ISSN 2282-0930 • Epidemiology Biostatistics and Public Health - XIII[th] SISMEC Congress - Vol. 20 Suppl. 1

Prediction of Risk of Disease in Children at Risk of Facioscapulohumeral Muscolar Dystrophy with Machine Learning Approach

Cuoghi Costantini Riccardo[(1)], Nuredini Andi[(2)], Pini Sara[(2)], Albano Noemi[(2)], Spano Maria Carlotta[(2)], D'Amico Roberto[(1)], Tupler Rossella[(2)]

(1) Università di Modena e Reggio Emilia, Dipartimento di Scienze Mediche e Chirurgiche Materno Infantili e dell'Adulto, Modena, Italy
(2) Università di Modena e Reggio Emilia, Dipartimento di Scienze Biomediche, Metaboliche e Neuroscienze, Modena, Italy

CORRESPONDING AUTHOR: *Cuoghi Costantini Riccardo, riccardo.cuoghicostantini@unimore.it*

INTRODUCTION

Facioscapulohumeral muscular dystrophy (FSHD) (MIM#158900) is one of the most prevalent forms of muscular dystrophy, characterized by progressive skeletal muscle weakness, primarily affecting the face, shoulders, and upper arms. Its genetic basis is complex and typically involves contractions of the D4Z4 repeat region on the 4q subtelomere, even though it might be still incompletely described [1].

As a hereditary disease, an accurate risk assessment is crucial for improving genetic counselling, especially in the context of pregnancy planning. However, due to the significant variability in clinical manifestation and progression and the age-dependent penetrance of the disease, predicting the probability and severity of FSHD in newborns poses several challenges [2], and no tools are available for clinicians for this purpose.

OBJECTIVES

The aim of this study was to develop a machine learning model aimed at enhancing FSHD disease risk prediction for child of D4Z4 alleles of reduced size (DRA) carriers. In particular, our study focused on designing a predictive tool which can estimate the probability of FSHD and the age of disease onset in newborns, given the information of parents and other family members.

METHODS

This predictive model was estimated on the basis of genetic, clinical and socio-demographic data collected in the Italian National Registry for FSHD [3]. Clinical data includes presence and severity of FSHD symptoms, measured using the FSHD Score [4], and a standardized description of clinical phenotypes, obtained through the Comprehensive Clinical Evaluation Form (CCEF) [5]. The availability of detailed family trees allowed the model to include the information carried by each family member, weighted by the degree of kinship with respect to subject involved in the genetic counselling.

To be included in this study, each family must be composed by a child, a DRA carrier parent, and may include one or more relatives. Families which did not fit in this structure were excluded. Since the expected FSHD onset age lies between 15 and 30 years of age, subject with age at visit less than 30 were also exclude, to limit misclassification.

For the development of the predictive model, we relied on a stacking approach: 4 base learners (a generalized regression model (GLM), a random forest (RF), a support vector machine (DVM) and a Bayesian network (BN)) provided a first-level individual prediction. Subsequently, these first-level predictions were combined by a random forest meta-learner to obtain the final predictions. Figure 1 outlines the model structure.

Leave-One-Out cross-validation was used to train each base learner (except BN) and the meta-learner. The parameters and hyperparameters of GLM, RF and SVM were estimated via grid search within each cross-validation loop, using the classification error as performance measure.

The BN learning procedure articulated in two steps: in first place, the structure of the network is established by defining the causal connections among all features, relying on expert knowledge. Then, the probability parameters that describe how the variables influence each other are estimated with a

DOI: 10.54103/2282-0930/

Bayesian approach. Non informative prior distributions were assumed at each non-deterministic node of the network. Since all variables have been previously categorized, each node likelihood was a Multinomial distribution, and a Dirichlet prior was used [6]. The network was embedded also with a set of deterministic nodes, which were introduced to process family trees with different structure and depth, and to reduce the model complexity.

The prediction accuracy of the model for each outcome (occurrence of FSHD and age at onset of first symptoms) was estimated using Leave-One-Out cross validation.

Based on the probabilities estimated from the model, the child of each family was predicted as with FSHD phenotype or as asymptomatic/healthy and assigned to an estimated age at onset class (No Onset, ≤22 years, <22 years). Youden Index was used to estimate the optimal probability cutoffs.

RESULTS

A total of 293 families were included in the study. Of these, 121 families contributed to the estimation of risk of disease base learners, whereas 104 families contributed to the age at onset base learners.

For the evaluation of risk of disease, the model showed an area under the ROC curve (AUC) equal to 0.89. With the selected probability cutoff, the sensibility was equal to 0.90 and the specificity 0.70. The accuracy was 0.75, with 91 out of 121 children correctly assigned to their actual clinical status.

For the estimation of age at disease onset, the model reached a multi-class AUC equal to 0.88, with an accuracy of 0.72 (75 out of 104 children's age at onset correctly predicted).

Overall, 70 children (67.3%) were correctly assigned to both their actual clinical status and their onset age class.

CONCLUSION

The developed predictive model was able to provide accurate estimates of disease probability in children of patients characterised by FSHD symptomatology, even though it was not able to discriminate between finer clinical categories. These findings further support the hypothesis that additional elements, such as other genetic variants and environmental factors, must be consider for predictive purposes.

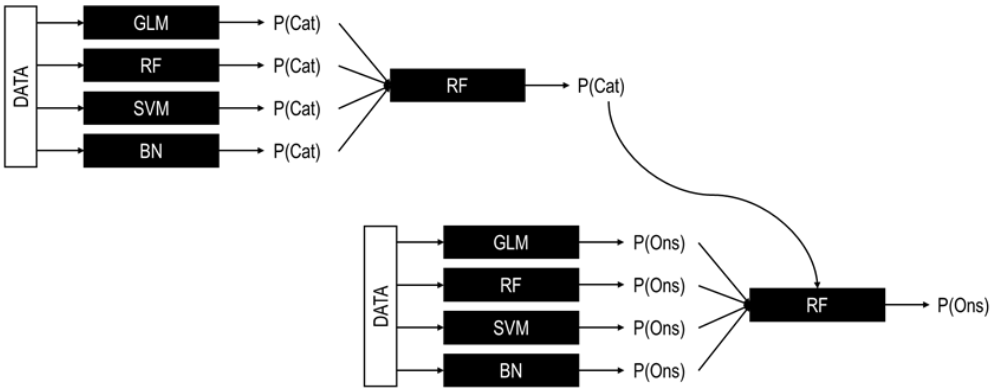
Nevertheless, this model can lead to significant advancements in FSHD genetic counselling and in implementing personalized medicine practices. Notably, our model is based on a very limited number of variables. So, it can be easily applied to provide tailored advice for families at risk of FSHD in real life scenarios.

REFERENCES

1. Ricci G., Scionti I., Sera F., et al. Large scale genotype-phenotype analyses indicate that novel prognostic tools are required for families with facioscapulohumeral muscular dystrophy. Brain. 2013 Nov;136(Pt 11):3408-17.
2. Tawil R., Mah J.K., Baker S., et al. Sydney Workshop Participants Clinical practice considerations in facioscapulohumeral muscular dystrophy. Neuromuscul Disord. 2015;26(7):462–471.
3. Bettio C., Salsi V., Orsini M. et al. The Italian National Registry for FSHD: an enhanced data integration and an analytics framework towards Smart Health Care and Precision Medicine for a rare disease. Orphanet J Rare Dis, 2021, 16, 470.
4. Lamperti C., Fabbri G., Vercelli L. et al, A standardized clinical evaluation of patients afected by facioscapulohumeral muscular dystrophy: the FSHD clinical score. Muscle Nerve, 2010, 42:213–217
5. Ricci G., Ruggiero L., Vercelli L. et al. A novel clinical tool to classify facioscapulohumeral muscular dystrophy phenotypes. J Neurol, 2016, 263:1204–1214
6. Azzimonti, L., Corani, G., Zaffalon, M. Hierarchical estimation of parameters in Bayesian networks, Computational Statistics & Data Analysis, 2019, 137, 67-91.

DOI: 10.54103/2282-0930/29363

Figure 1. Predictive model structure. Black boxes represent machine learning models, with inputs and outputs denoted by incoming and outgoing arrows, respectively. Abbreviations: P(Cat): Probability of occurrence of disease; P(Ons): Probability of age at onset; GLM: Generalized linear regression model; RF: Random Forest; SVM: Support vector machine; BN: Bayesian network



DOI: 10.54103/2282-0930/29363