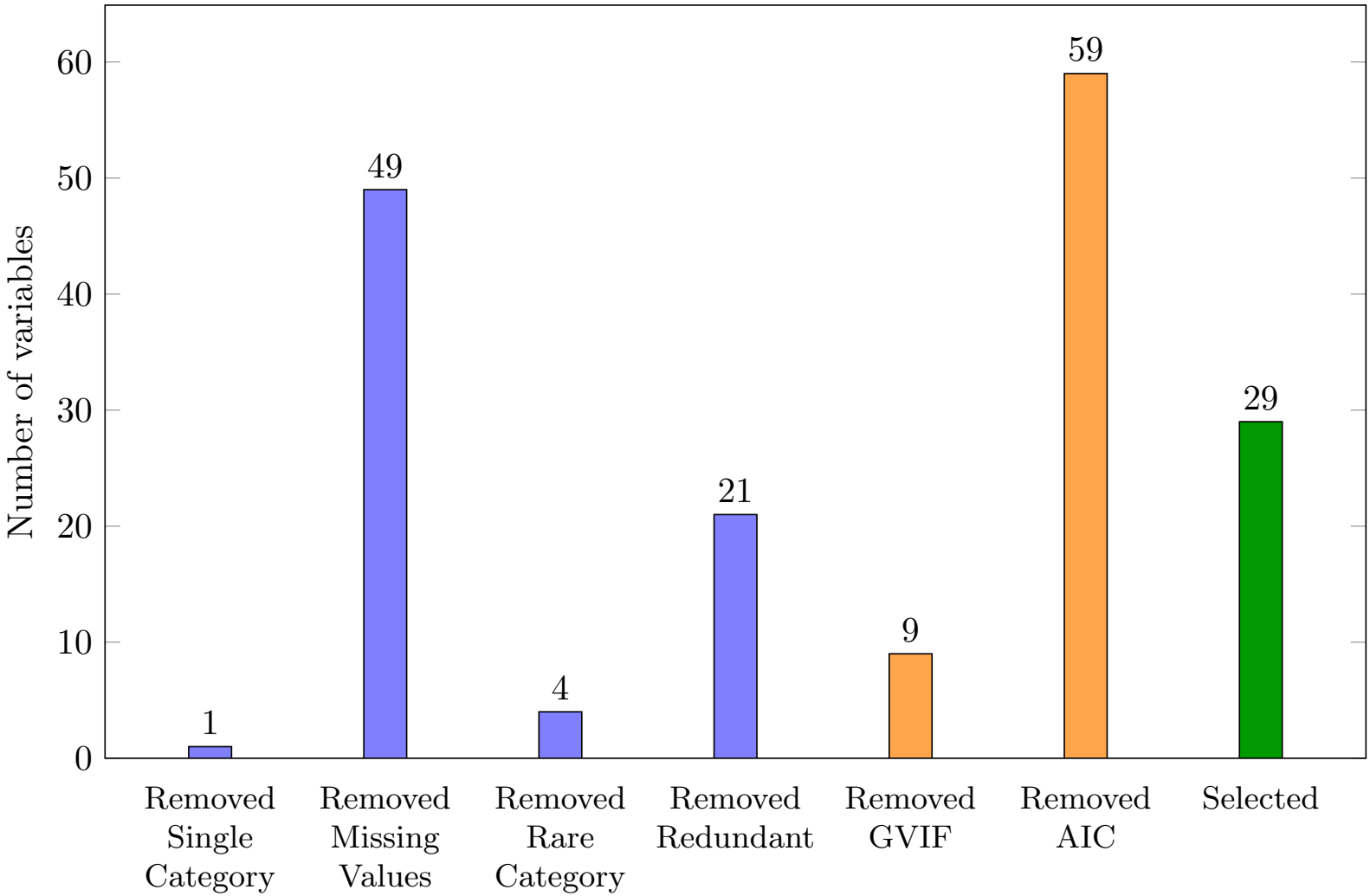
- **Outcomes**: hospital readmission or death, measured at three follow-up occasions (28 days, 3 months, 6 months after discharge)

- A wide set of patient-level information (172 variables), measured at baseline:
  - Demographic information
  - Clinical characteristics and comorbidities
  - Laboratory tests and treatments
  - Therapy data for 18 drugs (diuretics, inotropes, vasodilators, others)

Data preparation and variable selection

- **Some variables are eliminated**:
  - Variables assuming only one value in all the sample (non-informative)
  - Variables with more than 10% missing values
  - Binary variables with one category occurring in less than 0.5% of cases
  - Redundant variables (e.g., weight and height removed, keeping BMI)
- Missing data (49 variables; on average 2.7% of missing values) imputation performed using the **Spectral Regularization Algorithm**
- Removal of variables with excessively high Variance Inflation Factor (VIF) to assess **multicollinearity**
- Final selection of variables using the **Akaike Information Criterion** (AIC) to balance model fit and complexity

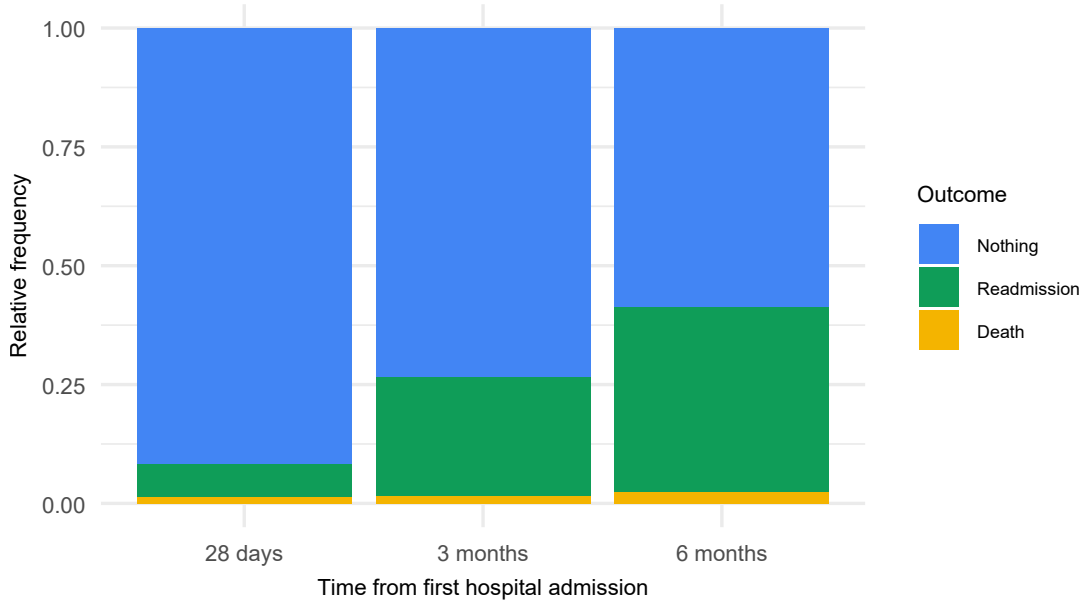
Number of removed and selected variables



Response variable

- **Categorical response variable** Y_{it}, $t = 1, \ldots, 3$, with **three categories**:

$$Y_i^{(t)} = \begin{cases} 2 & \text{if patient } i \text{ died at time } t \\ 1 & \text{if patient } i \text{ was readmitted at time } t \\ 0 & \text{otherwise} \end{cases}$$



Outline

HM model estimation

- After data cleaning, the sample includes **1,981 patients** with heart disease, observed over **3 time occasions**
- The HM model with covariates is estimated with a number of latent states k ranging from 1 to 5
- For each k, the model is estimated $20 \cdot k$ **times**, using both deterministic and random initializations to reduce the risk of local maxima
- The optimal number of states is selected using the **AIC**, which identifies **4 latent states** (AIC index = 5,682.662; number of free parameters = 52)
- Latent states may be interpreted as **patient risk profiles** according to the estimated support points

Estimated model parameters

State	$\hat{\alpha}_u$	$\hat{\pi}_u$
1	-5.265	0.728
2	12.750	0.086
3	13.597	0.172
4	27.233	0.015

Table: Estimated initial probabilities ($\hat{\pi}_u$) and support points ($\hat{\alpha}_u$)

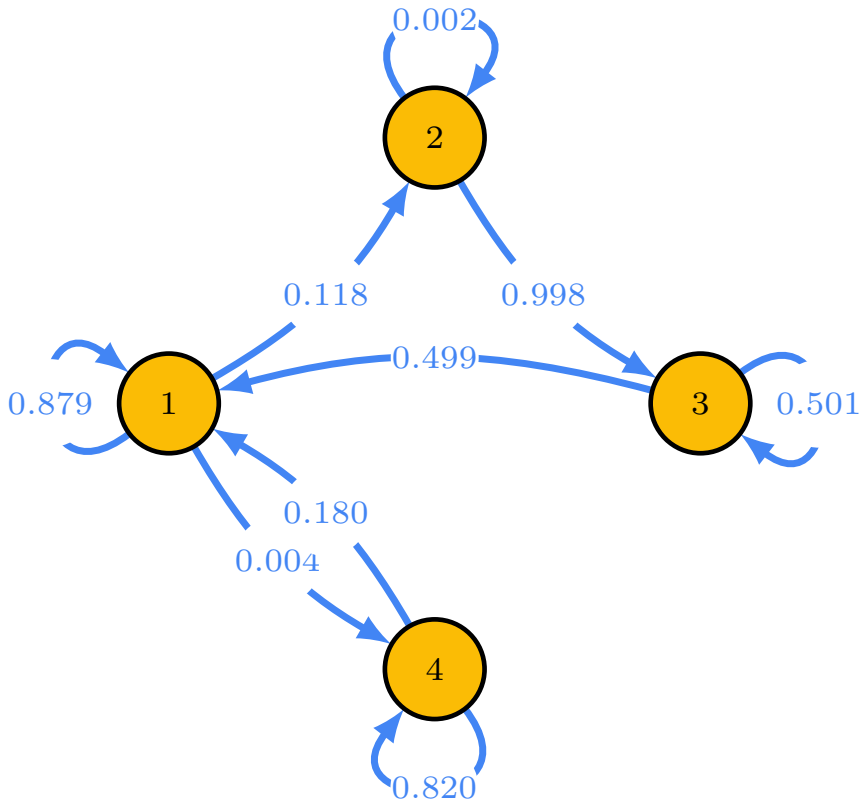


Figure: Estimated transition probabilities ($\hat{\pi}_{u|\bar{u}}$)

Estimated regression coefficients

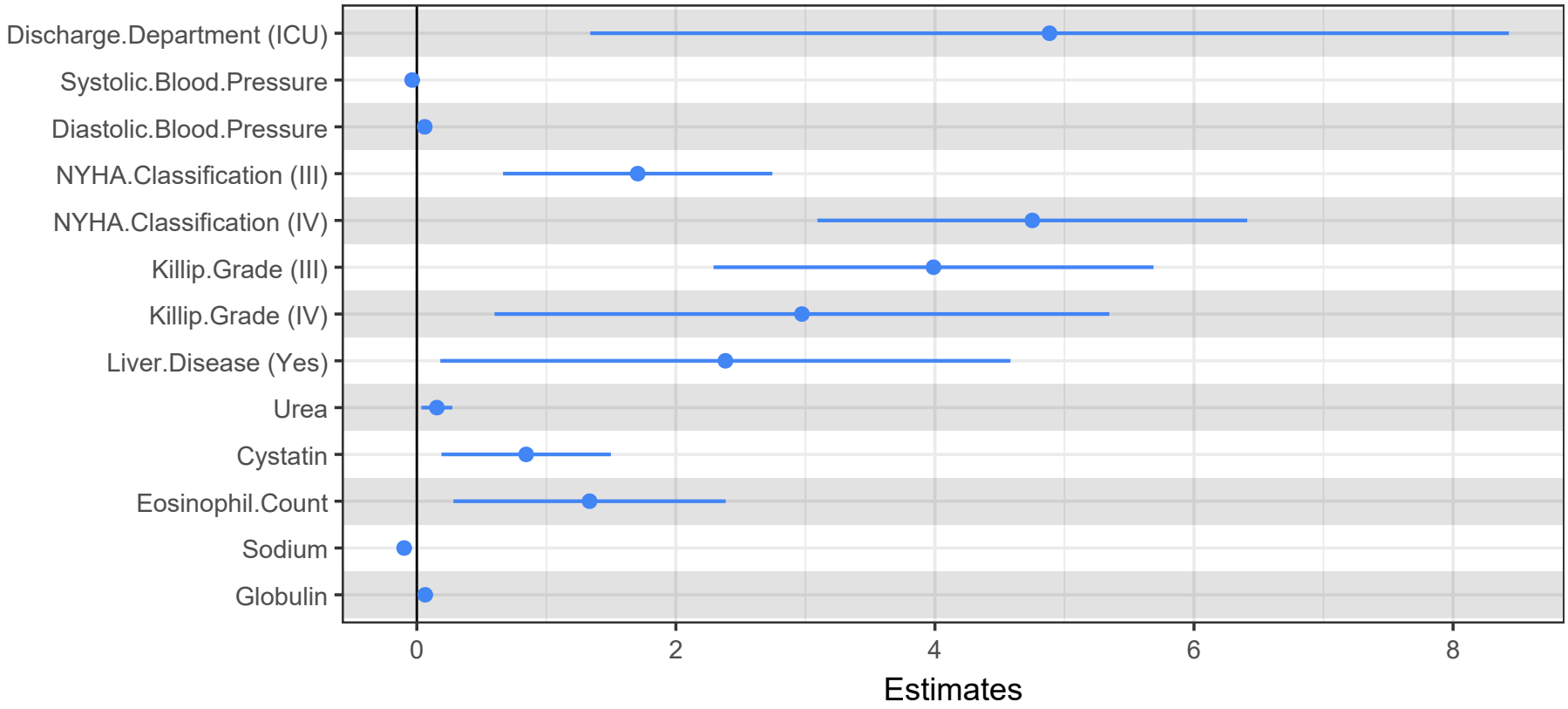
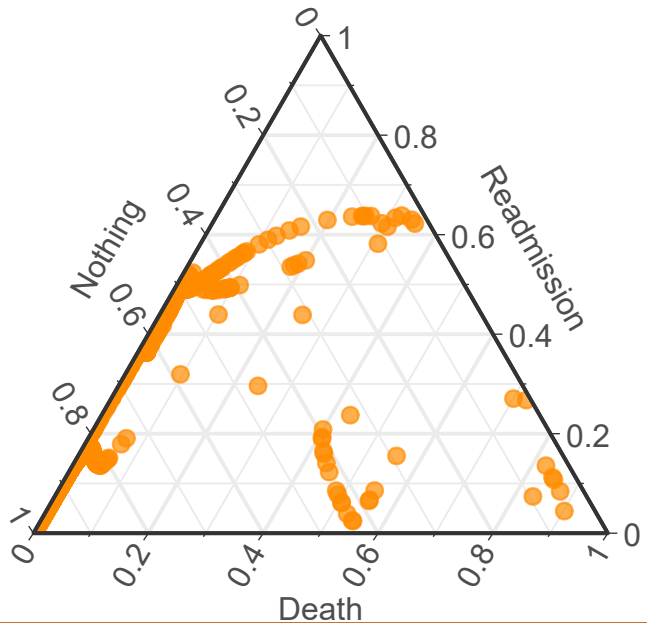


Figure: Point estimate ($\hat{\beta}$, interpretable in terms of log-odds) and corresponding confidence interval (confidence level: 95%) for the most relevant covariates

Forecasting

- The HM model (estimated using $k = 4$ latent states) is trained on data from **the first two time occasions** and used to forecast responses at the **last time occasion**
- For each patient a vector of forecasting probabilities is defined as:

$$\hat{\boldsymbol{p}}_i^{(t^*)} = (\hat{p}_{\text{Nothing}}^{(t^*)},\ \hat{p}_{\text{Readmission}}^{(t^*)},\ \hat{p}_{\text{Death}}^{(t^*)})$$



Forecasting results - Evaluation metrics

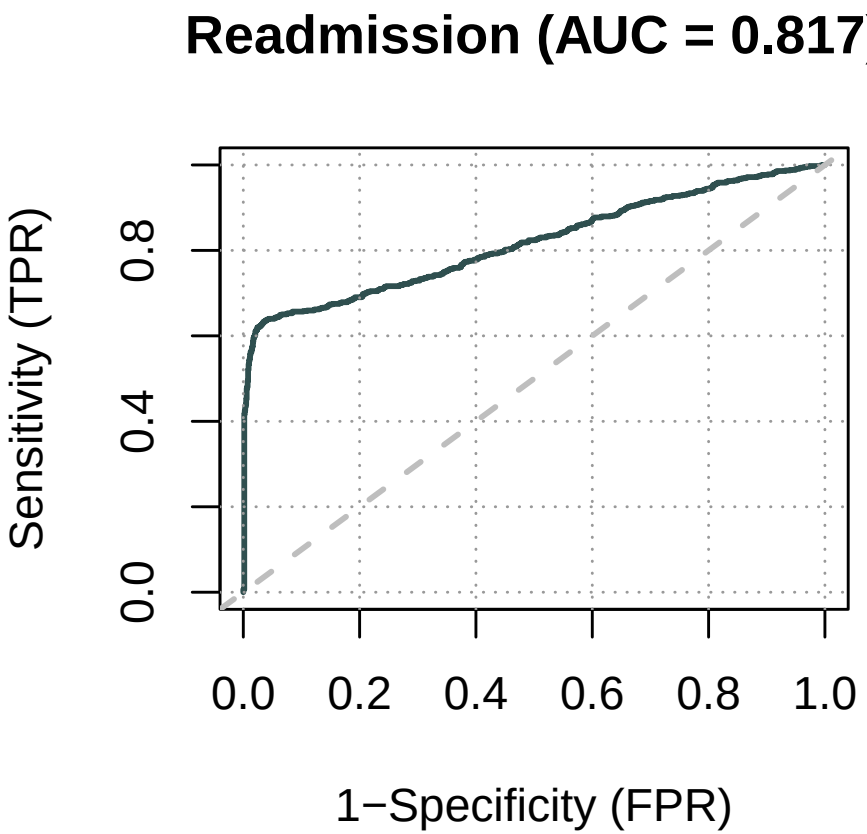
Cutoff selected through **F1 score**

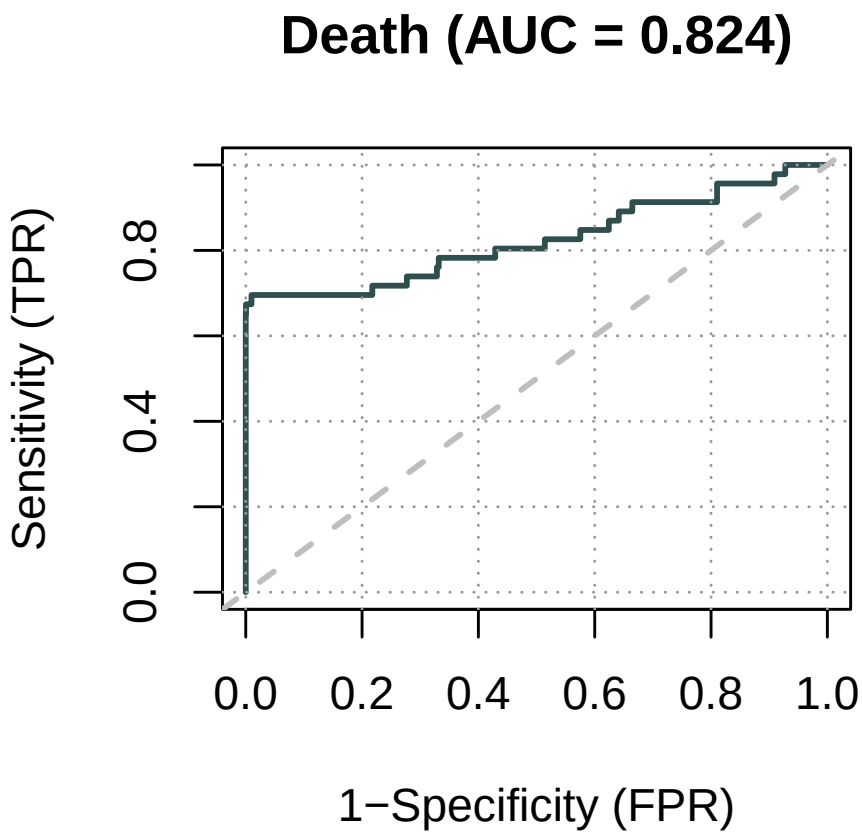
	Readmission	Death
Cutoff (c)	0.231	0.518
Specificity	0.983	1.000
Sensitivity/Recall	0.587	0.500
Precision	0.958	1.000
Accuracy	0.829	0.988
F1 score	0.728	0.667

Cutoff selected through **Youden's J statistics**

	Readmission	Death
Cutoff (c)	0.163	0.448
Specificity	0.942	1.000
Sensitivity/Recall	0.614	0.630
Precision	0.942	1.000
Accuracy	0.835	0.991
F1 score	0.744	0.773

Forecasting results - ROC curve





Outline

Conclusions

- **Main contributions**:
  - Use of a **HM model** for categorical variables with more than two levels, with covariates in the measurement model for predictive purposes
  - Data-driven **selection of the forecasting cutoff** through k-fold cross-validation, explicitly accounting for class imbalance
- **Future work**:
  - Exploration of alternative strategies for cutoff selection, such as **cost-sensitive approaches**
  - Simulation study to **measure the performance** and robustness of the proposed method
  - Extension to data with a **longer follow-up** and **time-varying covariates**
  - Inclusion of covariates with more than 10% of missing values, through the addition of dummy variables serving as missing indicators
  - Inclusion of drop-out and investigation of possible violation of the local independence assumption

Outline

References

Bartolucci, F. and Farcomeni, A. (2009). A multivariate extension of the dynamic logit model for longitudinal data based on a latent markov heterogeneity structure. *J. Am. Stat. Assoc.*, **104**, 816–831.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013). *Latent Markov models for longitudinal data*. Chapman & Hall/CRC, Boca Raton.

Braunwald, E., Zipes, D. P., and Libby, P. (2001). *Heart disease: a textbook of cardiovascular medicine (6th ed.)*. Saunders.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. Series B Stat. Methodol.*, **39**, 1–38.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, **11**, 2287–2322.

Welch, L. (2003). Hidden Markov models and the Baum-Welch algorithm. *IEEE Inform. Theory Soc. Newsl.*, **50**, 10–13.

Zhang, Z., Cao, L., Chen, R., Zhao, Y., Lv, L., Xu, Z., and Xu, P. (2021). Electronic healthcare records and external outcome data for hospitalized patients with heart failure. *Scientific Data*, **8**, 1–6.

Acknowledgments

Acknowledgment: L. Brusa, F. Pennoni, and F. Bartolucci acknowledge the financial support from the grant "Hidden Markov Models for Early Warning Systems" of Ministero dell'Università e della Ricerca (PRIN 2022TZEXKF) funded by European Union - Next Generation EU, Mission 4, Component 2, CUP J53D23004990006.