Semi-parametric models - An application in medicine

Cite as: AIP Conference Proceedings **2293**, 420104 (2020); https://doi.org/10.1063/5.0028556
Published Online: 25 November 2020

J. A. Pereira, A. L. Pereira, and T. A. Oliveira

Semi-Parametric Models - An Application in Medicine

Pereira J.A.[1,a)], Pereira A.L.[2, a)] and Oliveira T.A. [3,4]

[1]*Department of Periodontology - Dental Medicine School of University of Porto. Portugal*
[2]*MMMB, Open University, Portugal.*
[3]*Center of Statistics and Applications (CEAUL), Lisboa, Portugal*
[4]*Open University, Portugal*

[a)]lobopereiramail@gmail.com

Abstract. Modeling of medical data often requires the inclusion of non-linear forms of the predictors and, the Generalized Additive Models (GAMs) can provide an excellent fit in the presence of non-linear relationships and significant noise in the predictor variables. The accurate assessment of QT interval is of paramount importance since its prolongation (LQTS) is a life threatening condition. The QT interval is affected by heart rate and gender and may be adjusted to improve the detection of patients at increased risk. Bazett's formula is the most commonly used QT correction formula, and takes into account only the heart rate assessed by the RR interval, and cut-of values of corrected QT are defined according to gender. In this work we analyzed relevance of QRS, together with gender and RR to explain QT length using GAMs. Results showed that QRS and gender are significant to non-pathological QT modelling.

INTRODUCTION

The QT interval is the segment between the start of Q wave and the end of T wave and represents the time for ventricular depolarization and repolarization, effectively the period of ventricular systole from ventricular isovolumetric contraction to isovolumetric relaxation. Long QT syndrome (LQTS) is a genetic or acquired condition characterized by a prolonged QT interval on the surface electrocardiagram (ECG) and is associated with a high risk of sudden cardiac death due to ventricular tachyarrhythmias[1] and its prevalence is estimated at 1 in 5000, or 25000 individuals in the United States, with approximately 3000 deaths annually, principally in untreated children and young adults[2] and therefore accurate assessment of QT interval is of paramount importance, especially when evaluating cardiovascular risk in young athletes.

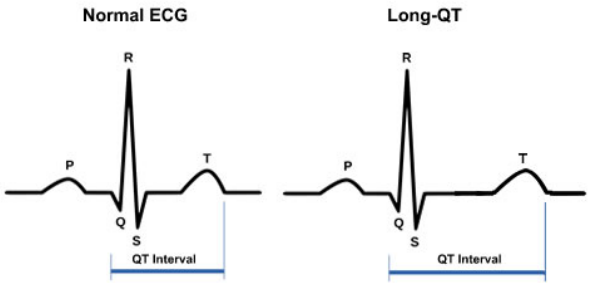

FIGURE 1. ECG trace normal and Long-QT

The generalized additive model (GAM) described by Hastie and Tibirani[3] is a nonparametric extension of the generalized linear model (GLM) to incorporate nonlinear forms of the predictors.

International Conference of Numerical Analysis and Applied Mathematics ICNAAM 2019
AIP Conf. Proc. 2293, 420104-1–420104-4; https://doi.org/10.1063/5.0028556
Published by AIP Publishing. 978-0-7354-4025-8/$30.00

In GLM, proposed by Nelder and Wedderburn[4] (1972), the model is described by form (1). The link function $g(\cdot)$ makes DV into a linear function of the predictors, $g(\cdot)$ may not be linear, but the $X_j\beta_j$ necessarily is.

$$g(\mu) = \alpha + X_j\beta_j = \eta_L \quad (1)$$

Where $g(\mu)$ is the link function that relates the mean μ of the dependent variable (DV) to the linear predictor, η_L, that includes the independent variables (IV). The GAM further generalize the GLM by including non-linear forms of the predictors providing an excellent fit in the presence of non-linear relationships and significant noise in the predictor variables[5] by a sum of smooth functions of DV.

From form (1) we change to form (2)

$$g(\mu) = \alpha + \sum_{j=1}^{p} f_j(X_j) = \eta_A \quad (2)$$

Thus, $g(\mu)$ is the link function that relates the mean μ of the dependent variable (DV) to the linear predictor, η_A, that is sum of smooth functions of X_j, for $j = 1, \dots, p$.

In this paper we fit GLM and GAMs to ECG data of young athletes to explore the relevance of the inclusion of QRS, together with gender and RR to explain QT length and we discuss the interpretability of the models.

METHODS AND RESULTS

The data were processed with package 'mgcv'[7] in R environment[8].

The significance level was set to 0.05.

Descriptive statistics of variables of interest from data of 522 boys and 303 girls, with ages ranging from 5 to 18 years old, who practice sports in the city of Porto are shown in Table 1.

TABLE 1. Descriptive statistics

	Age	**IMC (Kg/m²)**	**QT(ms)**	**RR(ms)**	**QRS(ms)**
Min	5	14.20	290	480	64
Mean	12.82	32.07	377	916	88
Max	18	64.50	468	1474	116

GAM regression models QT on gender, RR and QR, are described in Table 2, other combinations of smooth functions and methods were tried, yielding similar results. In this paper only the more relevant models are presented.

TABLE 2. Equations, smooth functions and method of GAM models

Model	**Equation Model**	**Smooth functions**	**Method**
GAM 1	$QT = Gender + RR + QRS$	none (parametric)	GVC
GAM 2	$QT = Gender + s(RR) + s(QRS)$	thin plate regression splines	
GAM 3	$QT = Gender + s(RR) + QRS$	thin plate regression splines	
GAM 4	$QT = Gender + te(RR, QRS)$	tensor product	

Summary of fitted GAM models are shown in Tables 3 to 6, where GAM 1 is a full parametric model (GLM) and the others are semi parametric.

Consider the following representations: **Inter** – intercept; **Est** – parameter estimator **SE** – standard error; **S-W res** – Shapiro-Wilk normality test; **edf** – effective degrees of freedom; **Ref. df** – Reference degrees of freedom; **PC** and **NPC** – parametric and non-parametric components. Thus, we have:

TABLE 3. Summary of GAM 1

	Inter	**SE**	**p-value**	**Gender**	**SE**	**p-value**	**RR**	**SE**	**p-value**	**QRS**	**SE**	**p-value**
GAM 1	233.97	6.81	0.00	12.33	1.33	0.00	0.13	0.00	0.00	0.19	0.08	0.01
Adjusted Rsq – 0.61; **AIC** – 7003; **GVC** – 283.85**;** **S-W res** – p-value =0.03												

TABLE 4. Summary of GAM 2

GAM 2	Inter			Gender			s (RR)			s (QRS)		
PC	**Est**	**SE**	**p-value**	**Est**	**SE**	**p-value**						
	371.96	0.75	0.00	12.69	1.31	0.00						
NPC							**edf**	**Ref.df**	**p-value**	**edf**	**Ref.df**	**p-value**
							2.84	3.61	0.00	1.96	2.38	0.01

Adjusted Rsq – 0.62; **AIC** – 6977; **GVC** – 275.10**;** **S-W res** – p-value =0.07

TABLE 5. Summary of GAM 3

GAM 3	Inter			Gender			s(RR)			QRS		
PC	**Est**	**SE**	**p-value**	**Est**	**SE**	**p-value**				**Est**	**SE**	**p-value**
	354.45	6.81	0.00	12.61	1.31	0.00				0.20	0.07	0.00
NPC							**edf**	**Ref.df**	**p-value**			
							2.91	3.70	0.00			

Adjusted Rsq – 0.62; **AIC** – 6979; **GVC** – 275.77**;** **S-W res** – p-value =0.05

TABLE 6. Summary of GAM 4

GAM 4	Inter			Gender			te (RR, QRS)		
PC	**Est**	**SE**	**p-value**	**Est**	**SE**	**p-value**			
	372.02	0.75	0.00	12.53	1.31	0.00			
NPC							**edf**	**Ref.df**	**p-value**
							5.95	7.58	0.00

Adjusted Rsq – 0.62; **AIC** – 6976; **GVC** – 274.65**;** **S-W res** – p-value =0.06

TABLE 7. Checking the adequacy of basis dimensions

	GAM 2		GAM 3	GAM 4
	s (RR)	**s (QRS)**	**s (RR)**	**te (RR, QRS)**
k'	9.00	3.00	9.00	24.00
edf	2.84	1.96	2.91	5.95
k-index	0.75	0.75	1.02	1.02
p-value	0.77	0.00	0.78	0.72

k' – maximum possible edf for the term; **edf** – effective degrees of freedom; **k-index** – estimate of the residual variance divided by the residual variance

TABLE 8. Concurvity assessment

	GAM 2			GAM 3		GAM 4	
	param	**s (RR)**	**s (QRS)**	**param**	**s (RR)**	**param**	**te (RR, QRS)**
worst	0.41	0.15	0.25	0.99	0.14	0.42	0.19
observed	0.41	0.14	0.21	0.99	0.14	0.42	0.06
estimate	0.41	0.10	0.24	0.99	0.10	0.42	0.02

worst, observed and estimate values of concurvity indices

The result shows that the introduction of QRS (edf>1, $p<0.05$) and gender ($p<0.05$) in all models helps to explain QT. Selection of the best model was based on adjusted R squared (Adjusted Rsq), Akaike information criteria (AIK), adequacy of bases dimensions (k), normality of the residuals and absence of concurvity. The lowest AIC was presented by GAM 4, as well lowest concurvity indices among all considered models and smaller than 1. The basis dimensions (k) for smooth terms are not so small that they force over-smoothing as suggested by k-index greater than 1 and a high p-value associated with **edf** far from k'. However, the plot showing the smooth components (Figure 2) of GAM

4 (a) are more difficult to interpret than those of GAM 2 (b), which yield similar results in all considered indicators (Tables 4, 7 and 8).

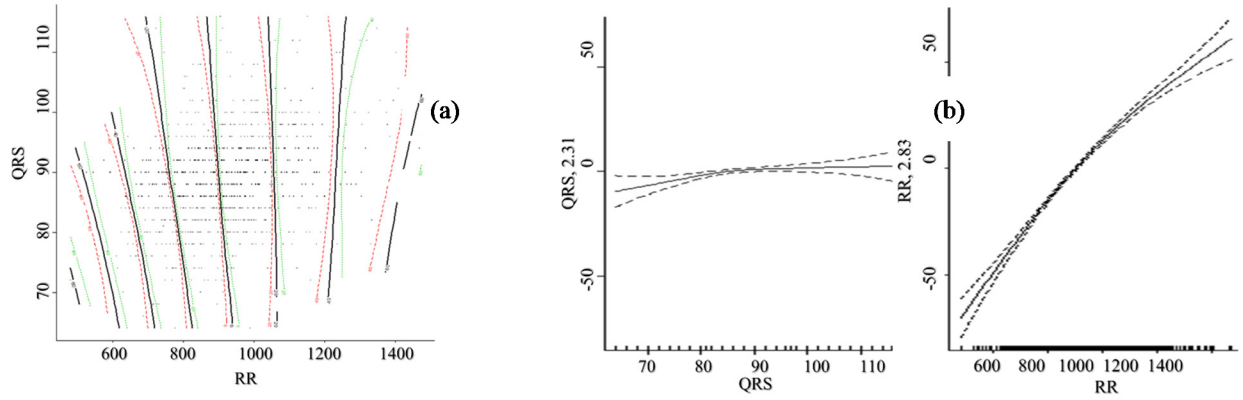

FIGURE 2. Plot of the smooth components of GAM 4 (a) and GAM 2 (b)

CONCLUSIONS

The semi-parametric approach to model medical data showed advantages over the full parametric one, by allowing to model non-linear relationships and getting better fittings. However, the selection of the better fit must be done carefully taking in account R squared (Adjusted Rsq), Akaike information criteria (AIK), adequacy of bases dimension, normality of the residuals and absence of concurvity, which interpretation is similar to GLM equivalents, except basis dimensions (k). The interpretation of smooth components plots using tensor product smooth, which allow the degree of smoothing to be different with respect to different variables (RR, QRS), is not straightforward. The modeling of QT in young athletes must be done including QRS and gender in the models.

REFERENCES

1. K.A. Mehmet and J.P. Daubert JP in BMJ Best Practice https://bestpractice.bmj.com
2. D,M. Roden and P.M. Spooner, "Inherited Long QT Syndromes: A Paradigm for Understanding Arrhythmogenesis", J Cardiovasc Electrophysiol, Vol. **10**, 1664-1683 (December 1999).
3. T. Hastie and R. Tibshirani "Generalized Additive Models", Journal of the Royal Statistical Society, Series B, **1**(3), 297-318 (1986).
4. J. Nelder and R.Wedderburn "Generalized Linear Models", Journal of the Royal Statistical Society. Series A (General), **135** (3): 370–384 (1972).
5. StatSoft, Inc. Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/ (2013).
6. S.N Wood, "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models", Journal of the Royal Statistical Society (B) **73**(1):3-36 (2011).
7. R Core Team, "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/(2013).