Tutorial in Biostatistics

Statistics in Medicine

Received 23 March 2009, Accepted 24 November 2009 Published online 19 January 2010 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3841

# Dose-response analyses using restricted cubic spline functions in public health research

**Loic Desquilbet[a]*† and François Mariotti[b]**

**Taking into account a continuous exposure in regression models by using categorization, when non-linear dose-response associations are expected, have been widely criticized. As one alternative, restricted cubic spline (RCS) functions are powerful tools (i) to characterize a dose-response association between a continuous exposure and an outcome, (ii) to visually and/or statistically check the assumption of linearity of the association, and (iii) to minimize residual confounding when adjusting for a continuous exposure. Because their implementation with SAS® software is limited, we developed and present here an SAS macro that (i) creates an RCS function of continuous exposures, (ii) displays graphs showing the dose-response association with 95 per cent confidence interval between one main continuous exposure and an outcome when performing linear, logistic, or Cox models, as well as linear and logistic-generalized estimating equations, and (iii) provides statistical tests for overall and non-linear associations. We illustrate the SAS macro using the third National Health and Nutrition Examination Survey data to investigate adjusted dose-response associations (with different models) between calcium intake and bone mineral density (linear regression), folate intake and hyperhomocysteinemia (logistic regression), and serum high-density lipoprotein cholesterol and cardiovascular mortality (Cox model). Copyright © 2010 John Wiley & Sons, Ltd.**

## 1. Introduction

How to take into account a continuous exposure in regression models has been widely described in the literature [1, 2]. There are mainly two domains where it is necessary to consider the coding of a continuous exposure that is included in regression models.

The first domain is the characterization of the dose-response curve representing the association between the continuous exposure and a health-related outcome (i.e. a curve showing the relationship between the amount, intensity, or duration of the exposure and the risk of the outcome [3]). This can be used either when there is no *a priori* hypothesis regarding the shape of the dose-response association or to test the assumption of linearity of the association before including the exposure into the model with appropriate recoding. Dose-response analyses are often carried out by categorizing the exposure, although several limitations have already been pointed out [4–6], including loss of information and reduction in power, and modeling a continuous dose–response association with a biologically implausible step function [7]. Alternatives to categorization include the use of fractional polynomials (FPs) [8, 9] or spline functions [10, 11]. Both methods use all data points to estimate the dose-response association between the exposure and the outcome [12], and have the advantage that they can fit both complex and linear associations.

The second domain is the adjustment for a potential confounder that is continuous. In this case, residual confounding may persist if not all information is taken into account in the model, which is clearly the case when using categorization [13, 14]. Alternatives to categorization have been proposed, including the use of spline functions [15].

Because linear associations in epidemiologic research can rarely be assumed *a priori*, there is a need for practical tools to explore adjusted dose-response associations between a continuous exposure and an outcome. Based on the strong theoretical advantages of restricted cubic spline (RCS) functions, the aim of this paper is to present an SAS macro that makes the use of RCS functions easier in practice when the objectives are to (i) graphically characterize the dose-response association between a continuous exposure and an outcome, (ii) test the assumption of linearity of the association, and (iii) quantify the association when the

[a] *AgroParisTech, UMR 1290 BIOGER-CPP, F-75005 Paris, France*

[b] *AgroParisTech, CRNH-IdF, UMR914 Nutrition Physiology and Ingestive Behavior, F-75005 Paris, France*

* *Correspondence to: Loic Desquilbet, National Veterinary School of Alfort, Laboratory of Epidemiology and Animal Infectious Diseases, F-94704 Maisons-Alfort, France.*

† *E-mail: ldesquilbet@vet-alfort.fr*

latter assumption is not valid. We illustrated the SAS macro by using the third National Health and Nutrition Examination Survey (NHANES III) data in three analyses (linear, logistic, and Cox regression analyses). We chose examples from nutritional epidemiology since dose-response associations in this field would not be expected to be linear *a priori* [16, 17]. The analyses were chosen based on the well-established associations between continuous exposures (calcium intake, folate intake, and serum high-density lipoprotein [HDL] cholesterol) and outcomes (bone mineral density (BMD), hyperhomocysteinemia (HHcy), and cardiovascular mortality) [18–21] that also proved to be *a posteriori* good examples for non-linear dose-response associations.

The paper is arranged as follows. In the next Section, we provide an overview of spline functions. In Section 3, we focus on RCS functions in linear, logistic, and Cox proportional hazard model [22]. In Section 4, we briefly present four studies that have used RCS functions to investigate dose-response associations or to minimize residual confounding. The SAS macro is presented in Section 5. The data that we used to illustrate the SAS macro are presented in Section 6, and the results of the analyses are presented in Section 7. In Section 8, we compare the GAM SAS procedure with our SAS macro. Section 9 are concluding remarks. The SAS version used for the examples presented in this paper was V9.1.3.

## 2. Overview of basic characteristics of spline functions

A spline regression includes a continuous exposure coded by using spline functions, i.e. piecewise functions whose 'pieces' are polynomials (splines) of low degrees (usually, $\leqslant 3$) defined over adjacent intervals [11, 23]. The junction between two intervals is called 'knot'. Typically, there is a small number of knots (between 3 and 8), which must be specified by the user; let $K$ be the number of knots. The general formula of an unadjusted spline regression is the following:

$$Y(v) = \alpha + \sum_{i=0}^{N} \delta_i \cdot S_i(v) \tag{1}$$

where $v$ is the value of a continuous exposure $V$, $\alpha$ the intercept, $S_i$ the $i$th spline, $\delta_i$ the estimate related to the $S_i$ spline, and $N$ such that $N+1$ is the total number of splines included into the regression. The total number of splines depends on the number of knots $K$ and the type of spline functions. There are at least six types of spline functions that are commonly used in epidemiology, such as binary (step), linear, quadratic, cubic, restricted quadratic, and RCS functions [11, 24].

Table I provides the formula of the $S_i$ splines, according to the type of spline functions, using the '+' function defined as

$$u_+ = u \quad \text{if } u>0$$
$$= 0 \quad \text{if } u\leqslant 0$$

In contrast to a binary spline function, a linear spline function is a sum of straight lines with a change in slope at each knot that allows the risk to vary within as well as between categories, without sudden jump in risk from one interval to the next one. Quadratic and cubic spline functions are a sum of polynomials with degrees of 2 and 3, respectively, with continuity and slope constraints at each knot. Compared with a linear spline regression, these constraints allow smoother risk characterizations by avoiding sharp bends at each knot. Restricted quadratic and cubic spline functions are quadratic and cubic spline functions with an additional constraint of linearity before the first knot and after the last knot. This feature avoids odd behaviors in open-ended tails of the exposure distribution and therefore makes the dose-response association more realistic. As shown in Table I, RCS regressions with $K$ knots contain $K$-1 estimates (intercept excepted), which is lower than the number of estimates in linear, quadratic, restricted quadratic, and cubic spline regressions ($K+1$, $K+2$, $K$, and $K+3$, respectively).

In a purpose of visual comparisons between types of spline functions, Figures 1(a) and (b) show the unadjusted relationship between beta-carotene intake (in mg/day) and systolic blood pressure (in mmHg) that can be obtained from NHANES III data, using the six types of spline functions described in Table I, with five knots located at 0.5, 1.0, 1.5, 2.0, and 2.5 mg/day. As shown in these figures, the RCS function allows a flexible curve without odd behaviors before the first knot and after the last knot.

In conclusion, RCS regressions have advantages of parsimony while allowing smooth and plausible dose-response curves with a wide range of possible shapes characterizing the association between the continuous exposure and an outcome. However, the

**Table I**. $S_i(v)$ expression in equation (1) according to the type of spline functions.

| Type of spline function | Value of $N$ | $S_i(v)$ expression |
|---|---|---|
| Binary | $K-1$ | 1 if $v>t_{i+1}$, 0 if $v\leqslant t_{i+1}$ |
| Linear | $K$ | $v$ if $i=0$; $(v-t_i)_+$ if $i>0$ |
| Quadratic | $K+1$ | $v$ if $i=0$; $v^2$ if $i=1$; $(v-t_{i-1})^2_+$ if $i>1$ |
| Cubic | $K+2$ | $v$ if $i=0$; $v^2$ if $i=1$; $v^3$ if $i=2$; $(v-t_{i-2})^3_+$ if $i>2$ |
| Restricted quadratic | $K-1$ | $v$ if $i=0$; $(v-t_i)^2_+ - (v-t_K)^2_+$ if $i>0$ |
| Restricted cubic | $K-2$ | $v$ if $i=0$; $(v-t_i)^3_+ - \frac{t_K-t_i}{t_K-t_{K-1}}(v-t_{K-1})^3_+ + \frac{t_{K-1}-t_i}{t_K-t_{K-1}}(v-t_K)^3_+$ if $i>0$ |

$N$, the maximum value of $i$ in the sum sign, in equation (1); $K$ the total number of knots; $S_i(v)$ the $i$th spline; $t_j$ the value of the $j$th knot ($j\in\{1,\ldots,K\}$; $u_+ = u$ if $u>0$, 0 otherwise.

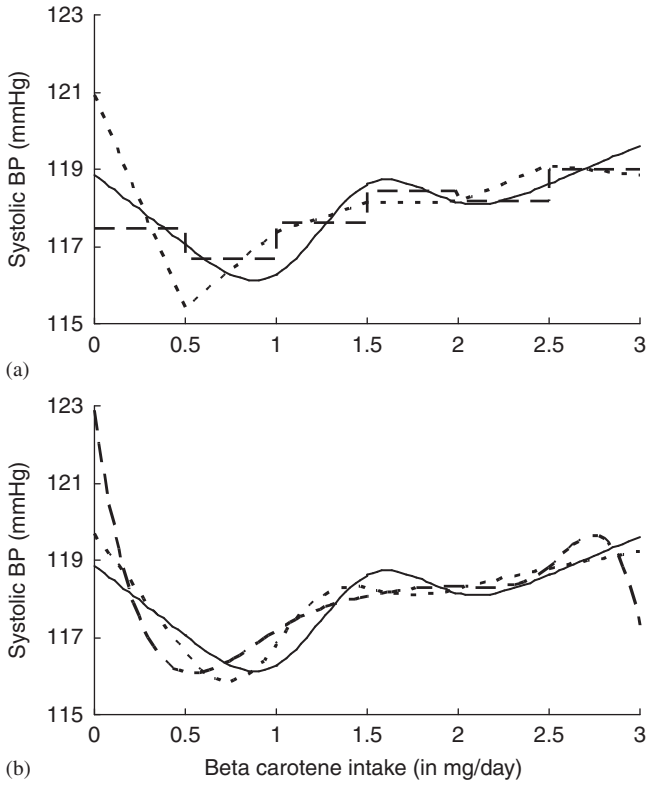


**Figure 1**. Unadjusted relationship between beta-carotene intake (in mg/day) and systolic blood pressure (BP, in mmHg) using NHANES III data, using spline functions with five knots, located at 0.5, 1, 1.5, 2, and 2.5 mg/day. In (a) and (b), the solid line is a restricted cubic spline function. In (a), short dashes represent a binary spline function, and long dashes represent a linear spline function; in (b), short dashes represent a restricted quadratic spline function, and long dashes represent a cubic spline function. The quadratic spline function (not shown) was virtually coinciding with the cubic spline function.

drawbacks of RCS regressions are (i) the complicated formula expression of an RCS function that could be a source of errors in programming statements, and (ii) the fact that values of the estimates $\delta_i$ related to each spline $S_{i,i>0}$ are virtually not interpretable. Practically, solving these two points was an important objective of the macro that we developed.

## 3. RCS functions in linear, logistic, and Cox proportional hazard models

In all following models in Section 3, $Y$ is the outcome (either continuous or binary), $V$ is a continuous exposure under study for the dose-response association with the outcome, and $K$ is the number of knots of the RCS function of $V$.

### 3.1. Linear model

Using notations in Table I for an RCS function, $Y(v)$ is a continuous outcome and $\alpha$ is the estimated mean value of $Y(V=0)$. If $\delta_0=\delta_1=\cdots=\delta_{K-2}=0$, there is no overall association between $Y$ and $V$. If $\delta_0\neq0$ and $\delta_1=\cdots=\delta_{K-2}=0$, the association between $Y$ and $V$ is linear (i.e. $\delta_0$ is the estimated increase in $Y$ per 1-unit increase in $V$). If $\delta_0\neq0$ and $\delta_{i,i>0}\neq0$, the association between $Y$ and $V$ is not linear.

### 3.2. Logistic model

In case-control or cross-sectional designs, let $D$ be a binary outcome that values 1 for cases and 0 for controls, and $P(D=1|V)$ be the probability of $D=1$ according to values of $V$. Using the notations in Table I for a logistic model, $Y(v)$ is Logit($P(D=1|V)$) and $\alpha$ is the estimated mean value of Logit($P(D=1|V=0)$). If $\delta_0=\delta_1=\cdots=\delta_{K-2}=0$, there is no overall association between the presence of $D$ and $V$. If $\delta_0\neq0$ and $\delta_1=\cdots=\delta_{K-2}=0$, the association between Logit($P(D=1|V)$) and $V$ is linear (i.e. $\delta_0$ is the estimated Ln(Odds Ratio [OR]) quantifying the risk for the presence of $D$ per 1-unit increase in $V$). If $\delta_0\neq0$ and $\delta_{i,i>0}\neq0$, the association between Logit($P(D=1|V)$) and $V$ is not linear.

### 3.3. Cox proportional hazard model

In a cohort design, let $h(t)$ be the hazard function of an event occurring during the follow-up in the cohort at time $t$, and $h_0(t)$ the hazard function at time $t$ for individuals with $V=0$. Using notations in Table I for a Cox proportional hazard model, $Y(v)$ is

Ln($h(t)$) and $\alpha$ is Ln($h_0(t)$). If $\delta_0=\delta_1=\cdots=\delta_{K-2}=0$, there is no overall association between the incidence of the event and $V$. If $\delta_0\neq0$ and $\delta_1=\cdots=\delta_{K-2}=0$, the association between Ln($h(t)$) and $V$ is linear (i.e. $\delta_0$ is the estimated Ln(Hazard Ratio [HR]) quantifying the risk for the incidence of the event per 1-unit increase in $V$). If $\delta_0\neq0$ and $\delta_{i,i>0}\neq0$, the association between Ln($h(t)$) and $V$ is not linear.

### 3.4. General comments

In any regression model, an RCS function with three knots includes the two splines $S_0$ and $S_1$ with their respective estimates $\delta_0$ and $\delta_1$. In this case, $S_1$ is the non-linear part of the dose-response association, and $\delta_1$ quantifies how much the dose-response association is far from linear. Therefore, with an RCS function with three knots, testing $\delta_1 = 0$ is equivalent to test whether the assumption of linearity is statistically rejected or not.

To visually check the assumption of linearity of the association between the exposure and the outcome with logistic of Cox models, the $Y$-axis must be Ln(OR) or Ln(HR), respectively, instead of OR or HR.

## 4. RCS regressions in literature

Many studies have used RCS regressions to investigate the dose-response association between continuous exposures and an outcome or to minimize residual confounding for continuous confounders. In this section, we present four studies in order to provide a short overview of potential applications for the present SAS macro. Other examples of the use of RCS functions can be found elsewhere [12, 25–27].

The objective of the study of Saraiya *et al.* was to investigate the association between prostate-specific antigen (PSA, continuous outcome), as a biomarker of early detection of prostate cancer, and age (continuous exposure) [28]. The study included 1320 men $\geqslant$40 years from NHANES 2001–2002 study. Linear regressions were performed, and age was included into the model by using an RCS function. The authors have shown a non-linear association with age among Whites non-Hispanic and Mexican Americans, with a low increase in PSA levels with increased ages for ages $\leqslant$65 years followed by a much steeper increase for ages $>$65 years.

The objective of the study of Jiang *et al.* was to investigate the association between myocardial ischemia (binary outcome) and depression, measured by the Center for Epidemiological Studies—Depression (CES-D) scale (continuous exposure) [29]. The study included 135 patients with documented ischemic heart disease (IHD). Logistic regressions included CES-D scores by using an RCS function, and were adjusted for age, resting left ventricular ejection fraction, and history of myocardial infarction. The authors have shown inverted U-shape associations, with lower risks for ischemia for patients with lowest or highest values of CES-D scores.

The objective of the study of Schernhammer *et al.* was to assess the association between folate intake and risk for colon cancer according to p53 expression, among 88 691 women included and followed in The Nurses' Health Study [30]. There were 399 incident cases of colon cancer accessible for p53 expression. Multivariate Cox proportional hazard regression models were performed, including folate and vitamin B6 intakes using RCS functions. Models were adjusted for several cofactors including age, energy intake, smoking, alcohol consumption, and body mass index. The authors have shown J-shape associations between folate or vitamin B6 intakes and risk for colon cancer, but only for p53-positive cancers, with higher risks for cancers for lowest and highest intakes.

The objective of the study of Troude *et al.* was to assess whether there was a change in HIV-1 virulence between 1996 and 2007, among 903 HIV-infected patients [31]. Patients were recruited in the French ANRS PRIMO cohort within 3–6 months after documented HIV infection. Multivariate linear regression models were performed to evaluate temporal trends in HIV virulence measured at enrolment by CD4 cell count, plasma HIV RNA, or intracellular HIV DNA levels. Year of infection was coded using an RCS function, as well as confounders, such as age and time, between biological measurements and HIV infection in order to minimize residual confounding. The authors have found no graphical evidence of deviation from linearity for any of the markers of HIV virulence, suggesting that HIV-1 virulence remained stable between 1996 and 2007.

## 5. The SAS macro

The SAS macro, named %RCS_Reg, addresses the two drawbacks of the RCS regressions, i.e. the complicated formula in a data set that may lead to errors in programming and the virtually impossible interpretation of the estimates related to each spline included in the model.

First, the SAS macro creates in an SAS datafile the RCS function with $K$ knots of one or more continuous exposures (i.e. it creates the $K-2$ splines described in Table I). The user can choose the number of knots (between $K=3$ and $K=5$) and the location of the knots for each continuous exposure. The user can choose the location of knots according to either the percentiles of the distribution of the exposure(s) or some arbitrary values of the exposure(s). By default, the SAS macro creates an RCS function with three knots, located at the 5th, 50th, and 95th percentiles; it also displays the value of the knots in the SAS LOG window, and some elements of the distribution of the continuous exposure(s) in the SAS OUTPUT window.

Second, but optionally, the SAS macro includes these splines into a regression model. The proposed regressions are linear, logistic, and the Cox proportional hazard regressions, as well as generalized estimating equations (GEE) models for repeated

measurements [32] with either an identity or a logit link. Regressions can be adjusted for potential confounders (including continuous confounders coded using RCS functions). The macro provides in the SAS OUTPUT window two statistical tests for the main continuous exposure (specified by the user): (i) a test where the null hypothesis is $\delta_0=\delta_1=\cdots=\delta_{K-2}=0$ (test of an overall association between the main exposure and the outcome), and (ii) a test where the null hypothesis is $\delta_1=\cdots=\delta_{K-2}=0$ (test of a non-linear association between the main exposure and the outcome). In the SAS GRAPH window, the SAS macro displays the (unadjusted or adjusted) dose-response association (with its 95 per cent confidence interval [CI]) between the main continuous exposure and the outcome, when comparing individuals with any value of the exposure (within the range of the exposure values in the datafile) with individuals with a reference value. This reference value is by default the median of the exposure distribution, but it can also be specified by the user. For linear regressions (including GEE linear regressions), the macro displays the estimated difference in the continuous outcome; for logistic regressions (including GEE logistic regressions), the macro displays Ln(OR) of having the outcome; for a Cox model, the macro displays Ln(HR) of presenting the outcome during follow-up. For logistic and Cox models, one option of the macro enables to display ORs or HRs, instead of Ln(ORs) or Ln(HRs). However, this option is not recommended for visual inspection of the assumption of a linear association between the exposure and the outcome in logistic or Cox models because this assumption must be visually checked on the logarithmic scale only. Optionally, the SAS macro provides in the SAS OUTPUT window the value of the estimated association (difference, Ln(OR), OR, Ln(HR), or HR), with its 95 per cent confidence interval, when comparing individuals with one or more specific values of the exposure (assigned by the user) with individuals with the reference value (the one that was used to display the dose-response association in the SAS GRAPH window). This option is useful to present the dose-response association in a table for some meaningful values of the exposure.

Upon request, the SAS macro provides the predicted value of the outcome for each observation of the sample on which the model was developed. For linear and logistic regressions, the macro also provides 95 per cent confidence intervals of the predicted values. For the Cox model, the macro provides the predicted value of the survival function $S(t)$ computed by using the product-limit method.

The SAS macro also allows 'by factor' analyses, i.e. analyses stratified on a categorical exposure. In this case and if the user chooses to define knots according to the percentiles of the distribution of quantitative variables, these percentiles will be specifically determined within each category of the categorical exposure. In a 'by factor' analysis, the macro provides all the statistical results, location of knots, and graphs for each category.

Confidence intervals are obtained by calculating the matrix product $\underline{d}'\otimes C\otimes\underline{d}$ where $\underline{d}=(\underline{z}(v)-\underline{z}(ref))$ is the vector of differences of the $\underline{z}(v)$ and $\underline{z}(ref)$ components, $\underline{z}(v)=(S_0(v),S_1(v),\ldots,S_{K-2}(v)),S_i(v)$ the $i$th spline described in equation (2), *ref* the reference value (assigned by the macro or chosen by the user), and $C$ the estimated covariance matrix [11].

GENMOD SAS procedure is used for linear and logistic regressions (including GEE models), PHREG SAS procedure is used for the Cox proportional hazard model, IML SAS procedure is used for matrix products calculations, and GPLOT SAS procedure is used to display graphs. The overall association and non-linear association tests are Wald $\chi^2$ tests with $K-1$ and $K-2$ degrees of freedom (DF), respectively, and are provided by using the WALD option in the CONTRAST statement of the GENMOD procedure, or by using the TEST statement in the PHREG procedure.

Akaike Information Criteria (AIC) [33] may provide useful information for model selection. The model that better explains the observations, while requiring the lower number of parameters, is the one with the lower AIC. It is calculated by the SAS macro according to the $-2\text{Log}(L)+2n$ formula, where $L$ is the maximum likelihood value and $n$ the number of parameters of the model (including the intercept). The AIC value is provided in the SAS LOG window.

In Appendix A, we provide a full description of the parameters required by the %RCS_REG SAS macro. In Appendix B, we provide the mode of availability of our SAS macro.

## 6. Application to NHANES III data

### *6.1. Continuous exposures and dependent variables*

To illustrate the %RCS_REG SAS macro, we used data from NHANES III and NHANES III Linked Mortality Public-use data. NHANES III is a cross-sectional examination survey of the representative US civilian non-institutionalized population aged 2 months and older, which was implemented from 1988 to 1994 (http://www.cdc.gov/nchs/about/major/nhanes/nh3data.htm). The NHANES III Linked Mortality File provides mortality follow-up data from the date of NHANES III survey participation through 31 December 2000, for examinees aged 17 years and older. The NHANES III Total nutrient intake File provides estimate for usual nutrient intake, based on 24-h dietary recalls. Serum total homocysteine concentrations was measured from phase 2 (1991–1994) blood sampling of the NHANES III [34]. BMD of the proximal femur (the hip region including the femoral neck, intertrochanter, trochanter, and total hip) was measured by dual X-ray absorptiometry [35].

To illustrate the SAS macro for the linear regression, we chose calcium intake as the main continuous exposure (in g/day [variable named CALCIUM in the SAS datafile]) and BMD at the femur neck region as the continuous outcome (in g/cm$^2$ [BMD_FEMUR]). For the logistic regression, we chose folate intake as the main continuous exposure (in mg/day [FOLATE]) and HHcy as the binary outcome (serum homocysteine $>13\,\mu$mol/l [19] [HYPER_H]). For the Cox model, we chose serum HDL cholesterol as the main continuous exposure (in mmol/l [HDL_CHOL]) and cardiovascular-related death as the event (ICD-10 codes ranging from I20 to I99, after exclusion of I52–I59 and I79 codes [CVD]) in addition to time from examination to event/censors (in months [SURVIVAL_T]).

*6.2. Number and location of knots, and reference values for the main continuous exposures*

Both the number and the location of knots have been arbitrarily chosen to illustrate the SAS macro, although also driven in part by previous recommendations [24]. To assess the dose-response association between calcium intake and BMD, CALCIUM was coded using an RCS function with three knots, located at the 5th, 50th, and 95th percentiles. The reference value for CALCIUM was chosen to be the median value. To assess the dose-response association between folate intake and presence of HHcy, FOLATE was coded using an RCS function with five knots, located at the 5th, 25th, 50th, 75th, and 95th percentiles. The reference value for FOLATE has been set at the estimated average requirement, i.e. 0.320 mg/day. To assess the dose-response association between HDL cholesterol and cardiovascular-related mortality, HDL_CHOL was coded using an RCS function with four knots, located at 0.8, 1.0, 1.5, and 2.0 mmol/l. We chose arbitrary values of the knots instead of resorting to the percentiles of the distribution of HDL cholesterol only to present the use of the AVK_MSV parameter (see Appendix A) of the SAS macro. These values had no biological significance; they approximately corresponded to the 5th, 25th, 75th, and 95th percentiles of HDL_CHOL distribution. The reference value for HDL_CHOL was set at 1 mmol/l, the lowest clinical threshold for low HDL cholesterol [36].

*6.3. Potential confounders included in models*

Since the purpose of this paper was to illustrate well-known associations with the SAS macro, and neither to investigate original associations nor to try to identify the exact relation, we chose a set of classical potential confounders, whatever the model under study: sex (SEX), race/ethnicity (White non-Hispanic [WHITE_NH], Black non-Hispanic [BLACK_NH], Hispanic [HISPANIC], other race/ethnicity [ETHN_OTHER]), smoking status (current versus no current smoker [CURR_SMOKER]), education ($<$12 years [INF_HS], 12 years [HS], $>$12 years [COLLEGE]), and age (in a 10-year unit [AGE]). The reference categories for race/ethnicity and education were WHITE_NH and HS, respectively. AGE was included using an RCS function with three knots located at the 5th, 50th, and 95th percentiles, in order to minimize residual confounding with more parsimony than with an RCS function with five knots. To present the use of some parameters of the SAS macro, we performed one additional Cox model for the HDL cholesterol and cardiovascular-related mortality analysis that included body mass index (in $kg/m^2$ [BMI]), besides the potential confounders cited above. BMI was coded using an RCS function with four knots arbitrarily located at 20, 23, 30, and 38 $kg/m^2$; these values approximately corresponded to the 5th, 25th, 75th, and 95th percentiles of BMI distribution.

*6.4. Study sample*

We selected all NHANES III examinees aged 20 years and older, who were included in the Linked Mortality File with at least 1 month of follow-up, and with no missing values on calcium and folate intake, HDL cholesterol, and on all potential confounders cited above. Pregnant women were excluded from the analyses. The total sample size was 14 757 individuals. Of these individuals, BMD measurements were available for 13 407 individuals and serum homocysteine was assayed among 6641 individuals.

## 7. Results

Let Initial_datafile be the original SAS datafile that contained all the necessary variables for the analyses, located at D:\RCS\Data. The splines of the continuous exposures were created on data with non-missing value on BMD (for calcium intake and BMD analysis) or on serum homocysteine (for folate intake and HHcy analysis).

*7.1. Calcium intake and BMD*

Among the 13 407 individuals, median calcium intake was 0.623 g/day and the median BMD at the femur neck region was 0.827 $g/cm^2$ (Table II). Besides the graph of the adjusted dose-response association between CALCIUM and BMD_FEMUR, we also wanted values of the difference in BMD_FEMUR for specific values of CALCIUM such as 0.5, 1.0, and 1.5 g/day. The SAS code was as follows:

```
%RCS_Reg(  infile= initial_datafile,  outfile= out_fin,
           Dir_data= "D:\RCS\Data",    where= BMD_FEMUR ne .,
           Main_spline_var= CALCIUM,
           Oth_spline_var1= AGE,
           typ_reg= lin,     dep_var= BMD_FEMUR,
           adjust_var=       SEX BLACK_NH HISPANIC ETHN_OTHER CURR_SMOKER
                             INF_HS COLLEGE,
  Y_ref_line= 1,             max_Xaxis= 2,
  specif_val= 0.5 1.0 1.5);
```

There was no need to specify the number ($K=3$) and the location (5th, 50th, and 95th percentiles) of knots for the splines of CALCIUM and AGE since they are the default values in the SAS macro. The SAS macro provided in the SAS LOG window the values of

**Table II**. Study populations according to the type of analyses, from NHANES III data.

| Characteristics | Linear analysis ($N=13407$) | Logistic analysis ($N=6641$) | | Survival analysis ($N=14757$) | |
|---|---|---|---|---|---|
| | | No HHcy ($N=5625$) | HHcy ($N=1016$)† | No CVD death ($N=13782$) | CVD death ($N=975$)† |
| Women, *n* (per cent) | 6926 (52) | 3297 (59) | 422 (42) | 7337 (53) | 424 (43) |
| Age (years)* | 45 (32-64) [20-90] | 41 (30–60) [20–90] | 63 (41–74) [20–90] | 43 (31–61) [20–90] | 76 (67–82) [20–90] |
| BMI ($kg/m^2$)* | | | | 26.3 (23.2–30.1) [11.7–79.6] | 25.9 (22.8–29.2) [14.4–67.3] |
| Race/ethnicity, *n* (per cent) | | | | | |
| White non-Hispanic | 5752 (43) | 2020 (36) | 469 (46) | 5585 (41) | 612 (63) |
| Black non-Hispanic | 3556 (27) | 1669 (30) | 292 (29) | 3772 (27) | 194 (20) |
| Hispanic | 3578 (27) | 1661 (30) | 211 (21) | 3866 (28) | 156 (16) |
| Other ethnicity | 521 (4) | 275 (5) | 44 (4) | 559 (4) | 13 (1) |
| Current smoker, *n* (per cent) | 3484 (26) | 1304 (23) | 289 (28) | 3644 (26) | 167(17) |
| Education, *n* (per cent) | | | | | |
| <12 years | 5361 (40) | 2113 (38) | 499 (49) | 5380 (39) | 590 (61) |
| 12 years | 4151 (31) | 1791 (32) | 288 (28) | 4324 (31) | 207 (21) |
| >12 years | 3895 (29) | 1721 (31) | 229 (23) | 4078 (30) | 178 (18) |
| Calcium intake (g/day)* | 0.623 (0.390–0.967) [0.002–9.870] | | | | |
| BMD ($g/cm^2$)* | 0.827 (0.717–0.941) [0.231–1.841] | | | | |
| Folate intake (mg/day)* | | 0.223 (0.138–0.352) [0.003–2.783] | 0.194 (0.123–0.291) [0–3.929] | | |
| HDL cholesterol (mmol/l)* | | | | 1.27 (1.03–1.53) [0.21–5.07] | 1.24 (1.01–1.55) [0.31–4.32]¶ |
| Follow-up (months)* | | | | 105 (87–124) [1–145] | 57 (31–85) [1–139] |

BMD, bone mineral density at the femur neck region; HDL, high-density lipoprotein; HHcy, hyperhomocysteinemia; *median (Interquartile range) [range]; †all differences were significant with *p*-values <0.01, except when the sign '¶' is present.





Statistics in Medicine

the knots for the two spline variables included into the regression, the reference value for CALCIUM (0.623), and the name of the created spline variables:

```
---------------------------- Miscellaneous information ----------------------------
Values of the 3 knots of CALCIUM (choice according to the percentiles):
Knot #1: 0.176     (5 th percentile)
Knot #2: 0.623     (50 th percentile)
Knot #3: 1.731     (95 th percentile)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
Values of the 3 knots of AGE (choice according to the percentiles):
Knot #1: 2.258      (5 th percentile)
Knot #2: 4.5       (50 th percentile)
Knot #3: 8.133      (95 th percentile)

Spline variables have been created in the "out_fin" SAS datafile, located in the SAS Working library
The reference value for CALCIUM has been assigned to its median (i.e., 0.623)
- - - - - - -
Name of the 2 spline variables of CALCIUM that have just been created:
Spline #1 = CALCIUM_RCS_lin
Spline #2 = CALCIUM_RCS_S1
- - - - - - -
Name of the 2 spline variables of AGE that have just been created:
Spline #1 = AGE_RCS_lin
Spline #2 = AGE_RCS_S1
- - - - - - -
The AIC value for the model is: -16274.81616
------------------------- End of miscellaneous information ------------------------
```

The results of the regression are presented in Table III. In this model, the two splines of calcium intake (CALCIUM_RCS_lin and CALCIUM_RCS_S1) were significantly different from 0. We cannot interpret the value of −0.0141 for CALCIUM_RCS_S1, whereas the value of 0.0258 for CALCIUM_RCS_lin is simply the slope in g/cm$^2$ of BMD per g/day of calcium intake for values of calcium intake lower than the first knot (here, 0.176 g/day): BMD increased of 0.0258 g/cm$^2$ per 1-g/day increase in calcium intake when calcium intake was lower than 0.176 g/day. Of note, the estimate for AGE_RCS_S1 was significantly different from 0, which means that the association between age and BMD was significantly not linear.

In the SAS OUTPUT window, the SAS macro also provided tests for overall and non-linear associations between calcium intake and BMD (Table IV). The test for the overall association between CALCIUM and BMD was significant, which means that, whatever

**Table III**. Estimates from the linear regression investigating associations with bone mineral density.

| | Using %RCS_REG SAS Macro | | | Using PROC GAM | | |
|---|---|---|---|---|---|---|
| Exposure | Estimate (SE) | Chi-sq | Pr>Chi-sq | Estimate (SE) | *t* value | Pr>Chi-sq |
| CALCIUM_RCS_lin | 0.0258 (0.0066) | 15.2 | <0.01 | | | |
| CALCIUM_RCS_S1 | −0.0141 (0.0044) | 10.1 | <0.01 | | | |
| AGE_RCS_lin | −0.0370 (0.0019) | 393.7 | <0.01 | | | |
| AGE_RCS_S1 | −0.0004 (0.0001) | 24.4 | <0.01 | | | |
| SEX | −0.0853 (0.0024) | 1314.1 | <0.01 | −0.0854 (0.0024) | −36.35 | <0.01 |
| BLACK_NH | 0.1033 (0.0030) | 1183.1 | <0.01 | 0.1031 (0.003) | 34.62 | <0.01 |
| HISPANIC | 0.0294 (0.0031) | 87.3 | <0.01 | 0.0298 (0.0031) | 9.48 | <0.01 |
| ETHN_OTHER | 0.0135 (0.0061) | 4.9 | 0.03 | 0.0136 (0.0061) | 2.23 | 0.03 |
| CURR_SMOKER | −0.0226 (0.0027) | 69.3 | <0.01 | −0.0222 (0.0027) | −8.25 | <0.01 |
| INF_HS | −0.0010 (0.0029) | 0.1 | 0.74 | −0.0014 (0.0029) | −0.48 | 0.63 |
| COLLEGE | −0.0045 (0.0030) | 2.3 | 0.13 | −0.0042 (0.003) | −1.41 | 0.16 |
| Linear (CALCIUM) | | | | 0.0060 (0.0023) | 2.65 | 0.01 |
| Spline (CALCIUM) | | | | NP | 9.56* | <0.01 |
| Linear (AGE) | | | | −0.0457 (0.0007) | −67.23 | <0.01 |
| Spline (AGE) | | | | NP | 22.13* | <0.01 |

SE, standard error; NP, not provided by PROC GAM; * Chi-square value with 1 DF.

**Table IV**. Tests for overall and non-linear associations between calcium intake and bone mineral density.

| Contrast | DF | Chi-square | Pr>Chi-sq |
|---|---|---|---|
| Overall_association | 2 | 17.15 | <0.01 |
| Non_lin_association | 1 | 10.13 | <0.01 |

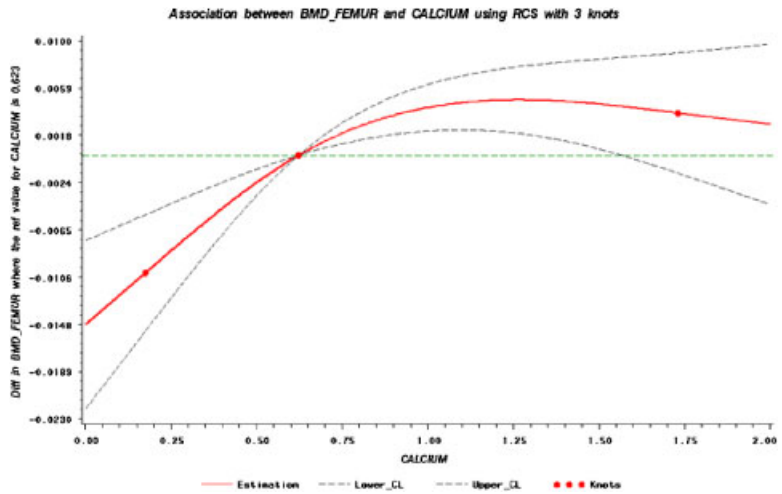


**Figure 2**. Adjusted dose-response association between calcium intake (g/day) and bone mineral density (BMD) in femur neck region (g/cm$^2$). Calcium intake was coded using an RCS function with three knots located at the 5th, 50th, and 95th percentiles of the distribution of calcium intake. *Y*-axis represents the difference (Diff) in BMD between individuals with any value of calcium intake with individuals with 0.623 g/day of calcium intake. Dashed lines are 95 per cent confidence intervals. Knots are represented by dots.

the shape of the association, calcium intake was significantly associated with BMD. The test for the non-linear association was also significant, which means that the association between calcium intake and BMD was significantly not linear. Of note, the value of the $\chi^2$ test for the non-linear association (10.13) was the same as the one for the $S_1$ spline of CALCIUM (Table III): with an RCS function with three knots, testing the non-linear association between the exposure and the outcome is equivalent to testing the null hypothesis that the value of the estimate of the $S_1$ spline equals 0.

Specifying 'SPECIF_VAL=0.5 1.0 1.5' provided, in the SAS OUTPUT window, the values of the adjusted differences in BMD_FEMUR, when comparing individuals with 0.5, 1.0, and 1.5 g/day of calcium intake with individuals with the reference value of calcium intake (0.623 g/day):

```
------------------------------------------------------------------------------------
List of the Differences [95% CI] for specific values of CALCIUM (ref value for CALCIUM: 0.623)
CALCIUM   Differences     Lower_CL     Upper_CL
 0.5    -.002389722   -.003543868   -.001235577
 1.0    0.004139711   0.002105576   0.006173846
 1.5    0.004457122   0.000530411   0.008383834
```

Compared to individuals with 0.623 g/day of calcium intake, individuals with 0.5, 1.0, and 1.5 g/day of calcium intake had differences of −0.0024 (95 per cent CI: −0.0035, −0.0012), +0.0041 (95 per cent CI: +0.0021, +0.0062), and +0.0045 (95 per cent CI: +0.0005, +0.0084) g/cm$^2$ in BMD at the femur neck region, respectively.

Since the maximum value of calcium intake was 9.870 g/day whereas the 95th percentile was 1.731 g/day, we specified MAX_XAXIS=2 to have the dose-response association displayed in the optimal way in the SAS GRAPH window (i.e. with the maximum value of 2 g/day on the *X*-axis). Figure 2 shows the dose-response association between calcium intake and BMD, where 0.623 g/day was the reference value for calcium intake. The reading is straightforward: BMD of individuals with calcium intake up to ∼1.5 g/day significantly differed from that of individuals with calcium intake of 0.623 mg/day.

### *7.2. Folate intake and HHcy*

Among individuals with no missing data on serum homocysteine levels, 1016 (15 per cent) presented HHcy (serum homocysteine >13 μmol/l). Individuals with HHcy differed significantly from individuals with no HHcy on all potential confounders (Table I).

**Table V**. Estimates from the logistic regression investigating associations with presence of hyperhomocysteinemia.

| | Using %RCS_REG SAS Macro | | | Using PROC GAM | | |
|---|---|---|---|---|---|---|
| Exposure | Estimate (SE) | Chi-sq | Pr>Chi-sq | Estimate (SE) | *t* value | Pr>Chi-sq |
| FOLATE_RCS_lin | −5.147 (1.963) | 6.87 | <0.01 | | | |
| FOLATE_RCS_S1 | 170.973 (107.635) | 2.52 | 0.11 | | | |
| FOLATE_RCS_S2 | −416.258 (253.164) | 2.7 | 0.10 | | | |
| FOLATE_RCS_S3 | 328.765 (186.616) | 3.1 | 0.08 | | | |
| SEX | −0.853 (0.076) | 125.67 | <0.01 | −0.8472 (0.075) | −11.3 | <0.01 |
| BLACK_NH | −0.005 (0.094) | <0.01 | 0.95 | −0.0183 (0.0922) | −0.2 | 0.84 |
| HISPANIC | −0.321 (0.104) | 9.56 | <0.01 | −0.3350 (0.103) | −3.25 | <0.01 |
| ETHN_OTHER | −0.111 (0.182) | 0.37 | 0.54 | −0.1260 (0.1811) | −0.7 | 0.49 |
| CURR_SMOKER | 0.479 (0.086) | 31.43 | <0.01 | 0.4625 (0.0847) | 5.46 | <0.01 |
| INF_HS | 0.083 (0.091) | 0.83 | 0.36 | 0.0985 (0.0902) | 1.09 | 0.28 |
| COLLEGE | −0.151 (0.1) | 2.28 | 0.13 | −0.1483 (0.0999) | −1.49 | 0.14 |
| AGE_RCS_lin | 0.048 (0.065) | 0.54 | 0.46 | | | |
| AGE_RCS_S1 | 0.014 (0.003) | 24.98 | <0.01 | | | |
| Linear (AGE) | | | | 0.3616 (0.0212) | 17.03 | <0.01 |
| Spline (AGE) | | | | NP | 21.76* | <0.01 |
| Linear (FOLATE) | | | | −1.2101 (0.1818) | −6.66 | <0.01 |
| Spline (FOLATE) | | | | NP | 22.69† | <0.01 |

SE, standard error; NP, not provided by PROC GAM; * Chi-square test with 1 DF; † Chi-square test with 3 DF.

**Table VI**. Tests for overall and non-linear associations between folate intake and presence of hyperhomocysteinemia.

| Contrast | DF | Chi-square | Pr>Chi-sq |
|---|---|---|---|
| Overall_association | 4 | 67.46 | <0.01 |
| Non_lin_association | 3 | 23.89 | <0.01 |

Median folate intake among HHcy individuals was 0.194 mg/day compared with 0.223 mg/day for non-HHcy individuals ($p<0.01$). Besides the graph of the adjusted dose-response association between FOLATE and HHcy, we also wanted values of the association for specific values of FOLATE such as 0.100, 0.200, and 0.500 mg/day. The SAS code was as follows:

```
%RCS_Reg( infile= initial_datafile,     outfile= out_fin,
          Dir_data= ''D:\RCS\Data'',      where= HYPER_H ne.,
          Main_spline_var= FOLATE, ref_val= 0.320,      knots_msv= 5 25 50 75 95,
         Oth_spline_var1= AGE,
         typ_reg= log,          dep_var= HYPER_H,
         adjust_var=            SEX BLACK_NH HISPANIC ETHN_OTHER CURR_SMOKER
                                INF_HS COLLEGE,
         Y_ref_line= 1,         max_Xaxis= 1,
         specif_val= 0.100 0.200 0.500);
```

The reference value of 0.320 mg/day of folate intake has been assigned by specifying REF_VAL = 0.320. In the SAS LOG window, the macro provided the values of the five knots of folate intake (0.063, 0.136, 0.218, 0.344, and 0.704 mg/day). The results of the logistic regression are presented in Table V.

Although the estimates of the three splines FOLATE_RCS_S1, FOLATE_RCS_S2, and FOLATE_RCS_S3 were not significantly different from 0 (Table V), the tests for overall and non-linear associations between folate intake and the presence of HHcy were both highly significant (Table VI). Therefore, after adjustment for sex, race/ethnicity, education, smoking status, and age, folate intake was significantly associated with the presence of HHcy, and the dose-response association was significantly not linear.

We may wonder whether an RCS function with four knots or three knots would have led to more adequate models than did an RCS function with five knots. The AIC values for models including FOLATE with a 4-knot RCS function (located at the 5th,

25th, 75th, and 95th percentiles) or with a 3-knot RCS function (located at the 5th, 50th, and 95th percentiles) were 5065.5 and 5064.6, respectively, whereas the AIC value for the above model with five knots was 5064.4, suggesting that the model with a 5-knot RCS function for FOLATE was slightly more adequate. Therefore, the fact that the estimates of spline are not significant (see Table V) does not necessarily imply that the continuous exposure must be model with more parsimony (i.e. with fewer knots).

Of note, the estimate of the linear spline of AGE (AGE_RCS_lin) was not significantly different from zero, whereas the estimate of the second spline of AGE was highly significant (Table V). This result is not paradoxical since the value of the estimate of the first spline of AGE is the slope before the first knot (in a logistic model, this slope is Ln(OR) for a 1-unit increase in the exposure). Therefore, the slope before AGE = 22.5 years was not significantly different from 0, which means that there was no significant association between age and the presence of HHcy for ages between 20 and 22.5 years. For ages greater than 22.5 years, the non-linear component of the association was highly significant.

Specifying 'SPECIF_VAL = 0.100 0.200 0.500' provided, in the SAS OUTPUT window, the value of the adjusted Ln(OR) for presenting HHcy, when comparing individuals with 0.100, 0.200, and 0.500 mg/day of folate intake with individuals with 0.320 mg/day of folate intake:

```
-------------------------------------------------------------------------------------------
List of the Ln_ORs [95% CI] for specific values of FOLATE (ref value for FOLATE: 0.320)
FOLATE    Ln_ORs    Lower_CL   Upper_CL
 0.1     0.48288    0.29260    0.67316
 0.2     0.29008    0.17216    0.40801
 0.5    -0.36615   -0.53566   -0.19665
```

As displayed in Figure 3(a), the SAS macro provided in the SAS GRAPH window Ln(ORs) where 0.320 mg/day was the reference value for folate intake. Had OR values been required in the SAS OUTPUT and GRAPH windows, EXP_BETA = 1 must have been specified. In contrast, the EXP_BETA parameter must be set to 0 (default value) in order to have Ln(ORs) on the *Y*-axis instead of ORs and to be able to visually check the assumption of linearity of the association. As displayed in the SAS OUTPUT window, the *p*-values for both the overall and the non-linear association tests were $<0.01$.

To illustrate the impact of the number and location of knots, Figures 3(b) shows Ln(ORs) when using an RCS function with five knots located at the 5th, 10th, 50th, 90th, and 95th percentiles (*p*-values for both the overall and the non-linear association tests were $<0.01$) and Figure 3(c) shows Ln(ORs) when using an RCS function with three knots located at the 5th, 50th, and 95th percentiles (*p*-values for both the overall and the non-linear association tests were $<0.01$). The shapes of the dose-response association shown in Figures 3(a) and 3(b) were very similar, and slightly differed with the shape shown in Figure 3(c). The AIC value corresponding to Figure 3(b) was 5065.0 (versus 5064.4 and 5064.6, for Figures 3(a) and 3(c), respectively).

### *7.3. HDL cholesterol and time to cardiovascular-related death*

Out of the 14 757 individuals, 975 (7 per cent) died from a cardiovascular event. These individuals differed significantly from the remainders on all potential confounders, but not on HDL cholesterol (median, 1.24 mmol/l versus 1.27 mmol/l, respectively; Table I). For this analysis, we wanted to display HRs (as opposed to Ln(HRs)) on the graph representing the dose-response association with HDL cholesterol; we also wanted values of the association for specific values of HDL_CHOL such as 0.50, 0.75, 1.50, 2.00, and 2.50 mmol/l. The SAS code was as follows:

```
%RCS_Reg(  infile= initial_datafile,  outfile= out_fin,
           Dir_data= "D:\RCS\Data",
           Main_spline_var= HDL_CHOL,   ref_val= 1.0,   AVK_msv= 1,   knots_msv= 0.8 1.0 1.5 2.0,
           Oth_spline_var1= AGE,
           typ_reg= cox,
           dep_var= CVD,     surv_time_var= SURVIVAL_T,
           adjust_var=       SEX BLACK_NH HISPANIC ETHN_OTHER CURR_SMOKER
                             INF_HS COLLEGE,
           Y_ref_line= 1,    max_Xaxis= 3,
           exp_beta= 1,      round= 0.01,
           specif_val= 0.50 0.75 1.50 2.00 2.50);
```

The results of the regression are presented in Table VII. As displayed in the SAS OUTPUT window, the *p*-values for both the overall and the non-linear association tests were $<0.01$. Therefore, after adjustment for sex, race/ethnicity, education, smoking status, and age, HDL cholesterol was significantly associated with cardiovascular-related death, and the dose-response association was significantly not linear.

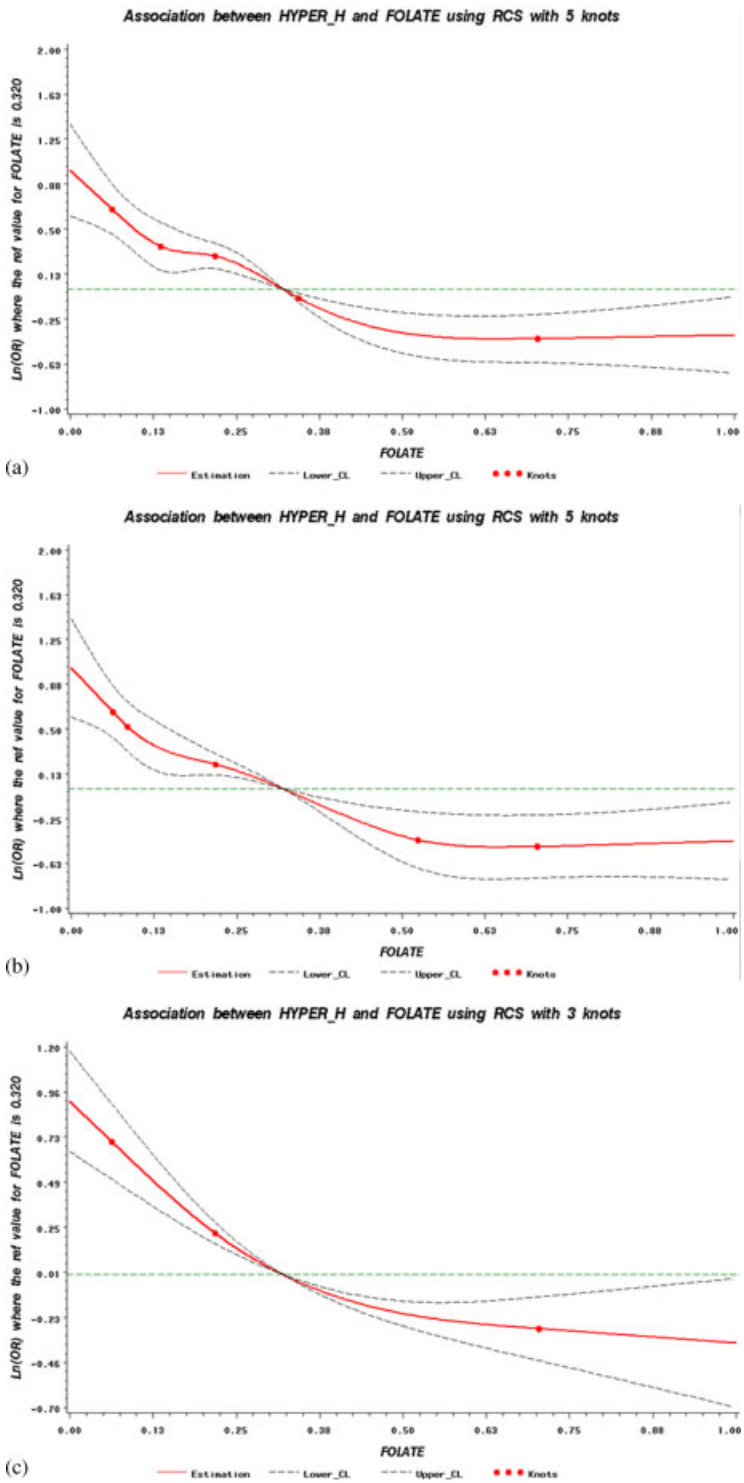


**Figure 3**. Adjusted dose-response association between folate intake and the presence of hyperhomocysteinemia, according to different locations and numbers of knots chosen for coding folate intake with the RCS function: (a) with five knots located at the 5th, 25th, 50th, 75th, and 95th percentiles; (b) with five knots located at the 5th, 10th, 50th, 90th, and 95th percentiles; (c) with three knots located at the 5th, 50th, and 95th percentiles. *Y*-axis represents the Ln(Odds Ratio) to present hyperhomocysteinemia for any value of folate intake compared to individuals with 0.320 mg/day of folate intake. Dashed lines are 95 per cent confidence intervals. Knots are represented by dots.

**Table VII**. Estimates from the Cox model investigating associations with cardiovascular mortality.

| Exposure | Estimate (SE) | Chi-square | Pr>Chi-sq |
|---|---|---|---|
| HDL_CHOL_RCS_lin | −1.37 (0.38) | 12.85 | <0.01 |
| HDL_CHOL_RCS_S1 | 4.50 (1.64) | 7.52 | 0.01 |
| HDL_CHOL_RCS_S2 | −6.54 (2.58) | 6.44 | 0.01 |
| SEX | −0.45 (0.07) | 42.75 | <0.01 |
| BLACK_NH | 0.14 (0.09) | 2.43 | 0.12 |
| HISPANIC | −0.10 (0.10) | 1.03 | 0.31 |
| ETHN_OTHER | −0.72 (0.28) | 6.59 | 0.01 |
| CURR_SMOKER | 0.33 (0.09) | 13.59 | <0.01 |
| INF_HS | 0.17 (0.08) | 3.86 | 0.05 |
| COLLEGE | −0.15 (0.1) | 2.05 | 0.15 |
| AGE_RCS_lin | 0.72 (0.12) | 33.16 | <0.01 |
| AGE_RCS_S1 | 0.009 (0.004) | 5.10 | 0.02 |

SE, standard error.

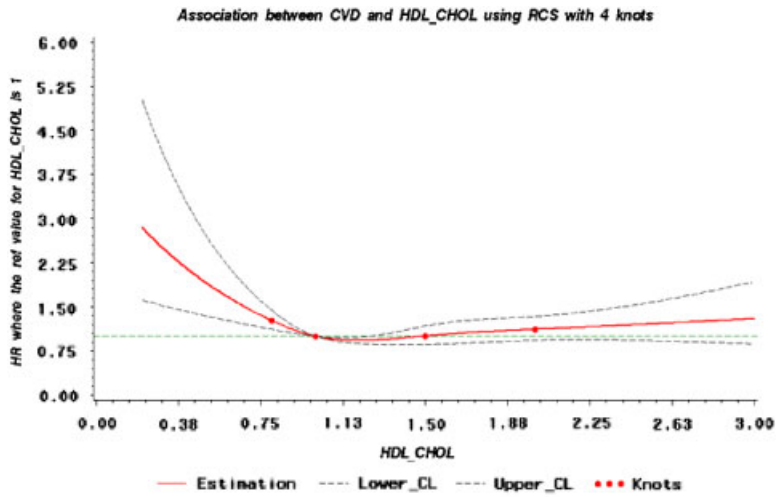


**Figure 4**. Adjusted dose-response association between serum high-density level (HDL) cholesterol and risk for cardiovascular-related death. HDL cholesterol was coded using an RCS function with four knots arbitrarily located at 0.8, 1.0, 1.5, and 2.0 mmol/l, which approximately corresponded to the 5th, 25th, 75th, and 95th percentiles of HDL cholesterol distribution. *Y*-axis represents the adjusted hazard ratio for cardiovascular-related death for any value of HDL cholesterol compared to individuals with 1 mmol/l of HDL cholesterol. Dashed lines are 95 per cent confidence intervals. Knots are represented by dots.

Specifying AVK_MSV=1 allowed an arbitrary definition of the location of the knots (instead of a definition according to the percentiles by default). Specifying EXP_BETA=1 enabled to provide HRs (instead of Ln(HRs) by default) in the SAS GRAPH window (see Figure 4, where 1.0 mmol/l was the reference value for HDL cholesterol since REF_VAL=1.0). Had we set the PRINT_OR_HR parameter to 1, a list of all of the HRs that were used for the graph would have been printed in the SAS OUTPUT window.

Specifying 'SPECIF_VAL=0.50 0.75 1.50 2.00 2.50' with EXP_BETA=1 provided, in the SAS OUTPUT window, the value of the adjusted HRs for cardiovascular-related death, when comparing individuals with HDL cholesterol of 0.50, 0.75, 1.50, 2.00, and 2.50 mmol/l with individuals with HDL cholesterol of 1 mmol/l. The parameter ROUND rounded the values of the HRs displayed in the SAS OUTPUT window:

```
-------------------------------------------------------------------------------
List of the HRs [95% CI] for specific values of HDL_CHOL (ref value for HDL_CHOL: 1.0)
HDL_CHOL    HRs   Lower_CL   Upper_CL
  0.50     1.91     1.35       2.71
  0.75     1.36     1.15       1.60
  1.50     1.00     0.86       1.17
  2.00     1.12     0.93       1.34
  2.50     1.20     0.93       1.56
```

Had we wanted to additionally adjust for BMI by using an RCS function with four knots arbitrarily located at 20, 23, 30, and 38 kg/m$^2$, the SAS code would have been as follows:

```
%RCS_Reg(  infile= initial_datafile,   outfile= out_fin,
           Dir_data= "D:\RCS\Data",
           Main_spline_var= HDL_CHOL,  ref_val= 1.0,     AVK_msv= 1,    knots_msv= 0.8 1.0 1.5 2.0,
           Oth_spline_var1= AGE,
           Oth_spline_var2= BMI,                         AVK_osv2= 1,   knots_osv2= 20 23 30 38,
           typ_reg= cox,
           dep_var= CVD,   surv_time_var= SURVIVAL_T,
           adjust_var=     SEX BLACK_NH HISPANIC ETHN_OTHER CURR_SMOKER
                           INF_HS COLLEGE,
           Y_ref_line= 1,  max_Xaxis= 3,
           exp_beta= 1,    round= 0.01,
           specif_val= 0.50 0.75 1.50 2.00 2.50);
```

The new HRs for the specific values of HDL_CHOL were similar to the ones without this additional adjustment for BMI, although slightly lower for higher values of HDL_CHOL:

```
-------------------------------------------------------------------------------------------
List of the HRs [95% CI] for specific values of HDL_CHOL (ref value for HDL_CHOL: 1.0)
HDL_CHOL    HRs   Lower_CL   Upper_CL
  0.50     1.92      1.35       2.73
  0.75     1.36      1.15       1.60
  1.50     0.97      0.83       1.14
  2.00     1.04      0.86       1.26
  2.50     1.09      0.83       1.43
```

## 8. GAM SAS procedure versus %RCS_REG SAS macro

### *8.1. Brief overview of GAM SAS procedure*

The GAM SAS procedure is an experimental procedure in SAS V9.1.3 that fits generalized additive models defined by Hastie and Tibshirani [37]. These models provide a flexible method for identifying non-linear associations between a continuous exposure and an outcome. They enable the mean of the outcome to depend on an additive predictor through a non-linear link function. Many widely used statistical models belong to this general class; they include additive models for Gaussian data, non-parametric logistic models for binary data, and non-parametric log-linear models for Poisson data.

Let $Y$ be a response random variable and $X_1$, $X_2$,…,$X_p$ be a set of exposures. A regression procedure can be viewed as a method for estimating the expected value of $Y$ given the values of $X_1$, $X_2$,…,$X_p$. The additive model generalizes the linear model by modeling the conditional expectation as

$$E(Y|X_1, X_2, \ldots, X_p) = \alpha + \sum_{i=1}^{p} \delta_i \cdot S_i(X_i)$$

where $S_i$ are smooth functions that will not be given a parametric form but instead will be estimated in a non-parametric fashion [37]. The GAM SAS procedure proposes three smoothing techniques: LOESS (local regression), cubic smoothing splines, and thin-plate smoothing splines. For these three techniques, the number of DF must be assigned (4 being the default number). When choosing the cubic smoothing spline technique, the total number of DF breaks down into one DF for the linear part and the remaining DF being for the smoothing non-linear component. Therefore, an RCS function with $K$ knots (1 DF for the linear part, and $K-2$ DF for the smoothing part) would be equivalent in terms of DF to a cubic smoothing spline technique with a total of $K-1$ DF.

The GAM SAS procedure provides in the SAS OUTPUT window a test for the linear part of the association between the continuous exposure and the outcome, as well as a test for the smoothing spline component. The experimental output delivery system (ODS) GRAPHICS statements are available with GAM SAS procedure. They provide in an additional file the graph displaying the values of the smoothing component (with their 95 per cent confidence interval) in the $Y$-axis according to the values of the continuous exposure in the $X$-axis. This graph, therefore, allows to visually check a potential departure from the linearity assumption. The main advantage of PROC GAM is its iterative process that selects the smoothing components according to the requested number of DF, without any intervention from the user. The main drawbacks of PROC GAM are the following: (i) it is not possible to display and/or quantify the association between the continuous exposure and the outcome (because SAS does not provide the estimated value of the parameters of the cubic smoothing spline function), and (ii) it is not available for the Cox models.

In the next two subsections, PROC GAM will be used to perform the same linear and logistic analyses that have been previously performed with %RCS_REG in this paper. In a purpose of comparison, the number of DF of the continuous potential confounder (age) and the main exposure (calcium and folate intakes for the linear and logistic analyses, respectively) will be chosen to be the same in PROC GAM and %RCS_REG, and the smoothing technique when using PROC GAM will be the cubic smoothing spline technique. Therefore, age will be included using the cubic smoothing spline technique with 2 DF; in the linear analysis, calcium intake will be included using the cubic smoothing spline technique with 2 DF; in the logistic analysis, folate intake will be included using the cubic smoothing spline technique with 4 DF. The ODS HTML and GRAPHICS statements will be used to display the values of the smoothing components according to the continuous exposure. The results for %RCS_REG for linear and logistic analyses are presented in Tables III and IV, and in Tables V and VI, respectively. The curve of the dose-response association provided by the %RCS_REG macro has been obtained without assigning a specific value to the MAX_XAXIS parameter (in order to visually have the curve for all values of the exposure, like the graph provided by PROC GAM).

### *8.2. Calcium intake and BMD analysis by using PROC GAM and %RCS_REG*

The SAS code for PROC GAM was as follows:

```
LIBNAME lib_rcs "D:\RCS\Data";
ods html;
ods graphics on;
proc gam data=lib_rcs.initial_datafile plots(clm);
model BMD_FEMUR =  param(SEX BLACK_NH HISPANIC ETHN_OTHER CURR_SMOKER
                   INF_HS COLLEGE) spline(AGE,df=2) spline(CALCIUM,df=2)
                   / dist = Gaussian ;
run;
ods graphics off;
ods html close;
```

The results for the regression model components when using PROC GAM are presented in Table III. Of note, PROC GAM provides Student's *t*-tests for the estimates (except for the spline component) whereas %RCS_REG provides chi-square tests. These results for the potential confounders are virtually identical to those obtained by using RCS functions. The value of the estimate of Linear(AGE) (−0.0457) must not be compared with the one for AGE_RCS_lin (−0.0370) since their interpretations are different. When using PROC GAM, the value of the estimate of Linear(AGE) is the linear part of the association between AGE and BMD on the overall range of AGE values, whereas the value of the estimate of AGE_RCS_lin is the (linear) slope quantifying the association between AGE and BMD but only before the first knot of AGE (22.58 years).

The results for the smoothing model components when using PROC GAM are also presented in Table III. The cubic smoothing components are significantly different from 0 for both AGE and CALCIUM, which means that there was a significant non-linear association between AGE or CALCIUM and BMD. The statistical tests for the non-linear component of CALCIUM when using PROC GAM can be compared with the 'Non_lin_association' statistical test provided by the %RCS_REG macro, since they both test the departure from linearity with the same number of DF (here, 1 DF): $\chi^2 = 9.56$ for 'Spline(CALCIUM)' (Table III) and 10.13 for 'Non_lin_association' for CALCIUM when using %RCS_REG (Table IV).

The graph displayed in Figure 5 provided by PROC GAM is somewhat difficult to compare with the graph displayed in Figure 6 by %RCS_REG. To obtain the overall shape of the association between calcium intake and BMD shown in Figure 6 by using the results and figure provided by PROC GAM, one needs to mentally add the shape of the values of the cubic smoothing component displayed in Figure 5 with the estimate of the linear effect (slope) of calcium (i.e, +0.006 in Table III). This mental exercise is unfortunately necessary since PROC GAM does not provide the table of the values of the cubic smoothing component displayed in Figure 5, at least with SAS V9.1.3.

### *8.3. Folate intake and HHcy analysis by using PROC GAM and %RCS_REG*

The SAS code for PROC GAM was as follows:

```
LIBNAME lib_rcs "D:\RCS\Data";
ods html;
ods graphics on;
proc gam data=lib_rcs.initial_datafile plots(clm);
model HYPER_H =   param(SEX BLACK_NH HISPANIC ETHN_OTHER CURR_SMOKER
                  INF_HS COLLEGE) spline(AGE,df=2) spline(folate,df=4)
                  / dist = binomial;
run;
ods graphics off;
ods html close;
```

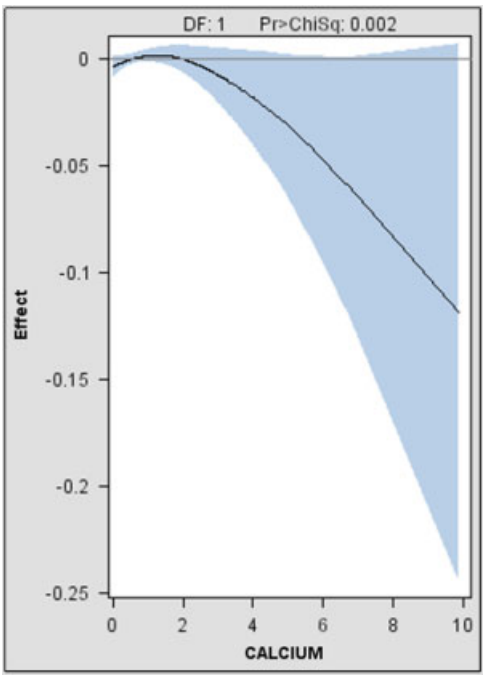


**Figure 5**. Values of the cubic smoothing component of CALCIUM in the linear regression model according to values of CALCIUM, by using PROC GAM. The shape refers to 95 per cent confidence interval.

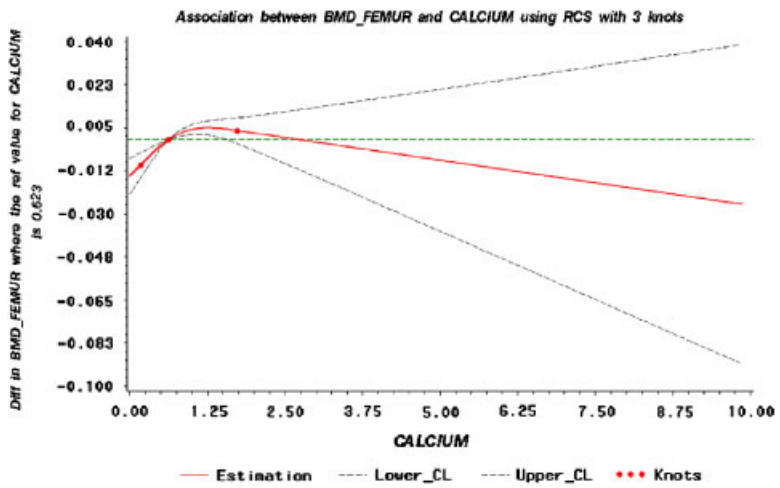


**Figure 6**. Dose-response association between calcium intake and bone mineral density, by using %RCS_REG macro. Dashed lines refer to 95 per cent confidence interval. Knots are represented by dots.

The results for the regression model components when using PROC GAM are presented in Table V. These results for the potential confounders are virtually identical to those obtained by using RCS functions.

The results for the smoothing model components when using PROC GAM are also presented in Table V. The cubic smoothing components are significantly different from 0 for both AGE and FOLATE, which means that there was a significant non-linear association between AGE or FOLATE and presence of HHcy. The statistical test for the non-linear component of FOLATE when using PROC GAM can be compared with the 'Non_lin_association' statistical test provided by the %RCS_REG macro, since they both test the departure from linearity with the same number of DF (here, 3 DF): $\chi^2 = 22.69$ for 'Spline(FOLATE)' (Table V) and 23.89 for 'Non_lin_association' for FOLATE when using %RCS_REG (Table VI).

## 9. Concluding remarks

RCS functions are useful tools to characterize a dose-response association between a continuous exposure and an outcome for *a priori* non-linear associations [11, 38], and/or to check the assumption of linearity before introducing a continuous exposure in

a model with appropriate recoding. Their use has also been recommended when adjusting for continuous exposures in order to minimize residual confounding [15]. Here, we provided an SAS macro that (i) creates RCS functions with the usual recommended numbers of knots (3 to 5, [23]) of which the location can be set freely (i.e. according to either percentiles or arbitrary values, all chosen by the user), and (ii) includes the splines of an RCS function in the most used regression models in epidemiology, such as linear, logistic, and Cox models, as well as in GEE linear and logistic models for repeated measurements. Eventually, the SAS macro provides the estimated dose-response association in the SAS GRAPH window (with its 95 per cent confidence interval), with statistical tests for overall and non-linear associations, and (optionally) values of the association for a list of specified values of the main continuous exposure in the SAS OUTPUT window.

One disadvantage of RCS functions is that the shape of the dose-response association generally depends on the location and the number of knots. However, the location of knots has a little impact on the shape of the dose-response association [1], as illustrated here (Figures 3(a) and 3(b). Durrleman and Simon proposed the 5th, 25th, 50th, 75th, and 95th percentiles for a 5-knot RCS function, the 5th, 25th, 75th, and 95th percentiles for a 4-knot RCS function, and the 5th, 50th, and 95th percentiles for a 3-knot RCS function [24] and these locations are the most widely used with RCS functions. Conversely, the number of knots has a non-negligible impact on the shape of the dose-response association. The choice of the adequate number of knots depends on the objective. For statistically testing the assumption of linearity of a dose-response association, a 3-knot RCS function, which is much smoother than a 5-knot RCS function, would be more powerful to detect departure from linearity (Wald $\chi^2$ test with 1 DF compared with 3 DF for a 5-knot RCS function). For adjustment for a continuous confounder by using an RCS function, a 3-knot RCS function would also be preferable to a 5-knot RCS function since it is more parsimonious and would remove most of the residual confounding [15]. A 5-knot RCS function would be desirable for explanatory analyses, when one wants to characterize the dose-response association more precisely (i.e. closer to the data), on the *a priori* assumption that the association might be more complex than that accounted by a 3-knot RCS function. In this context, the value of AIC may help to determine if little 'jumps' that can be observed with a 5-knot RCS function would be spurious or relevant. For instance, the values of AIC for Figures 3(a) and 3(c) were 5064.4 and 5064.6, respectively, suggesting a slightly more adequate modeling by using a 5-knot RCS function than a 3-knot RCS function. The value of AIC is provided for all types of regressions proposed by the SAS macro, except for GEE models where AIC methods cannot be used directly [39]. If necessary, strategies to determine the optimal modeling can easily be implemented with serial comparison of AIC values, an objective criteria of fitting.

All our examples were based on a very large data set. In smaller data sets, attention should be paid to adequately choose the variables to be included into a model, and avoid including more variables than the 10th of the number of observations [2]. In this regard, when coding a continuous exposure by using an RCS function with $K$ knots, the user must remember that the number of spline variables (including the linear spline) is equal to $K-1$. Furthermore, when defining knots according to percentiles of the distribution of a continuous exposure, the values of two consecutive knots may be the same for small data sets (for instance, the 5th and the 10th percentiles) or for continuous exposures with a few numbers of different values. In the case of equal consecutive knots defined according to percentiles, we recommend choosing the location of knots according to arbitrary values of the exposure.

We did not provide examples for GEE models since the use of the SAS macro as well as the interpretation of the results in the SAS OUTPUT and GRAPH windows are exactly the same as for usual linear and logistic models. The only differences are the use of SUBJECT_VAR and WORK_CORR_MATRIX parameters.

Other SAS macros dealing with splines (RCSs or B-splines) have already been published, but their use was limited either to Cox models [40] or logistic regressions [41], and, as far as we know, there is no SAS macro dealing with splines for GEE models.

FPs are also useful tools to characterize dose-response associations between a continuous exposure and an outcome [8, 42]. An SAS macro using FPs has already been published for linear, logistic, and Cox models [43]. However, the SAS macro proposed in this latter paper does not display the shape of the dose-response association between the continuous exposure and the outcome, with its 95 per cent confidence intervals.

Our macro %RCS_REG, unlike PROC GAM, enables to (i) investigate dose-response associations with smooth curves in Cox models, and (ii) display and/or quantify the association between the exposure and the outcome, besides providing a visual check for the linearity assumption and a statistical test for this assumption that provides very similar values compared with PROC GAM. One drawback of using %RCS_REG over PROC GAM is that the user must choose the number and the location of knots for the RCS function, while when using PROC GAM with a cubic smoothing spline technique, the user must choose the number of DF only (which would be equivalent to choosing the number of knots for RCS functions). Because the impact of the location of the knots is limited, it is reasonable in most cases to use the classical locations suggested by Durrleman and Simon [24]. Therefore, once the number of knots has been chosen, the location of knots is automatic according to Durrleman and Simon's suggestions, leading finally to a similar ease of use between %RCS_REG and PROC GAM. However, the option to define location of knots at some arbitrary values of the continuous exposure can be very useful (e.g. when specifying threshold values already cited in the literature) and this feature adds some versatility to the %RCS_REG SAS macro.

## Appendix A: Description of the parameters of the RCS_REG SAS macro

The '**' sign indicates that the parameter must always be specified. The '*' sign indicates that the parameter must be specified if a regression model is requested. Otherwise, the parameter is optional. In this case and if the parameter is not specified, the default value is used.

INFILE**: name of the original SAS datafile that contains the continuous exposure(s) that one wants to code using an RCS function.

OUTFILE: name of the new SAS datafile, created in the SAS Working library, that will contain the splines of the continuous exposure(s). If not specified, the name of this outfile datafile is the name of the original SAS datafile followed by '_rcs'.

MAIN_SPLINE_VAR**: name of the main continuous exposure that one wants to code using an RCS function. The main continuous exposure is the continuous exposure for which one (optionally) wants to display the curve of the dose-response association with the outcome. If the name of this variable contains more than 12 characters, its name will be truncated to 11 characters, and the 12th character will be '_'.

AVK_MSV: by default, the knots of the main continuous exposure are defined according to the percentiles of its distribution (AVK_MSV=0). If AVK_MSV is set to 1, the user chooses to define Arbitrary Values of the Knots.

KNOTS_MSV: if AVK_MSV=0, KNOTS_MSV is the list of the values of the percentiles of the distribution of the main continuous exposure. These values must be separated by a space and be provided in the ascendant order (the lowest percentile comes first). For instance, if AVK_MSV=0, 'KNOTS_MSV=5 25 75 95' means that the user chooses the values of the 5th, 25th, 75th, and 95th percentiles of the main continuous exposure as the values of the knots. If AVK_MSV=0, the default list is '5 50 95'. If AVK_MSV=1, KNOTS_MSV must be specified, and it is the list of arbitrary values of the main continuous exposure that define the knots. For instance, if the main continuous exposure is the age of individuals, and if AVK_MSV=1, 'KNOTS_MSV=23 29 45' means that the knots are located at 23, 29, and 45 years old.

OTH_SPLINE_VAR1: first other continuous exposure that one wants to include into the regression model using an RCS function, as a potential confounder. If the name of this variable contains more than 12 characters, its name will be truncated to 11 characters, and the 12th character will be '_'.

AVK_OSV1: same as AVK_MSV for the first other continuous exposure that is specified in OTH_SPLINE_VAR1 parameter.

KNOTS_OSV1: same as KNOTS_MSV for the first other continuous exposure that is specified in OTH_SPLINE_VAR1 parameter.

OTH_SPLINE_VAR2-OTH_SPLINE_VAR10: second to (max) 10th other continuous exposures that one wants to include into the regression model using an RCS function, as potential confounders.

KNOTS_OSV2-KNOTS_OSV10: same as KNOTS_OSV1 for the second to the (max) 10th other continuous exposures that are specified in OTH_SPLINE_VAR2-OTH_SPLINE_VAR10 parameters.

AVK_OSV2-AVK_OSV10: same as AVK_OSV1 for the second to the (max) 10th other continuous exposures that are specified in OTH_SPLINE_VAR2-OTH_SPLINE_VAR10 parameters.

DIR_DATA: name of the input directory that contains the original SAS datafile specified in the INFILE parameter. The name must be typed between quotes (for instance, DIR_DATA='C:\SAS\data'). By default, the input directory is the SAS Working library.

TYP_REG: specifies the type of the regression. Assign 'lin' (without quotes) to perform a linear regression, 'log' for a logistic regression, or 'cox' for a Cox model. If TYP_REG is not specified, the macro only creates the splines of the continuous exposure(s) in the outfile datafile specified in the OUTFILE parameter, without performing any regression.

DEP_VAR*: name of the dependent variable (outcome) in the regression model. For a linear regression, DEP_VAR must be continuous. For a logistic regression, DEP_VAR must value 1 for cases and 0 for controls. For a Cox model, DEP_VAR must value 1 for events and 0 for censored individuals. If the name of this variable contains more than 16 characters, its name will be truncated to 15 characters, and the 16th character will be '_'.

SURV_TIME_VAR: name of the survival time variable in the case of a Cox model. SURV_TIME_VAR must be specified if TYP_REG=cox is specified.

ADJUST_VAR: list of the adjusted variables included into the regression model, except spline variables specified in the OTH_SPLINE_VAR*k* parameters ($k \in \{1,\ldots,10\}$). The list elements must be separated by a space. If a continuous exposure is specified in this ADJUST_VAR parameter, its association with the outcome is assumed to be linear. If ADJUST_VAR and OTH_SPLINE_VAR*k* are not specified, the regression will display an unadjusted dose-response association between the main continuous exposure and the outcome.

PRGM_STATEMENTS_COX: in the case of a Cox model, the macro uses the PROC PHREG SAS procedure. This procedure enables to create new explanatory variables or modify the values of explanatory variables in the MODEL statement of the procedure. The parameter PRGM_STATEMENTS_COX can be used to do this/theses same task(s) when performing a Cox model with the macro (for instance, to create time by covariate interactions).

BY_FACTOR: name of a categorical variable (which must be numerical) used for performing a 'by factor' analysis according to this variable, i.e. an analysis stratified on this categorical variable. If a 'by factor' analysis is requested, the name of the outfile datafile (specified in the OUTFILE parameter) will be followed by '_*k*', where *k* is the rank of the category (the lowest category being the first); there will, therefore, be as many outfile datafiles as the number of categories of the categorical variable.

MISSING_BF: if set to 1, a missing value on the variable specified in BY_FACTOR parameter will be treated as a non-missing value (and assigned to −99). In this case, the first outfile datafile will be for the missing values of the categorical variable specified in the BY_FACTOR parameter. By default, the analysis is not performed for missing data on this variable (MISSING_BF=0).

SUBJECT_VAR: in case of multiple records per subject, the macro can perform linear or logistic GEE models by using the 'REPEATED SUBJECT' command in PROC GENMOD SAS procedure. To perform a GEE model (and only to do so), assign the name of the subject variable.

WORK_CORR_MATRIX: structure of the working correlation matrix in case of GEE models, which will be used by the PROC GENMOD procedure. A list of available matrix structures can be found at http://support.sas.com/onlinedoc/913/docMainpage.jsp.

For instance, WORK_CORR_MATRIX = mdep(3) requests a 3-dependent working correlation matrix. This parameter must be specified if SUBJECT_VAR parameter is specified.

REF_VAL: reference value of the main continuous exposure that will be used to display the dose-response association with the outcome when comparing individuals with any value of the main continuous exposure with individuals with the value assigned in REF_VAL. By default, REF_VAL is the median of the main continuous exposure.

SPECIF_VAL: the macro can provide in the SAS OUTPUT window the estimated value of the association (with its 95 per cent confidence interval) between the main continuous exposure and the outcome for one or more specific values of the exposure (assigned as a list in SPECIF_VAL) when compared with the reference value (assigned in REF_VAL). The list of these specific values must be separated by a space. If SPECIF_VAL is specified, a data set named 'List_specif_values' is created in the SAS Working library. If a 'by factor' analysis is requested, the name of this data file will be followed by '_*k*', where *k* is the rank of the category of the variable specified in BY_FACTOR parameter (the lowest category being the first).

ROUND: specifies if the value of the association provided by using SPECIF_VAL parameter has to be rounded. For instance, if set to 0.01, it means that the value of the association (and the ones in the 95 per cent confidence interval) will be rounded to the nearest 0.01 value when displayed in the SAS OUTPUT window. By default, values are not rounded.

EXP_BETA: when set to 0, the macro displays Ln(ORs) or Ln(HRs); when set to 1, the macro displays ORs or HRs. The default value is 0. Of note, this parameter has no effect if TYP_REG = lin (linear regression).

PRINT_OR_HR: if set to 1 and if EXP_BETA is set to 1, the macro provides in the SAS OUTPUT window ORs or HRs (with their 95 per cent confidence intervals). In this case, a datafile named 'List_or_hr' is created in the SAS Working library. If a 'by factor' analysis is requested, the name of this datafile will be followed by '_*k*', where *k* is the rank of the category of the variable specified in BY_FACTOR parameter (the lowest category being the first). By default, ORs or HRs are not provided (PRINT_OR_HR = 0). This parameter has no effect if EXP_BETA = 0.

WHERE: this parameter enables to create the splines and to perform regressions in a selection of individuals, like the usual WHERE statement in all SAS procedures.

OUTPUT_PRED: if set to 1, the macro provides the predicted value of the outcome for each observation, in the outfile datafile (specified in OUTFILE parameter). This parameter uses the 'OUTPUT OUT' command of PROC GENMOD and PROC PHREG. For the linear regression, the macro provides the predicted value of the continuous outcome as well as its lower and upper 95 per cent confidence limits; for the logistic regression, the macro provides the predicted value of the probability of the binary outcome as well as its lower and upper 95 per cent confidence limits; for the Cox model, the macro provides the predicted value of the survival function $S(t)$ computed by using the product-limit method (default method in the PHREG SAS procedure). By default, the predicted values are not provided (OUTPUT_PRED = 0).

HISTOGRAM: if set to 1, the macro displays the distribution of the main continuous exposure using an histogram (HISTOGRAM statement in the PROC UNIVARIATE procedure). The default value is 0 (no histogram is displayed).

NO_GRAPH: if set to 1, no graph is displayed in the SAS OUTPUT window. The graph is displayed by default (NO_GRAPH = 0).

DISPLAY_KNOTS: if set to 1, the macro displays the knots on the dose-response curve as dots. If set to 0, the knots are not displayed on the curve. The default value is 1.

X_REF_LINE: if set to 1, a vertical dashed green line is displayed to materialize the reference value of the main continuous exposure. There is no vertical line by default (X_REF_LINE = 0).

Y_REF_LINE: if set to 1, a horizontal dashed green line is displayed to materialize the null hypothesis H0. The *Y*-coordinate of the line is 0 when a linear regression is requested, or when a logistic or a Cox model with EXP_BETA=0 is requested. The *Y*-coordinate of the line is 1 when a logistic or a Cox model with EXP_BETA=1 is requested. There is no horizontal line by default (Y_REF_LINE = 0).

MIN_XAXIS: by default, the graph displaying the dose-response curve starts, on the *X*-axis, at the minimum value of the main continuous exposure. Assigning a value to MIN_XAXIS enables the graph to start at this assigned value.

MAX_XAXIS: by default, the graph displaying the dose-response curve ends, on the *X*-axis, at the maximum value of the main continuous exposure. Assigning a value to MAX_XAXIS enables the graph to end at this assigned value.

NO_TITLE: if set to 1, there is no title on the graph. By default, there is a title (NO_TITLE = 0).

NO_LABEL_X: if set to 1, there is no label on the *X*-axis. By default, there is a label of the *X*-axis (NO_LABEL_X = 0), and it is the name of the main continuous exposure.

NO_LABEL_Y: if set to 1, there is no label on the *Y*-axis. By default, there is a label of the *Y*-axis (NO_LABEL_Y = 0), of which the text depends on the type of the regression.

NO_LEGEND: if set to 1, there is no legend under the *X*-axis. By default, the legend is below the *X*-axis (NO_LEGEND = 0).

PRINT_COVAR_MAT: if set to 1, the estimated regression parameter covariance matrix is displayed in the SAS OUTPUT window. By default, this matrix is not displayed (PRINT_COVAR_MAT = 0).

NO_DELETE_FILES: by default, the macro deletes all temporary files that have been created in the SAS Working library (NO_DELETE_FILES = 0). To keep these files, NO_DELETE_FILES must be assigned to 1.

## Appendix B: Availability of the %RCS_REG SAS macro and NHANES III data

The %RCS_Reg SAS macro as well as the NHANES III data can be sent by email upon request at loic.desquilbet@gmail.com.

## References

1. Steenland K, Deddens JA. A practical guide to dose-response analyses and risk assessment in occupational epidemiology. *Epidemiology* 2004; **15**:63–70.
2. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine* 2007; **26**:5512–5528.
3. Maclure M, Greenland S. Tests for trend and dose response: misinterpretations and alternatives. *American Journal of Epidemiology* 1992; **135**:96–104.
4. Zhao LP, Kolonel LN. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology* 1992; **136**:464–474.
5. Altman DG, Royston P. The cost of dichotomising continuous variables. *British Medical Journal* 2006; **332**:1080.
6. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology* 1995; **6**:450–454.
7. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine* 2006; **25**:127–141.
8. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics* 1994; **43**:429–467.
9. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**:964–974.
10. Smith PL. Splines as a useful and convenient statistical tool. *The American Statistician* 1979; **33**:57–62.
11. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* 1995; **6**:356–365.
12. Govindarajulu US, Spiegelman D, Thurston SW, Ganguli B, Eisen EA. Comparing smoothing techniques in Cox models for exposure–response relationships. *Statistics in Medicine* 2007; **26**:3735–3752.
13. Becher H. The concept of residual confounding in regression models and some applications. *Statistics in Medicine* 1992; **11**:1747–1758.
14. Brown CC, Kipnis V, Freedman LS, Hartman AM, Schatzkin A, Wacholder S. Energy adjustment methods for nutritional epidemiology: the effect of categorization. *American Journal of Epidemiology* 1994; **139**:323–338.
15. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997; **8**:429–434.
16. Byers T, Lyle B. The role of epidemiology in determining when evidence is sufficient to support nutrition recommendations, Summary statement. *The American Journal of Clinical Nutrition* 1999; **69**:1365S–1367S.
17. Potischman N, Weed DL. Causal criteria in nutritional epidemiology. *The American Journal of Clinical Nutrition* 1999; **69**:1309S–1314S.
18. Welten DC, Kemper HC, Post GB, van Staveren WA. A meta-analysis of the effect of calcium intake on bone mass in young and middle aged females and males. *The Journal of Nutrition* 1995; **125**:2802–2813.
19. Jacques PF, Selhub J, Bostom AG, Wilson PW, Rosenberg IH. The effect of folic acid fortification on plasma folate and total homocysteine concentrations. *The New England Journal of Medicine* 1999; **340**:1449–1454.
20. Pekkanen J, Linn S, Heiss G, Suchindran CM, Leon A, Rifkind BM, Tyroler HA. Ten-year mortality from cardiovascular disease in relation to cholesterol level among men with and without preexisting cardiovascular disease. *The New England Journal of Medicine* 1990; **322**:1700–1707.
21. Rywik SL, Manolio TA, Pajak A, Piotrowski W, Davis CE, Broda GB, Kawalec E. Association of lipids and lipoprotein level with total mortality and mortality caused by cardiovascular and cancer diseases (Poland and United States collaborative study on cardiovascular epidemiology). *The American Journal of Cardiology* 1999; **84**:540–548.
22. Cox DR. Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
23. Harrell Jr FE, Lee KL, Pollock BG. Regression models in clinical studies: determining relationships between predictors and response. *Journal of the National Cancer Institute* 1988; **80**:1198–1202.
24. Durrleman S, Simon R. Flexible regression models with cubic splines. *Statistics in Medicine* 1989 **8**:551–561.
25. Marrie RA, Dawson NV, Garland A. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. *Journal of Clinical Epidemiology* 2009; **62**:511–517 e511.
26. Friedrich N, Milman N, Volzke H, Linneberg A, Jorgensen T. Is serum ferritin within the reference range a risk predictor of cardiovascular disease? A population-based, long-term study comprising 2874 subjects. *The British Journal of Nutrition* 2009; **102**:594–600.
27. Weiner DE, Tighiouart H, Elsayed EF, Griffith JL, Salem DN, Levey AS, Sarnak MJ. The relationship between nontraditional risk factors and outcomes in individuals with stage 3 to 4 CKD. *American Journal of Kidney Diseases* 2008; **51**:212–223.
28. Saraiya M, Kottiri BJ, Leadbetter S, Blackman D, Thompson T, McKenna MT, Stallings FL. Total and percent free prostate-specific antigen levels among U.S. men, 2001–2002. *Cancer Epidemiology, Biomarkers and Prevention* 2005; **14**:2178–2182.
29. Jiang W, Babyak MA, Rozanski A, Sherwood A, O'Connor CM, Waugh RA, Coleman RE, Hanson MW, Morris JJ, Blumenthal JA. Depression and increased myocardial ischemic activity in patients with ischemic heart disease. *American Heart Journal* 2003; **146**:55–61.
30. Schernhammer ES, Ogino S, Fuchs CS. Folate and vitamin B6 intake and risk of colon cancer in relation to p53 expression. *Gastroenterology* 2008; **135**:770–780.
31. Troude P, Chaix ML, Tran L, Deveau C, Seng R, Delfraissy JF, Rouzioux C, Goujard C, Meyer L. No evidence of a change in HIV-1 virulence since 1996 in France. *AIDS* 2009; **23**:1261–1267.
32. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**:121–130.
33. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; **19**:716–723.
34. Jacques PF, Rosenberg IH, Rogers G, Selhub J, Bowman BA, Gunter EW, Wright JD, Johnson CL. Serum total homocysteine concentrations in adolescent and adult americans: results from the third national health and nutrition examination Survey. *The American Journal of Clinical Nutrition* 1999; **69**:482–489.
35. Looker AC, Wahner HW, Dunn WL, Calvo MS, Harris TB, Heyse SP, Johnston Jr CC, Lindsay R. Updated data on proximal femur bone mineral levels of US adults. *Osteoporosis International* 1998; **8**:468–489.
36. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Third report of the National Cholesterol Education Program (NCEP) expert panel on detection, evaluation and treatment of high blood cholesterol in adults (adult treatment panel III) final report. *Circulation* 2002; **106**:3143–3421.
37. Hastie T, Tibshirani R. Generalized additive models. *Statistical Science* 1986; **1**:297–310.

38. Gilboa SM, Correa A, Alverson CJ. Use of spline regression in an analysis of maternal prepregnancy body mass index and adverse birth outcomes: does it tell us more than we already know? *Annals of Epidemiology* 2008; **18**:196–205.
39. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; **57**:120–125.
40. Heinzl H, Kaider A. Gaining more flexibility in cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine* 1997; **54**:201–208.
41. Gregory M, Ulmer H, Pfeiffer KP, Lang S, Strasak AM. A set of SAS macros for calculating and displaying adjusted odds ratios (with confidence intervals) for continuous covariates in logistic B-spline regression models. *Computer Methods and Programs in Biomedicine* 2008; **92**:109–114.
42. Cui J, de Klerk N, Abramson M, Del Monaco A, Benke G, Dennekamp M, Musk AW, Sim M. Fractional polynomials and model selection in generalized estimating equations analysis, with an application to a longitudinal epidemiologic study in Australia. *American Journal of Epidemiology* 2009; **169**:113–121.
43. Sauerbrei W, Meier-Hirmer C, Benner A, Royston P. Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs. *Computational Statistics and Data Analysis* 2006; **50**:3464–3485.