Generalized Additive Mixed Models in medicine: a case study on LDL cholesterol in people living with HIV under different antiretroviral regimens

Michele Bellomo[1], Nicola Gianotti[1], Camilla Muccini[1], Hamid Hasson[1], Diana Canetti[1], Silvia Nozza[1,2], Monica Guffanti[1], Emanuela Messina[1], Vincenzo Spagnuolo[1], Riccardo Lolatto[1], Sara Diotallevi[1], Antonella Castagna[1,2], and Laura Galli[1]

[1] IRCCS San Raffaele Scientific Institute, Infectious Diseases Unit, Milan, Italy
bellomo.michele@hsr.it
[2] Vita-Salute San Raffaele University, Milan, Italy

Abstract. Generalized Additive Models (GAMs) are statistical methods that enable the modeling of variables through smooth effects. GAMs are powerful and flexible models, well-consolidated among statisticians. However, their application in medicine, especially in infectious diseases, is still limited. In this study, we applied GAMs to evaluate the dynamic of low-density lipoprotein (LDL) cholesterol in people living with HIV (PLHIV) who have switched to an antiretroviral regimen based on doravirine, rilpivirine, dolutegravir or bictegravir. Data were collected at IRCCS San Raffaele Scientific Institute and involve 2742 individuals and 54552 observations. A Generalized Additive Mixed Model (GAMM) and a quasi-Poisson GAMM were trained considering several risk factors and accounting for temporal dynamics through random effects and autoregressive covariance structures. Doravirine exhibited the most favorable effect on LDL cholesterol, with an associated average reduction in the first two years of use of -5.43 mg/L per year according to the GAMM and of -5.41% per year in the quasi-Poisson GAMM.[3]

Keywords: Generalized Additive Mixed Models, mgcv, HIV, HDL cholesterol

1 Background

Generalized Additive Models (GAMs) are statistical methods that generalize linear models by allowing the nonparametric modeling of variables and estimating their effect on the response variable through a generic smooth function. This relaxes the assumption of linearity and captures nonlinear relationships, while maintaining good model interpretability.

[3] Supported in part by a research grant from Investigator-Initiated Studies Program of Merck Sharp & Dohme LLC. The opinions expressed in this paper are those of the authors and do not necessarily represent those of Merck Sharp & Dohme LLC.

GAMs were introduced in the 1980s [1] and have become a well-known and established methodology among the statistical community. Over time, several libraries and packages have been developed for their convenient use [2]. Nevertheless, their application in the medical field, particularly in infectious diseases, remains rare.

In this study, we compared the dynamic of low-density lipoprotein (LDL) cholesterol in people living with HIV (PLHIV) who have switched to an antiretroviral regimen based on doravirine (DOR), rilpivirine (RPV), dolutegravir (DTG) or bictegravir (BIC). Dyslipidemia is a common issue in PLHIV, caused by multiple factors and, sometimes, worsened by side effects of antiretroviral drugs. Dyslipidemia is a clinical condition that leads to an increased risk of cardiovascular events and thus higher mortality. Understanding the effects of different antiretroviral drugs on this condition is crucial for physicians to choose the best therapy and improve the quality of life of PLHIV.

2 Methods

2.1 Study design and cohort

Observational cohort study on data collected in a large database during routine visits of PLHIV followed at the Infectious Diseases Unit of the IRCCS San Raffaele Scientific Institute (Milan, Italy). The study inclusion criteria are: adult PLHIV, antiretroviral experienced, who switched to a antiretroviral regimen containing DOR, RPV, DTG, or BIC, naïve to the studied drugs at the therapy switch, with at least 2 LDL cholesterol determinations before the switch and at least 4 determinations after the switch. Follow-up accrued from the first LDL cholesterol determination within a 5 year interval before the switch, until the last available determination after the switch or the start of an antiretroviral regimen not including any of the considered drugs (DOR, RPV, DTG, BIC). The same individual may take more than one drug under study during follow-up, either at different times, due to a subsequent change in therapy, or simultaneously (an antiretroviral therapy is generally constituted by a mix of antiretroviral drugs).

2.2 Covariates

In addition to DOR/RPV/DTG/BIC intake, the following variables were considered: age, sex, body mass index (BMI), presence of comorbidities (diabetes, cardiovascular events), CD4+ cell count (type of white blood cell essential for the immune system and targeted by HIV, whose count indicates immune health in PLHIV), viremia (count of HIV copies present in a millimeter of blood, crucial for assessing the level of viral suppression in PLHIV), use of statins (drugs used to lower cholesterol levels in the blood), use of antiretroviral drugs other than DOR/RPV/DTG/BIC, grouped by class (NRTI, NNRTI, PI, INSTI, boosters), with the exception of tenofovir disoproxil fumarate (TDF) and tenofovir alafenamide (TAF), that were considered separately. TDF has a well-known lipids-lowering effect, while the role of TAF on lipids is still debated. Since TDF and

TAF are often given in combination with the study drugs and in a mutually exclusive manner, taking their influence into account is essential to remove possible confounding factors.

2.3 Statistical modeling

Generalized Additive Mixed Models (GAMM) were used to study the LDL cholesterol dynamic. Covariates related to antiretroviral drugs and statins were coded as time-dependent variables, counting the number of years of continuous use. Continuous variables were modeled non-parametrically via smooth effect, except for the study's target drugs, which were modeled parametrically via β regression coefficient to facilitate comparisons.

Individual characteristics and temporal dynamics were modeled using random effects (random intercept + random slope on the years from the first switch) and an autoregressive covariance structure (first-order AR(1)).

Among the four studied antiretroviral drugs, some are more recent than others. For example, RPV was put on the market in 2011, while DOR in 2018. As a consequence, follow-up duration greatly varies across the four drugs, and therefore we decided to estimate slopes and confidence intervals related to these drugs distinguishing two intervals: the first 2 years of continuous use and longer periods (2 years is the longest period for which there were sufficient data to estimate the effect of the drug with shorter follow-up).

Finally, we trained a quasi-Poisson GAMM to deal with the overdispersion of GAMM residuals. A quasi-Poisson is a model trained using quasi-likelihood, a quantity with properties similar to likelihood, but which only requires assumptions on the mean-variance relationship for its definition. In a quasi-Poisson model, the variance of the response variable y_i has the form $V(\mu) = \phi\mu$, where V is the variance function, μ the mean of y_i and $\phi > 0$ a proportionality factor. The quasi-Poisson model can be seen as a generalization of the Poisson model (in the Poisson model the variance and the mean of y_i are equal, case $\phi = 1$). The ϕ parameter indicates the degree of dispersion of the quasi-Poisson compared to the standard Poisson, and the case $\phi > 1$ is suitable for count data that are more dispersed than a Poisson distribution. Since no specific assumptions are made on the distribution, the quasi-Poisson is appropriate also for continuous data. In the quasi-Poisson model we used the default log-link function, and therefore estimates are interpreted in terms of percentage changes.

All the models were fitted using R version 4.3.0 and the `mgcv` package [3].

2.4 Results

Overall, 2742 participants met the inclusion criteria, for a total of 54552 determinations.

All smooth effects are significant at the 5% level, except for NRTI other than TDF and TAF, INSTI other than BIC and DTG, and PI. Figure 1 shows the estimated smooth effects of the significant variables.

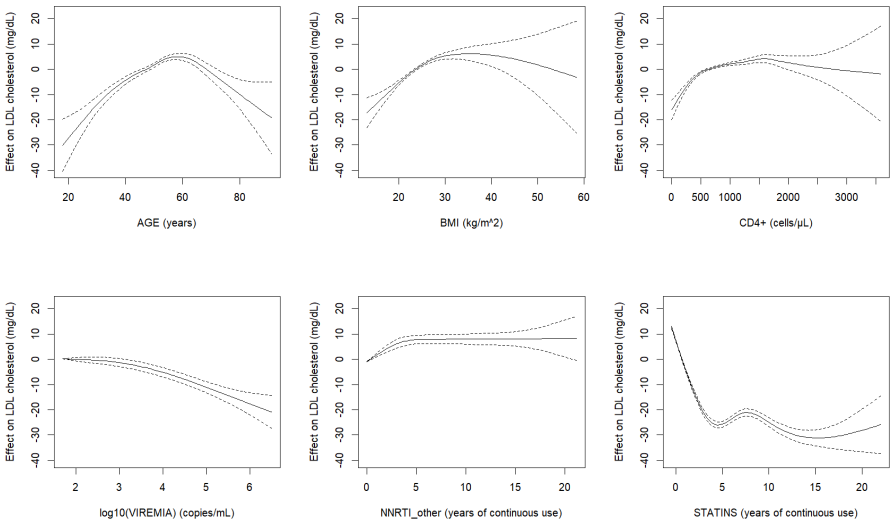


Fig. 1: Estimated significant smooth effects

Estimated effects have non-linear behaviors, which justifies the use of GAMs and their non-parametric modeling. LDL cholesterol appears to increase with age up to about 60 years, and then start to decrease. The LDL cholesterol is higher in people with higher BMI till 35 kg/m^2, where the effect of BMI reaches a plateau (for BMI values greater than 35 kg/m^2, the estimated smooth effect has higher margins of uncertainty due to the small number of individuals with such BMI values). LDL cholesterol decreases for CD4+<500 cells/μL and viremia>10^3 copies/mL, indicating that people more immunocompromised and with active infection have lower cholesterol. Statins have an important lowering effect, which peaks after 5 years of use and settles around a total average decrease of 25 mg/dL. We interpret the oscillations around the steady-state value as noise. Boosters (not shown in Figure 1) and NNRTI other than DOR and RPV have similar increasing effects, which peak after 4-5 years of continuous use and settle around a total average increase of 10 mg/dL.

Concerning parametric modeled variables, the estimated intercept is 119.75 mg/dL ($p<0.001$). There are no significant differences in LDL cholesterol between men and women (mean difference 0.17 mg/dL, $p=0.905$). A diabetes diagnosis leads to a mean decrease in LDL cholesterol of -4.97 mg/dL ($p=0.002$), while the occurrence of a cardiovascular event leads to a mean decrease of -14.00 mg/dL ($p<0.001$). These effects can be justified by healthier lifestyles and introduction of permanent lipid-lowering therapies following these diagnoses. The lipid-reducing effect of TDF is confirmed in the first 2 years of use (mean reduction of -3.77 mg/dL per year, $p<0.001$), while for periods > 2 the effect changes (average increase of 0.44 mg/dL per year, $p<0.001$). Regarding TAF, a slight

decreasing effect is detected for periods longer than 2 years (average decrease of -1.28 mg/dL per year, p=0.004), while there are no significant effects for the first 2 years of use (average increase of 0.28 mg/dL per year, p=0.569). β coefficients, 95% confidence intervals and p-values related to the drugs under study are reported in Table 1. As the median follow-up for DOR was 1.2 years (IQR 0.6-1.7), no slope could be estimated for the time interval > 2 years.

	Mean change per year 0-2 years	**Mean change per year >2 years**
DOR	−5.43 mg/L [−9.78, −1.08] $p = 0.014$	NA
RPV	1.01 mg/L [0.26, 1.76] $p = 0.008$	0.69 mg/L [0.26, 1.12] $p = 0.002$
DTG	1.10 mg/L [0.43, 1.77] $p = 0.001$	−0.31 mg/L [−0.76, 0.15] $p = 0.184$
BIC	−1.59 mg/L [−2.90, −0.27] $p = 0.018$	1.11 mg/L [−1.15, 3.37] $p = 0.335$

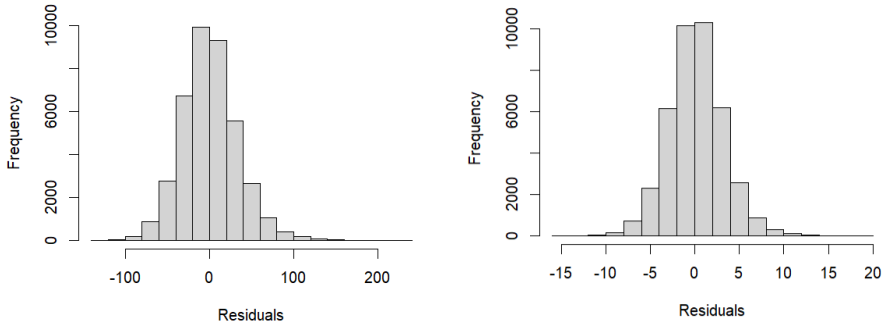
Table 1: β estimates, 95% confidence intervals and p-values for the mean annual changes in LDL cholesterol related to the drugs under study

RPV, DTG, and BIC have statistically significant but modest effects on LDL cholesterol in the first 2 years of use. DOR, instead, has an important reduction effect of −5.43 mg/L per year in the first 2 years of use. For periods greater than 2 years, a statistically significant effect is detected only for RPV (slight increase of 0.69 mg/L per year).

Checking the model assumptions, we notice some overdispersion in the right tail of the residuals. We tackle this problem by training a quasi-Poisson GAMM. In this way, the overdispersion is eliminated and the distribution of deviance residuals is symmetric, as shown in Figure 2. Shapes of smooth effects, estimated coefficients and p-values obtained in the quasi-Poisson GAMM are consistent with the GAMM with Gaussian distribution. In particular, DOR is associated with a mean decreasing effect of −5.41% per year in the first two years of use (p=0.003).

3 Conclusions

We applied GAMs to study LDL cholesterol in PLHIV under different antiretroviral regimens. Smooth effects enabled a more detailed modeling of variables and a consequent better estimate of errors and variances. At the same time, by modeling the study's target drugs parametrically via β coefficient, we obtained the main results in a format familiar to physicians and more suitable for comparisons and dissemination. The flexibility of the `mgcv` library enabled the inclusion



(a) GAMM (b) quasi-Poisson GAMM

Fig. 2: Histogram of residuals for model diagnostic

in the models of random effects, covariance structures and distributions alternative to the Gaussian. The quasi-Poisson GAMM showed a better fit to the data thanks to a less dispersion of the residuals. However, we presented the results of both models to demonstrate the robustness of our findings and, acknowledging the differing interpretations of estimated coefficients due to the distinct link functions, to offer a more comprehensive and general presentation of the results.

Among the drugs targeted in the study, DOR showed the best outcome on LDL cholesterol, with an estimated average reduction in the first two years of use of -5.43 mg/L per year in the GAMM and of -5.41% per year in the quasi-Poisson GAMM.

We believe this work can motivate more researchers to explore GAMs, essential instruments in the toolkit of a modern biostatistician.

References

1. Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.
2. Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.
3. Simon N Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36, 2011.