Generative Models for Screening Genetic Mutations in Marfan Syndrome

Antonin Della Noce and Vesna Lukic

Journées de Biostatistique, 22 November 2024

Collaborators

Nadine Hanna, Bichat

Hakim Benkirane, CS

Paul-Henry Cournède, CS

Pauline Arnaud, Bichat

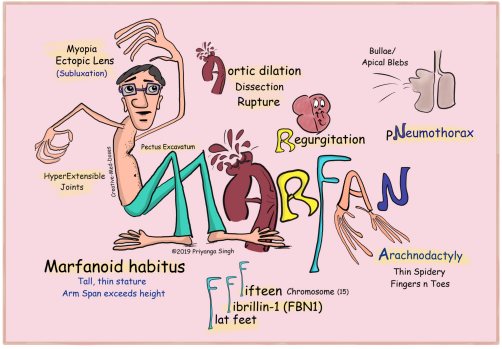
Olivier Milleron, Bichat

Guillaume Jondeau, Bichat

Marfan and Related Syndromes. Cardiovascular Risks

Main genetic tissue disorders

- Marfan syndrome
  (prevalence $\approx$ 1 in 10,000)
- Loeys-Dietz syndrome
- Ehlers-Danlos syndrome

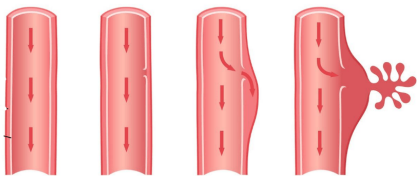


Creative-Med-Doses

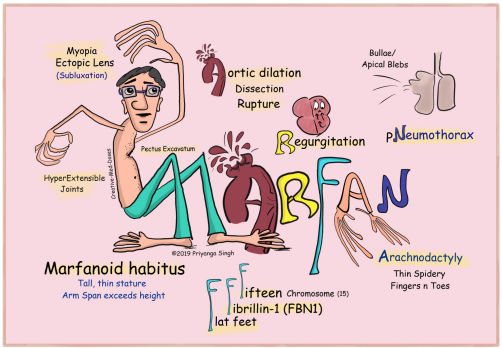
Marfan and Related Syndromes. Cardiovascular Risks

Main genetic tissue disorders

- Marfan syndrome
  (prevalence $\approx$ 1 in 10,000)
- Loeys-Dietz syndrome
- Ehlers-Danlos syndrome



Aortic dissection (*Center for vascular awareness*)



Creative-Med-Doses

Typical journey of a patient in France before being diagnosed with Marfan

Prophylactic treatment for aortic dissection

There are different levels of risk for aortic dissection according to the mutation and the phenotype of the patient.

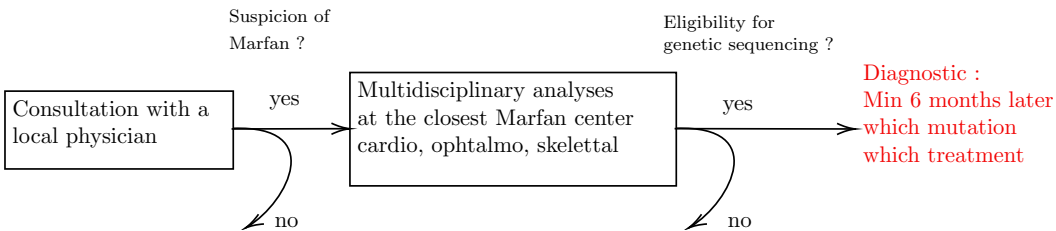
- **for high-risk patients**: surgery, elective aortic root replacement
- **for moderate-risk patients**: pharmacological treatment, β-blockers.

Typical journey of a patient in France before being diagnosed with Marfan

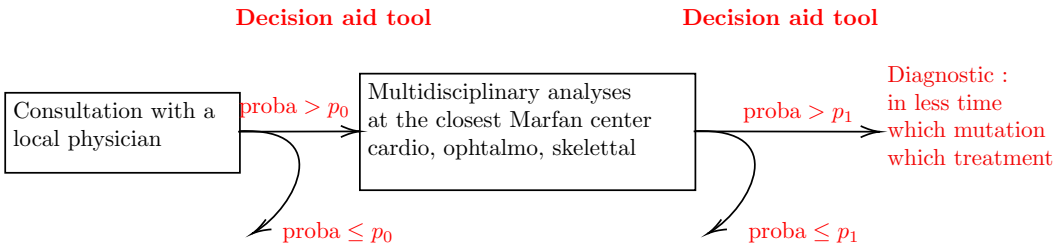
Prophylactic treatment for aortic dissection

There are different levels of risk for aortic dissection according to the mutation and the phenotype of the patient.

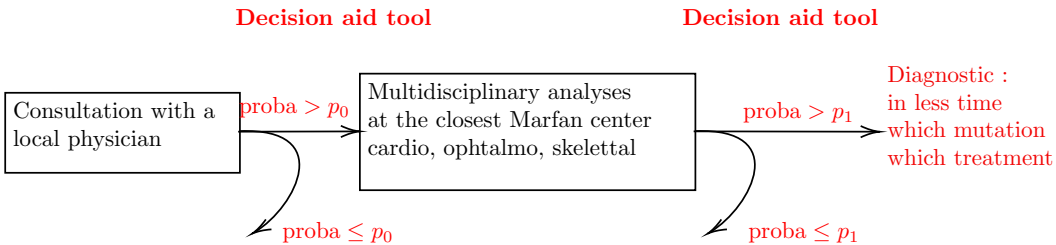
- **for high-risk patients**: surgery, elective aortic root replacement
- **for moderate-risk patients**: pharmacological treatment, β-blockers.



Decision aid tool for screening patients



Decision aid tool for screening patients

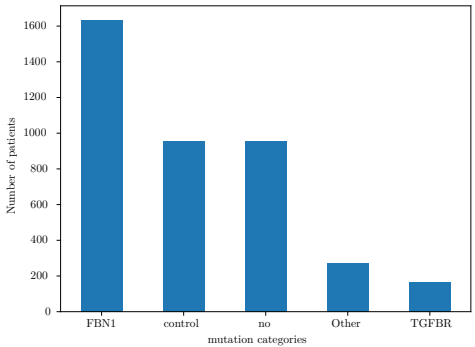


Objective

Design a decision aid tool for the two steps of the patient pathway:

- **outside the specialized centers**: help local physicians decide whether to refer a patient to a reference center
- **within the specialized centers**: personalize the set of tests based on a patient's characteristics

Data description



Cohort description (Bichat hopsital, Paris)

- 3,982 patients
- 19 clinical variables
  - age, sex
  - morphological features
  - cardiovascular features
  - ophthalmic features
- 5 genetic categories: *FBN1, TGFBR, Other, No, Control.*

Data type and missingness

- 20 columns in total: 8 continuous, 12 categorical;
- high missing rate: only 17% of the lines are complete.

From a classifier to a generative model

A first approach: screening as a classification problem

Learn the conditional distribution $\mathbb{P}(\text{category} \mid x_S)$ for all x_S, where S is a subset of the observed variables.

From a classifier to a generative model

A first approach: screening as a classification problem

Learn the conditional distribution $\mathbb{P}(\text{category} \mid x_S)$ for all x_S, where S is a subset of the observed variables.

Problem: $\hat{\mathbb{P}}(\text{category} = \textit{FBN1} \mid \text{age} = 30, \text{height} = 1m80) = 0.55$!!

From a classifier to a generative model

A first approach: screening as a classification problem

Learn the conditional distribution $\mathbb{P}(\text{category} \mid x_S)$ for all x_S, where S is a subset of the observed variables.
Problem: $\hat{\mathbb{P}}(\text{category} = \textit{FBN1} \mid \text{age} = 30, \text{height} = 1m80) = 0.55$!!

Solution: Prior distribution $\mathbb{P}$ on the mutation categories

$$\text{category} \sim 0.9999\ \delta_{\text{control}} + \varepsilon_1\ \delta_{\text{FBN1}} + \varepsilon_2\ \delta_{\text{no}} + \varepsilon_3\ \delta_{\text{other}} + \varepsilon_4\ \delta_{\text{TGFBR}}$$

From a classifier to a generative model

A first approach: screening as a classification problem

Learn the conditional distribution $\mathbb{P}(\text{category} \mid x_S)$ for all x_S, where S is a subset of the observed variables.
Problem: $\hat{\mathbb{P}}(\text{category} = \textit{FBN1} \mid \text{age} = 30, \text{height} = 1m80) = 0.55$!!

Solution: Prior distribution $\mathbb{P}$ on the mutation categories

$$\text{category} \sim 0.9999\ \delta_{\text{control}} + \varepsilon_1\ \delta_{\text{FBN1}} + \varepsilon_2\ \delta_{\text{no}} + \varepsilon_3\ \delta_{\text{other}} + \varepsilon_4\ \delta_{\text{TGFBR}}$$

Unknown phenotype distribution p

The phenotype is represented by a vector $x \in \mathcal{X} \subset \mathbb{R}^d$.

$$x \mid \text{age}, \text{sex}, \text{category} \sim p(x \mid \text{age}, \text{sex}, \text{category})\lambda(\mathrm{d}x)$$

where λ is a measure on $\mathcal{X}$.

From a classifier to a generative model

A classifier given by Bayes' theorem

$$\mathbb{P}(\text{cat} = c \mid x, \text{age}, \text{sex}) = \frac{p(x \mid \text{age}, \text{sex}, \text{cat} = \text{c})\mathbb{P}(\text{cat} = \text{c})}{\sum_{c'} p(x \mid \text{age}, \text{sex}, \text{cat} = \text{c}')\mathbb{P}(\text{cat} = \text{c}')}$$

From a classifier to a generative model

A classifier given by Bayes' theorem

$$\mathbb{P}(\text{cat} = c \mid x, \text{age}, \text{sex}) = \frac{p(x \mid \text{age}, \text{sex}, \text{cat} = \text{c})\mathbb{P}(\text{cat} = \text{c})}{\sum_{c'} p(x \mid \text{age}, \text{sex}, \text{cat} = \text{c}')\mathbb{P}(\text{cat} = \text{c}')}$$

Requirements for the generative model

- handle tabular data (continuous and categorical variables)
- deal with high missing rate in the data
- be able to evaluate the probability of a given phenotype, and all the conditional distribution probabilities

From a classifier to a generative model

A classifier given by Bayes' theorem

$$\mathbb{P}(\text{cat} = c \mid x, \text{age}, \text{sex}) = \frac{p(x \mid \text{age}, \text{sex}, \text{cat} = \text{c})\mathbb{P}(\text{cat} = \text{c})}{\sum_{c'} p(x \mid \text{age}, \text{sex}, \text{cat} = \text{c}')\mathbb{P}(\text{cat} = \text{c}')}$$

Requirements for the generative model

- handle tabular data (continuous and categorical variables)
- deal with high missing rate in the data
- be able to evaluate the probability of a given phenotype, and all the conditional distribution probabilities

Two generative paradigms explored

- Conditional Variational Auto-Encoder (CVAE) by Vesna
- Chained Equations probabilistic neural networks (inspired from MICE) by Antonin

Reconstruction performances and examination recommender

Reconstruction R^2 of MICE generative model

Variables	R^2
Span	0.99
Size	0.95
Weight	0.73
Ascending aorta	0.98
Sino-tubular	0.99
Valsalva sinus	0.77
Aortic arch	0.95
Annulus	0.43
Thumb sign	0.56
Wrist sign	0.54
Ectopia	0.18
Bifid uvula	0.15
Ogival palate	0.22
Pectus	0.16
Elbow extension	0.07

Definition of the reconstruction error

$$R^2 = 1 - \frac{\text{error of the model}}{\text{baseline error}}$$

Recommendation of the next exam by entropy minimization

Given partial observations x_S of a patient ($S \subsetneq [\![1, d]\!]$), the next exam to be conducted is the one minimizing the conditional entropy:

$$\ell^* = \operatorname*{argmin}_{\ell \notin S} \mathbb{E}\left[\text{ent}(\text{cat} \mid x_S, x_\ell, x_c)\right]$$

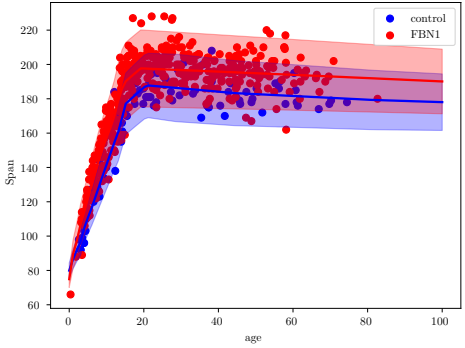
$$\text{where } \text{ent}(\text{cat} \mid x_S, x_\ell, x_c) = -\sum_c \mathbb{P}(\text{cat} = c \mid x_S, x_\ell, x_c) \log \mathbb{P}(\text{cat} = c \mid x_S, x_\ell, x_c)$$

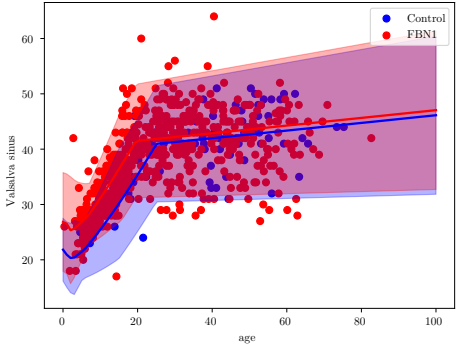
Recommendation of the next exam by entropy minimization

Example

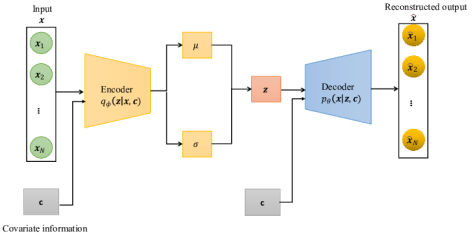
$$\mathbb{E}\left[\text{ent}(\text{cat} \mid \text{age} = 30, \text{sex} = \text{M}, \text{span})\right] < \mathbb{E}[\text{ent}(\text{cat} \mid \text{age} = 30, \text{sex} = \text{M}, \text{V. sinus})]$$

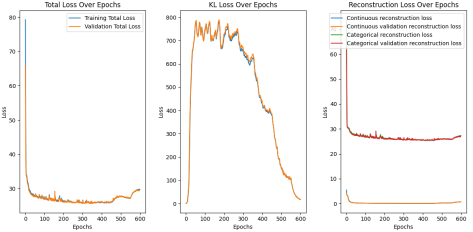
$$\text{but } \mathbb{E}\left[\text{ent}(\text{cat} \mid \text{age} = 10, \text{sex} = \text{M}, \text{span})\right] > \mathbb{E}[\text{ent}(\text{cat} \mid \text{age} = 10, \text{sex} = \text{M}, \text{V. sinus})]$$





Conditional VAE





Preprocessing and data splitting:

- Fill missing values (NA) with 0
- Create a mask for missing values (observed = 1, unobserved = 0)
- Use `StandardScaler` on numerical data
- Apply One Hot Encoding for categorical data
- Split data into training, validation, and test subsets

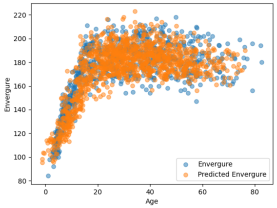
Model setup and training:

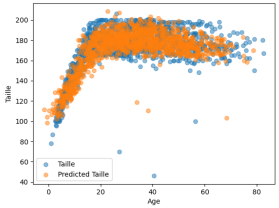
- Loss: Reconstruction (continuous & categorical) + KL loss
- Annealing KL loss: Beta starts at 0, increases to 0.1
- 600 epochs of training
- Uses Gumbel-softmax for categorical variables

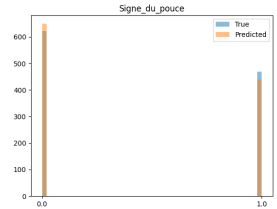
CVAE Results

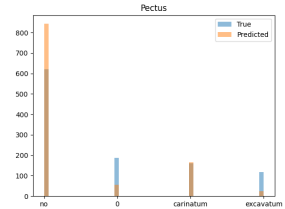
Numerical features	R^2
Anneau	0.79
Sinus_de_valsalva	0.83
Jonction_sino_tubulaire	0.78
Aorte_ascendante	0.79
Crosse_aorte	0.82
Taille	0.91
Poids	0.79
Envergure	0.72

Categorical features	R^2
Luette_bifide	0.38
Palais_ogival	0.51
Signe_du_pouce	0.81
Signe_poignet	0.79
Malocclusion_dentaire	0.39
Degre_ext_des_coudes	0.45
Ectopie	0.60
Pectus	0.35









Conclusion

Comparison of the generative models

CVAE

- better reconstruction of the full joint distribution
- requires extra computation to compute the probabilities
- a single training for all the variables

MICE with PNN

- better reconstruction of the conditional distributions
- straightforward computation of the probabilities
- multiple trainings for all variables

Perspectives

Dynamic calibration of the threshold probabilities according to the saturation state and the center ressources.

MICE with probabilistic neural network

Joint distribution with cycle 2 MICE

$$x \mid x_c \sim \int_{\mathcal{X}} \prod_{i=2}^{d} p_i^{(1)}(x_i' \mid x_c)$$

MICE with probabilistic neural network

Joint distribution with cycle 2 MICE

$$x \mid x_c \sim \int_{\mathcal{X}} \prod_{i=2}^{d} p_i^{(1)}(x'_i \mid x_c) \times p_1^{(2)}(x_1 \mid x'_{-1}, x_c) p_2^{(2)}(x_2 \mid x_1, x'_{-2}, x_c) ... p_d^{(2)}(x_d \mid x_{-d}, x_c) \lambda(\mathrm{d}x')$$

MICE with probabilistic neural network

Joint distribution with cycle 2 MICE

$$x \mid x_c \sim \int_{\mathcal{X}} \prod_{i=2}^{d} p_i^{(1)}(x'_i \mid x_c) \times p_1^{(2)}(x_1 \mid x'_{-1}, x_c) p_2^{(2)}(x_2 \mid x_1, x'_{-2}, x_c) ... p_d^{(2)}(x_d \mid x_{-d}, x_c) \lambda(\mathrm{d}x')$$

Neural networks as elementary univariate probabilistic models

- for continuous variables: $p_i^{(k)}(x_i \mid x') = \mathcal{N}(x_i; \mu(x'; \theta_{ik}), \sigma(x'; \theta_{ik})^2)$
- for categorical variables: $p_i^{(k)}(x_i \mid x') = \mathrm{Cat}(x_i; \mathrm{softmax}(f(x'; \theta_{ik})))$

MICE with probabilistic neural network

Joint distribution with cycle 2 MICE

$$x \mid x_c \sim \int_{\mathcal{X}} \prod_{i=2}^{d} p_i^{(1)}(x'_i \mid x_c) \times p_1^{(2)}(x_1 \mid x'_{-1}, x_c) p_2^{(2)}(x_2 \mid x_1, x'_{-2}, x_c) ... p_d^{(2)}(x_d \mid x_{-d}, x_c) \lambda(\mathrm{d}x')$$

Neural networks as elementary univariate probabilistic models

- for continuous variables: $p_i^{(k)}(x_i \mid x') = \mathcal{N}(x_i; \mu(x'; \theta_{ik}), \sigma(x'; \theta_{ik})^2)$
- for categorical variables: $p_i^{(k)}(x_i \mid x') = \mathrm{Cat}(x_i; \mathrm{softmax}(f(x'; \theta_{ik})))$

Properties: analytical marginalization

In this setting, all conditional and marginal distributions can be approximated by Gaussian mixtures.