frontiers | Frontiers in Epidemiology

TYPE Methods
PUBLISHED 18 August 2022
DOI 10.3389/fepid.2022.975380

OPEN ACCESS

EDITED BY
Rolf H. H. Groenwold,
Leiden University Medical
Center, Netherlands

REVIEWED BY
Michael Kammer,
Medical University of Vienna, Austria
Thaddäus Tönnies,
German Diabetes Center
(DDZ), Germany
Sanjay Rampal,
University of Malaya, Malaysia
Li Xiao,
University of Science and Technology
of China, China

*CORRESPONDENCE
Noah A. Schuster
n.schuster@amsterdamumc.nl

SPECIALTY SECTION
This article was submitted to
Research Methods and Advances in
Epidemiology,
a section of the journal
Frontiers in Epidemiology

RECEIVED 22 June 2022
ACCEPTED 01 August 2022
PUBLISHED 18 August 2022

CITATION
Schuster NA, Rijnhart JJM, Twisk JWR and Heymans MW (2022) Modeling non-linear relationships in epidemiological data: The application and interpretation of spline models. *Front. Epidemiol.* 2:975380. doi: 10.3389/fepid.2022.975380

Modeling non-linear relationships in epidemiological data: The application and interpretation of spline models

Noah A. Schuster[1,2]*, Judith J. M. Rijnhart[1,2], Jos W. R. Twisk[1,2] and Martijn W. Heymans[1,2]

[1]Amsterdam UMC location Vrije Universiteit Amsterdam, Epidemiology and Data Science, Amsterdam, Netherlands, [2]Amsterdam Public Health, Methodology, Amsterdam, Netherlands

Objective: Traditional methods to deal with non-linearity in regression analysis often result in loss of information or compromised interpretability of the results. A recommended but underutilized method for modeling non-linear associations in regression models is spline functions. We explain spline functions in a non-mathematical way and illustrate the application and interpretation to an empirical data example.

Methods: Using data from the Amsterdam Growth and Health Longitudinal Study, we examined the non-linear relationship between the sum of four skinfolds and VO_2max, which are measures of body fat and cardiorespiratory fitness, respectively. We compared traditional methods (i.e., quadratic regression and categorization) to spline methods [1- and 3-knot linear spline (LSP) models and a 3-knot restricted cubic spline (RCS) model] in terms of the interpretability of the results and their explained variance (r^2_{adj}).

Results: The spline models fitted the data better than the traditional methods. Increasing the number of knots in the LSP model increased the explained variance (from $r^2_{adj} = 0.578$ for the 1-knot model to $r^2_{adj} = 0.582$ for the 3-knot model). The RCS model fitted the data best ($r^2_{adj} = 0.591$), but results in regression coefficients that are harder to interpret.

Conclusion: Spline functions should be considered more often as they are flexible and can be applied in commonly used regression analysis. RCS regression is generally recommended for prediction research (i.e., to obtain the predicted outcome for a specific exposure value), whereas LSP regression is recommended if one is interested in the effects in a population.

KEYWORDS

linear splines, restricted cubic splines, non-linearity, regression analysis, epidemiological methods, spline functions

Introduction

In epidemiological research, regression analysis is often used to examine the association between an outcome and an exposure (1). A principal assumption of regression analysis is that the continuous exposure is linearly related to the outcome. In other words, a one-unit difference in the exposure is associated with a fixed difference in the outcome, regardless of the values of the exposure (2). However, linearity should not be assumed without assessing that the association is indeed linear (3–5). If the linearity assumption is violated and associations are estimated as linear nonetheless, then the effect estimate might not be a good representation of the true underlying effect and bias might be introduced. In order to obtain unbiased effects, the non-linear association requires explicit modeling. Failing to estimate a truly non-linear relationship as non-linear may lead to over- or underestimation of the exposure effect. However, it is important to note that the estimation of complex models may come at cost of increase uncertainty, especially in small samples. Therefore, in practice, one may want to consider the balance between model complexity and model uncertainty when choosing an appropriate method to model non-linear relationships.

There are different methods available to model non-linear associations. Simple methods such as polynomial regression (e.g., quadratic or cubic regression) and categorization of the exposure variable are widely used, largely due to historical precedent (6). With quadratic regression, for instance, the outcome is modeled as a quadratic function of the exposure (i.e., as a function of exposure x and the quadratic term x^2) (2, 7, 8). Adding higher order terms (such as a quadratic term) to a basic linear function increases the flexibility of the model, but simultaneously complicates the interpretability of the results as the regression coefficients of the terms cannot be interpreted separately from each other.

With categorization, the exposure variable is grouped (e.g., based on percentile values) and subsequently analyzed as a categorical variable with one of the groups as the reference category. However, categorization is associated with multiple issues, such as loss of information, discontinuity in the estimated average outcome value when moving from one category to the other, and difficulties with comparing results across studies as the cut-off points may be data dependent (2, 6, 8–13). Filardo et al. found that study findings were inconsistent under different exposure categorization schemes identified in the literature, which suggests that the way the exposure is categorized may impact conclusions (14). This emphasizes the importance of correctly modeling non-linear relationships.

A different approach to model non-linear associations is the use of spline functions in the regression model (2, 3, 8, 11, 12, 15, 16). Spline functions are transformations of the continuous exposure variable and can be added to any regression analysis. They are available in different forms, such as simple linear spline (LSP) functions, more complex restricted cubic spline (RCS) functions and B-splines (2). Spline functions estimate exposure effects for specific intervals of the exposure variable and are subject to continuity restrictions (i.e., the interval functions meet at the common interval edges so that—in contrast to categorization—there are no jumps in the line at these points) (17). In this paper, we focus on LSP and RCS functions. LSP functions assume that the exposure effects within each interval follow a linear shape, but across the intervals the effect may be non-linear. Therefore, LSP functions are more flexible than simple linear regression and categorization. RCS functions assume that the exposure effects within each category are cubic functions, allowing for more flexibility than other methods.

Although spline functions are broadly accessible in the software packages commonly used by epidemiologists, they are not widely used (3, 18). Most papers published on spline functions present these as complex mathematical functions (15, 19, 20) and do not discuss their interpretation. This may be one of the reasons that researchers default to less optimal methods for estimating non-linear effects, such as quadratic terms and categorization.

The aim of this paper is to describe linear and restricted cubic spline functions in a step-by-step and non-mathematical manner, and to demonstrate the advantages of these methods over simple linear regression, quadratic regression and categorization using an empirical data example. First, we provide an introduction into spline regression and describe linear- and restricted cubic spline regression in the context of an empirical data example. Then, we illustrate the application of traditional methods and spline methods to model non-linear relationships to that same data example. Finally, we discuss the interpretation of the effect estimates from different methods and describe the context in which the use of LSP and RCS models may be relevant.

Methods

Example dataset

Spline functions will be explained by using an empirical data example from the Amsterdam Growth and Health Longitudinal Study (AGHLS). The AGHLS is an ongoing cohort study that was set up to examine the growth, health and lifestyle among teenagers (21). We use data from the third round of measurements, when the participants were 15 years old, because it contains a clear non-linear relationship.

Throughout this paper, we analyze the non-linear relationship between the sum of four skinfolds (SFS) and cardiorespiratory fitness (VO_2max). SFS is an often used estimate of body fat and is calculated by summing the biceps-, triceps-, subscapular-, and suprailiac skinfolds (in millimeters) (22). VO_2max is defined as the absolute maximal oxygen uptake in centiliter per kilogram bodyweight (21). The relationship

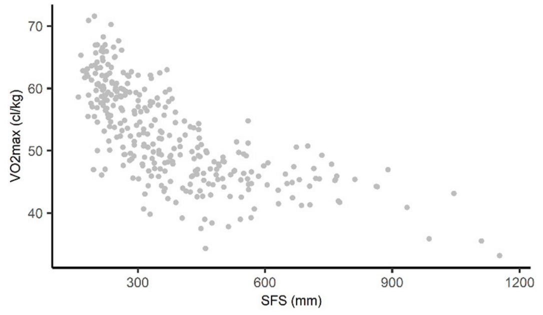

FIGURE 1
Non-linear relationship between SFS and VO_2max in AGHLS data. SFS, sum of four skinfolds; AGHLS, Amsterdam Growth and Health Longitudinal Study.

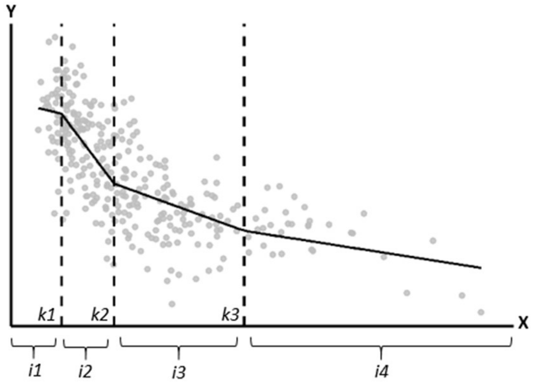

FIGURE 2
Graphical depiction of the important properties of a spline model. The gray points represent the observed data, and the black line is the fitted linear spline model. The vertical dotted lines represent the knots (located at *k1*, *k2*, and *k3*). *i1*, *i2*, *i3*, and *i4* represent the four intervals for which the exposure effect is estimated.

between SFS and VO_2max in our data is shown in Figure 1. Only subjects with complete data on both variables were included in the analysis ($n = 315$, 6 subjects were excluded because of incomplete data).

Spline functions

Splines can be applied to any statistical model that linearly relates the exposure to the outcome, such as linear, logistic, and Cox regression. With spline models, the continuous independent variable is divided into multiple intervals, and for each interval the relationship between the exposure and outcome is estimated separately. The relationship between the exposure and the outcome in each interval can, for example, be estimated with a linear function (resulting in linear spline regression) or with a cubic function (resulting in cubic spline regression). The use of so-called *spline basis functions* makes it possible to estimate the relationship between the exposure and the outcome for each of the intervals in the same model. The values of the exposure based on which the intervals are created are referred to as *knots*. Thus, each knot defines the end of one interval and the start of the next. In 3-knot models, the exposure is divided into four intervals. Subsequently, for each interval the exposure effect is estimated, resulting in four spline coefficients. Corresponding confidence intervals can, for example, be calculated with the standard errors or be obtained by bootstrapping (23).

In general, a small number of knots (i.e., 3 to 5) is sufficient to model a non-linear relationship. If the sample size is large and the relationship that is studied changes quickly, then more knots might be required (2, 24, 25). Increasing the number of knots generally improves the fit of the model, but may also lead to overfitting of the model to the data. If that is the case, the fitted function does not only follow the main features of the data but also small and random fluctuations (2, 7, 25). Wand presents an overview of statistical methods for establishing the number of knots (26).

Often, the locations of the knots are pre-specified based on the quantiles of the independent variable. For 3-knot models, Harrell recommends knots at the 10th, 50th, and 90th percentile. For 4-knot models, they are recommended at the 5th, 35th, 65th, and 95th percentile (2). In some cases, knot locations are suggested by theory or by study design (e.g., an interrupted time series design). However, generally the fit of a spline model is more dependent on the number of knots than on the knot locations (25).

In this paper, for illustrational purposes, we demonstrate 1- and 3-knot linear spline models and a 3-knot restricted cubic spline model using the knot locations recommended by Harrell. Figure 2 shows the most important properties of a spline model. The gray points in Figure 2 represent the observed data, and the black line is the fitted 3-knot linear spline model. The vertical dotted lines represent the three knots (labeled as $k1$, $k2$, and $k3$) and the lines in between the knots represent the estimated exposure effect for the four intervals between the knots. Spline models are based on continuity restrictions, which ensures that the line is smooth at the knots. For example, the line for the first interval is smoothly connected to the line of the second interval, and the line of the second interval is smoothly connected to the line of the third interval, etcetera. An interactive visualization of LSP and RCS models and the influence of the continuity restrictions, number of knots and location of knots on the estimated line can be found elsewhere (27, 28).

Linear spline models

In the 1-knot LSP model, the knot is located at the 50th percentile ($SFS = 330\ mm$). The corresponding linear spline model is

$$VO2max = \beta_0 + \beta_1 * SFS + \beta_2^* * (SFS - 330)_+ + \varepsilon \quad (1)$$

where β_0 represents the intercept and ε represents an error term. To provide valid inference via e.g., confidence intervals for coefficients, it is assumed that the error terms for each observation are uncorrelated and follow a Gaussian distribution with expected value of zero. The term $(SFS - 330)_+$ represents the *spline basis function*. This function is assigned a value of zero when $SFS - 330 \leq 0$. Because of this, coefficient β_1 represents the exposure effect estimate for individuals whose SFS is ≤330 mm. Coefficient β_2^* represents the difference in the effect estimates between the individuals whose SFS is ≤330 mm and those whose SFS is >330 mm. Thus, for individuals whose SFS is >330 mm, their exposure effect estimate is represented by $\beta_1 + \beta_2^*$. The 95% confidence interval corresponding to β_2^* can be used to assess whether the slopes for the two intervals of SFS are statistically significantly different.

In the 3-knot LSP model, the knots are located at the 10th, 50th, and 90th percentiles, i.e., at SFS = 212, 330, and 621.4 mm, respectively. The corresponding LSP model is

$$\begin{aligned} VO2max = {} & \beta_0 + \beta_1 * SFS + \beta_2^* * (SFS - 212)_+ \\ & + \beta_3^* * (SFS - 330)_+ + \beta_4^* * (SFS - 621.4)_+ + \varepsilon \end{aligned} \quad (2)$$

In Equation 2, spline coefficient β_2^* is only used whenever an individuals' SFS value is larger than 212, otherwise it is multiplied by zero and thus plays no role in the equation. For coefficient β_3^* this is for $SFS > 330$ and for coefficient β_4^* this is for $SFS > 621.4$, respectively. Thus, coefficient β_1 represents the exposure effect estimate for individuals whose SFS is ≤212 mm, while $\beta_1 + \beta_2^*$ represents the effect estimate for individuals whose SFS is >212 and ≤330 mm. The exposure effect estimates for individuals in the third and fourth interval (i.e., individuals whose SFS is >330 mm and ≤621.4 mm, and individuals whose SFS is >621.4 mm) are represented by $\beta_1 + \beta_2^* + \beta_3^*$ and $\beta_1 + \beta_2^* + \beta_3^* + \beta_4^*$, respectively.

For both the 1- and 3-knot LSP models, fitting the spline models is straightforward once the spline basis functions have been established. Appendix A contains a step by step description of how to estimate these models, including R software code.

Restricted cubic spline models

Although LSP models can approximate many relationships, they do not draw smooth lines and do not fit highly curved relationships well. This can be resolved by fitting a cubic spline model, which joins smoothly at the knot locations because the slopes are restricted to be equal at the boundaries (8). To improve the performance of the spline model in the tails of the exposure variable, where little data is located, additional constraints are imposed in *restricted* cubic spline models. In RCS models, the spline functions are linear in the tails (i.e., before the first and after the last knot) (2, 29). Whereas, in LSP models each interval is represented by a spline basis function, in RCS models $k - 2$ spline variables are fitted, where k is the number of knots. Thus, in a 3-knot restricted spline function, a single spline basis function is fitted (Equation 3).

$$VO2max = \beta_0 + \beta_1 * SFS + \beta_2^\dagger * SFS_2^\dagger + \varepsilon \quad (3)$$

where $SFS_2^\dagger$ and $\beta_2^\dagger$ represent the spline basis function and corresponding cubic spline coefficient (2). Each participant's value for the spline basis function is estimated as a function of the observed exposure value and the knot locations (i.e., SFS = 212, 330, and 621.4, respectively). The exact formula with which spline basis function $SFS_2^\dagger$ is calculated is presented in Appendix B. Equation 3 can also be expressed as Equation 4, which contains the interval functions and has the same form as the 3-knot LSP model. The only difference between the LSP and RCS models is that for RCS regression all spline basis functions are raised to the power of three:

$$\begin{aligned} VO2max = {} & \beta_0 + \beta_1 * SFS + \beta_2^* * (SFS - 212)_+^3 \\ & + \beta_3^* * (SFS - 330)_+^3 + \beta_4^* * (SFS - 621.4)_+^3 \# + \varepsilon \end{aligned} \quad (4)$$

Equation 5 to 7 can be used to convert cubic spline coefficient $\beta_2^\dagger$ into regression coefficients for each of the intervals:

$$\beta_2^* = \frac{\beta_2^\dagger}{(621.4 - 212)^2} \quad (5)$$

$$\beta_3^* = \frac{\beta_2^* * (212 - 621.4)}{(621.4 - 330)} \quad (6)$$

$$\beta_4^* = \frac{\beta_2^* * (212 - 330)}{(330 - 621.4)} \quad (7)$$

In Equation 5, β_2^* represents the coefficient for the interval between the first and the second knot and $\beta_2^\dagger$ is the cubic spline basis function coefficient from Equation 3. In Equation 6, β_3^* represents the coefficient for the interval between the second and third knot and β_2^* is the regression coefficient from Equation 4. In Equation 7, β_4^* represents the coefficient for the interval after the third and β_2^* is the regression coefficient from Equation 4. Subsequently, coefficients β_2^*, β_3^* and β_4^* can be plugged into Equation 4.

Like in quadratic regression, the exposure effect estimates differ across exposure values, which makes it less straightforward to interpret the coefficients from an RCS model.

TABLE 1 Regression- and interval coefficients for the relationship between VO_2max and SFS derived from linear- and quadratic regression, categorization, 1- and 3-knot linear spline regression and 3-knot restricted cubic spline regression.

Estimate	Regression coefficient	Interval coefficient
Linear regression		
β_0	64.0658	
β_1	−0.0304	
Quadratic regression		
β_0	73.2212	
β_1	−0.0746	
β_2	0.00004	
Categorization		
β_0	60.1339	
β_1	−4.9870	
β_2	−10.0695	
β_3	−15.1727	
1-knot linear spline regression		
β_0	77.5648	
β_1	−0.0810	$SFS \leq 330: -0.0810$
β_2^*	0.0632	$SFS > 330: -0.0810 + 0.0632 = -0.0178$
3-knot linear spline regression		
β_0	64.1788	
β_1	−0.0156	$SFS \leq 212: -0.0156$
β_2^*	−0.0671	$212 < SFS \leq 330: -0.0156 - 0.0671 = -0.0827$
β_3^*	0.0601	$330 < SFS \leq 621.4: -0.0156 - 0.0671 + 0.0601 = -0.0226$
β_4^*	0.0128	$SFS > 621.4: -0.0156 - 0.0671 + 0.0601 + 0.0128 = -0.0098$
3-knot restricted cubic spline		
β_0	75.9306	
β_1	−0.0738	
$\beta_2^\dagger$	0.0740	$\beta_2^*: 0.0000004$ $\beta_3^*: -0.0000006$ $\beta_4^*: 0.0000002$

β_2^*, β_3^*, and β_4^* represent spline coefficients that correspond to spline basis functions. $\beta_2^\dagger$ represents the cubic spline coefficient that corresponds to spline variable $SFS_2^\dagger$.

Results

We illustrate the interpretation and compare the performance of different methods to model non-linear relationships using the data example from the AGHLS. Table 1 presents the regression coefficients for each method. For the spline models, these regression coefficients are used to calculate the effects for each interval of SFS. These effects are presented under "interval coefficient." Table 2 presents the adjusted r^2 (i.e., the proportion of variance in VO_2max explained by SFS) of each method (30).

TABLE 2 Explained variance of each model.

Model	Adjusted r^2
Linear regression	0.487
Quadratic regression	0.558
Categorization	0.537
1-knot linear spline regression	0.578
3-knot linear spline regression	0.582
3-knot restricted cubic spline regression	0.591

For illustrative purposes we first estimated a simple linear regression model. Linear regression fits a straight line to the data (Figure 3A) and assumes that the effect of the exposure on the outcome is the same for every value of the exposure. In our data, the exposure effect estimate was −0.0304, meaning that a 1 mm difference in SFS was associated with a 0.0304 cl/kg lower VO_2max, regardless of the compared values of SFS. Naturally, this regression line was not a good representation of the relationship between SFS and VO_2max, which was also reflected in the lowest explained variance ($r^2_{adj} = 0.487$) of all estimated models.

Quadratic regression

With quadratic regression, VO_2max was estimated by SFS and the quadratic term SFS^2. As shown in Figure 3B and reflected in the explained variance ($r^2_{adj} = 0.558$), the quadratic model fitted the form of the relationship between SFS and VO_2max quite well relative to the other models. However, the regression coefficients do not have a straightforward interpretation because the effect of SFS on VO_2max is a function of both regression coefficients. That is, the effect of a one unit difference in SFS on VO_2max differs across SFS. For example, the average difference in VO_2max was −0.0506 cl/kg when SFS changed from 300 to 301 [i.e., $\left(-0.0746 * 301 + 0.00004 * 301^2\right) - \left(-0.0746 * 300 + 0.00004 * 300^2\right)$], while the average difference in VO_2max was −0.0266 cl/kg when SFS changed from 600 to 601 [i.e., $\left(-0.0746 * 601 + 0.00004 * 601^2\right) - \left(-0.0746 * 600 + 0.00004 * 600^2\right)$]. Compared to simple linear regression (Figure 3A), the confidence interval for the line estimated using quadratic regression becomes wider for higher values of SFS (Figure 3B). This reflects the additional uncertainty in the effect estimates from quadratic regression for higher SFS values. However, the wider confidence interval does not affect the conclusion that SFS is associated with VO_2 max.

Categorization

We divided SFS into four intervals based on quartiles. Because we used the lowest quartile as the reference category, the

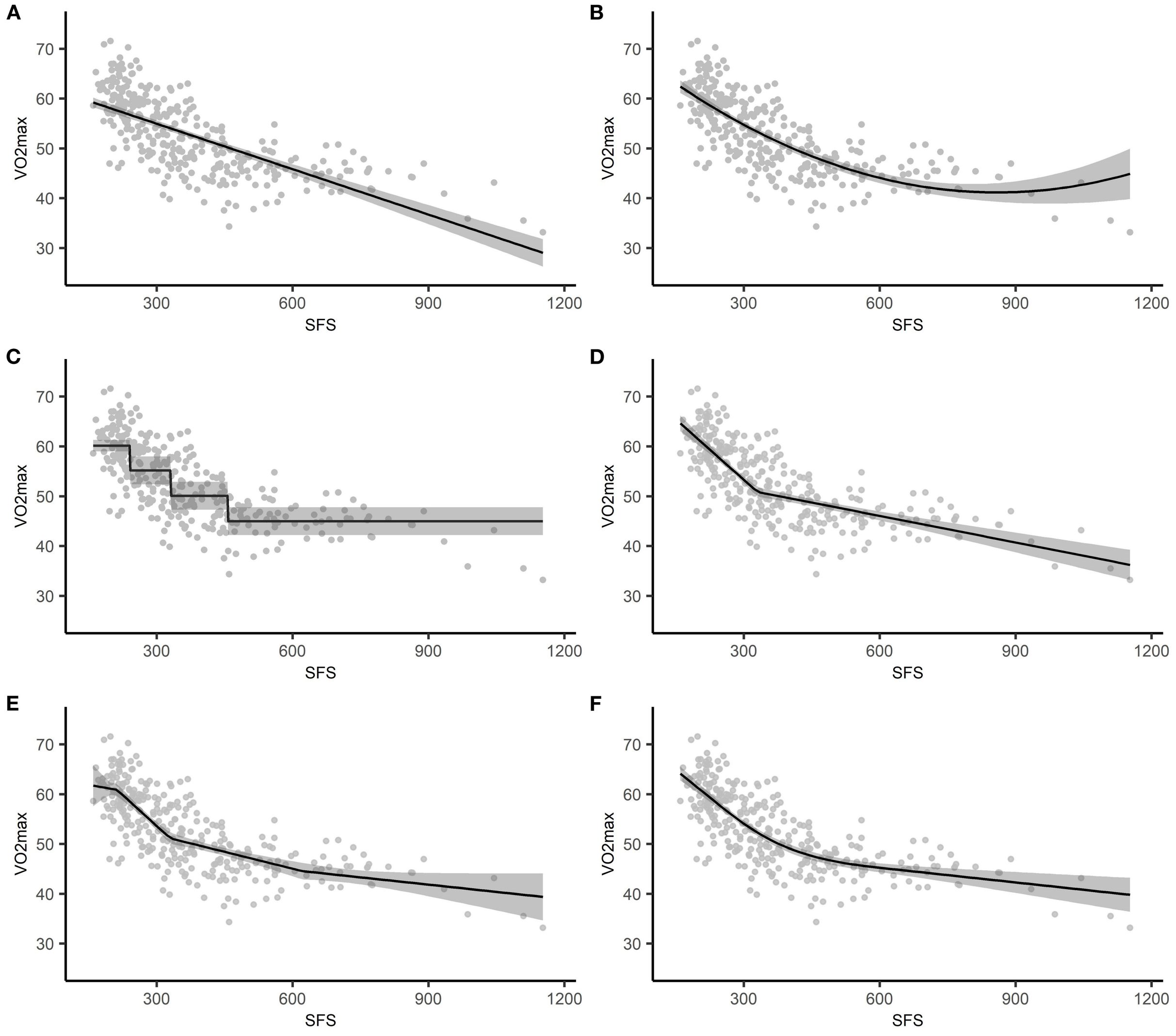

FIGURE 3
The estimated association between SFS and VO$_2$max plotted against the observed values, with the shading representing the 95% confidence intervals based on standard errors. **(A)** simple linear regression, **(B)** polynomial regression, **(C)** categorization, **(D)** 1-knot LSP regression, **(E)** 3-knot LSP regression, **(F)** 3-knot RCS regression. SFS, sum or four skinfolds; LSP, linear spline; RCS, restricted cubic spline.

intercept represented the mean VO$_2$max in cl/kg for individuals in that interval. The regression coefficients represented the mean difference in VO$_2$max between individuals in the lowest quartile and the other quartiles. For example, −4.9870 was the mean difference in VO$_2$max in cl/kg between subjects in the first and second quartile. The explained variance was slightly lower relative to the other models ($r^2_{adj} = 0.537$).

Figure 3C illustrates the assumed homogeneity within groups and the discontinuity in VO$_2$max (i.e., the change in average VO$_2$max value) when moving from one quartile to the next. For example, measures of SFS in the last quartile ranged between 458 and 1,153 mm, but all individuals had the same estimated VO$_2$max of 44.9612 cl/kg (i.e., 60.1339 − 15.1727).

1-knot linear spline model

For individuals whose SFS was equal to or <330 mm, a 1 mm difference in SFS was associated with a 0.0810 cl/kg lower VO$_2$max. The mean difference in the effect estimate between individuals in both intervals was 0.0632, meaning that for individuals whose SFS was >330 mm, a 1 mm difference in SFS was associated with a 0.0178 cl/kg lower VO$_2$max (i.e., −0.0810 + 0.0632). Thus, for individuals whose SFS was >330 mm the association between SFS and VO$_2$max was less strong than for individuals whose SFS was equal to or <330 mm. This is also illustrated in Figure 3D. The $r^2_{adj}$was 0.578. This indicates that the 1-knot linear spline model is a better fit to the data than both quadratic regression and categorization.

3-knot linear spline model

For individuals whose SFS was $\leq$212 mm, a 1 mm difference in SFS was associated with a 0.0156 cl/kg lower VO_2max. For individuals whose SFS was between 213 and 330 mm, a 1 mm difference in SFS was associated with a 0.0827 cl/kg lower VO_2max (i.e., $-0.0156 - 0.0671$). The interval coefficients for the other intervals can be found in Table 1.

Increasing the number of knots from 1 to 3 resulted in a slightly higher explained variance ($r^2_{adj} = 0.578$ vs. $r^2_{adj} = 0.582$, respectively). Furthermore, compared to simple linear regression (Figure 3A) and the 1-knot model (Figure 3D), the confidence interval for the line estimated using a 3-knot model becomes wider for higher values of SFS (Figure 3E). This reflects the additional uncertainty in the effect estimates from the 3-knot model for higher SFS values. However, the wider confidence interval based on the 3-knot model does not affect the conclusion that SFS is associated with VO_2 max.

3-knot restricted cubic spline regression

Like with quadratic regression, separate interpretation of the coefficients is of no practical value with RCS regression, as the effect of SFS on VO_2max is a function of multiple regression coefficients. For example, the average decrease in VO_2max was 0.0644 cl/kg when SFS changed from 300 to 301 mm [i.e., $\left(75.9306 - 0.0738 \ast 301 + 0.0000004 \ast (301 - 212)^3\right) - \left(75.9306 - 0.0738 \ast 300 + 0.0000004 \ast (300 - 212)^3\right)$], while the average decrease in VO_2max was 0.0244 cl/kg when SFS changed from 600 to 601 mm [i.e., $(75.9306 - 0.0738 \ast 601 + 0.0000004 \ast (601 - 212)^3 - 0.0000006 \ast (601 - 330)^3) - (75.9306 - 0.0738 \ast 600 + 0.0000004 \ast (600 - 212)^3 - 0.0000006 \ast (600 - 330)^3)$]. Figure 3F illustrates the restrictions (i.e., the function is linear for $SFS \leq 212$ and $SFS > 621.4$) and shows that the model fits the data quite well. This is also reflected in the explained variance ($r^2_{adj} = 0.591$).

Discussion

The aim of this paper was to explain linear and restricted cubic spline functions in a step-by-step and non-mathematical manner and to demonstrate the advantages of these methods over simple linear regression, quadratic terms and categorization using an empirical data example. Although spline regression is easy to implement with most statistical programs, epidemiologists still often apply traditional methods (e.g., quadratic regression and categorization) to model non-linear relationships.

In the data example, the spline models resulted in higher explained variance than the traditional methods. Both categorization and spline regression divided the continuous exposure variable into intervals. Categorization only allows for variation between categories, so that the estimated outcome is the same for each individual in an interval regardless of their individual exposure value. This explains the stepwise pattern in Figure 3C. Spline regression, on the other hand, allows for variation between and within intervals. As a result, the regression line shifts between knot locations, and regression lines meet at the knot locations. Although polynomial regression is easy to model, it suffers from a lack of smoothness and can lead to implausible curvatures, in particular at the edges. Splines provide a good alternative as they control for this curvature via the continuity restrictions. In addition, RCS models are linear before and after the last knot. LSP models provide a good balance between modeling the non-linear association and providing results that are relatively easy to interpret. Furthermore, RCS models provide a flexible method for modeling the non-linearity of an association, but come at the cost of regression coefficients that are less easy to interpret than LSP models. In our data example, the explained variance in the LSP model and the RCS model were comparable. If one is interested in reporting the association between sum of four skinfolds and VO_2max, then LSP models provide easier interpretations than the RCS models.

For both quadratic regression and RCS models, the increased complexity of the interpretation of the regression coefficients makes it less straightforward to summarize the exposure effect at the population level, because the exposure effect estimates differs in magnitude across exposure values. However, this is not necessarily a problem when the aim of a study is to make individual-level predictions of the outcome, as it remains relatively straightforward to compute the predicted outcome value for a specific exposure value using Equation 7 (8). Thus, in our data example, if one is interested in predicting VO_2max based on specific values of the sum of four skinfolds, then RCS models may be preferred. Two things that might help with interpreting the results are the reporting of figures (such as Figure 3) and calculating the effect for a number of different exposure contrasts (i.e., the two exposure values that are being compared). The latter was done for the interpretation of the 3-knot RCS model, and showed that the decrease in VO_2max was greater when SFS changed from 300 to 301 mm, then when it changed from 600 to 601 mm.

A strength of this paper are the non-mathematical explanations of LSP and RCS models. Although there are many other sources that describe spline models, most of these sources contain a high level of mathematical detail, which may discourage applied researchers from learning about these methods. In this paper, we tried to explained spline functions in a non-mathematical manner and in the context of an empirical data example. Furthermore, although we illustrated the application of spline models using cross-sectional data and within a linear regression context, the spline functions presented can be applied to all kinds of regression models, for example logistic and Cox regression. Further, they can also be used in longitudinal models such as generalized linear mixed models (GLMM) and generalized estimation equations (GEE).

Besides the methods discussed in the present paper, there are also other methods available that can be used to estimate non-linear associations. A method that we did not discuss is quadratic spline regression, in which the spline basis functions are quadratic functions. Although quadratic splines are often overlooked and not mentioned in known reference books (2), like cubic splines they result in smooth functions at the knot locations and can occur in restricted and unrestricted form. When the number of degrees of freedom are the same and the knots are located at comparable exposure values, restricted quadratic and cubic spline models might even yield similar results (31). SAS code for the estimation of restricted quadratic splines is provided by Howe et al. (31). Furthermore, we also did not discuss generalized additive models (GAMs), LOESS smoothing, penalized splines and fractional polynomials (32, 33), which are all capable of capturing non-linear relationships. However, these methods are relatively complicated and therefore, not much used in practice.

In this paper, we explained spline models based on a single exposure. However, in practice, researchers may want to adjust their association model for potential confounders of the exposure-outcome association. Most researchers are unaware that, if these confounders are continuous, then the linearity assumption also applies to these variables (34). Failing to explicitly model a non-linear confounder-outcome association may result in an under- or overestimation of the true exposure effect. Therefore, the linearity assumption should be assessed for each continuous confounder in a regression model, and splines can be applied when necessary.

Spline regression is easy to implement with most statistical software programs often used by epidemiologists. Table 3 contains a (non-exhaustive) overview of packages and macros available in different software programs. The analyses in this paper were conducted using the R programming language version 4.0.3 (35) and the "rms" package by Harrell (23). The R package "splines" is part of the basic distribution of R (29). Other frequently downloaded packages include "gss" (36) and "polspline" (37). An overview of spline methods and other R packages that may be used to fit spline models is presented elsewhere (29). In STATA, spline functions can be fitted using, among others, the STATA package "rmkspline" and the user-made package "RCsplines" (38). In SPSS, spline functions have to be fitted by hand and can be applied using the REGRESSION procedure. In SAS, the "effect" statement in "proc glimmix" provides an automated implementation for fitting splines. Documentation including syntax commands are available from the IBM support page (39) and the SAS Help Center (40).

TABLE 3 Spline regression options by software program.

Software program	Packages/procedures
R	rms, splines, gss, polspline
STATA	mkspline, RCsplines
SPSS	REGRESSION
SAS	TRANSREG

Although splines are easy to implement, they require certain choices to be made by the researcher. This concerns, for example, the number and location of the knots and the type of basis function (2). In addition, not all non-linear relations are "equally harmful" and the choice of spline model (e.g., linear or cubic) might depend on what's considered more important: LSP models might be used to model relations that only have a slight bend and that can be approximated by piecewise linear functions, whereas RCS might be used for maximum model accuracy. Another thing to consider is that some choices, such as increasing the number of knots, might introduce additional uncertainty to the model, especially in small samples. If the number of knots is too large, then the model overfits the data: it then describes the random error rather than the relationship between the variables. This affects the generalizability of the model outside of the data that it is based on (29). In our example, the confidence intervals were generally wider for more complex models, illustrating the additional model uncertainty introduced by more complex models. In some situations, the additional uncertainty might be a reason to use a more simple model.

Conclusion

Spline functions should be considered more often in the analysis of non-linear relationships as they allow for more flexibility in estimating non-linear associations than traditional methods such as quadratic regression and categorization and can be used in all kinds of regression analyses. With RCS models the exposure effect estimates differ across exposure values, making them more suitable for prediction (i.e., to obtain the predicted outcome for a specific exposure value). If one is interested in the effects in a population, then LSP models are more suitable due to the straightforward interpretation of the regression coefficients.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary materials, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by VU Medical Center. The patients/participants provided their written informed consent to participate in this study.

Author contributions

NS and MH designed the study. NS performed the statistical analyses and drafted the manuscript. All authors contributed to data interpretation, critically revised the manuscript, and approved the final version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fepid.2022.975380/full#supplementary-material

References

1. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*. 4 ed. Lippincott Williams & Wilkins (2021).

2. Harrell FE. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Switzerland: Springer International Publishing AG (2015).

3. Marrie RA, Dawson NV, Garland A. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. *J Clin Epidemiol.* (2009) 62:511–7.e1. doi: 10.1016/j.jclinepi.2008.05.015

4. Philippe P, Mansi O. Nonlinearity in the epidemiology of complex health and disease processes. *Theoret Med Bioethics.* (1998) 19:591–607. doi: 10.1023/A:1009979306346

5. Rapoport J, Teres D, Lemeshow S, Avrunin JS, Haber R. Explaining variability of cost using a severity-of-illness measure for ICU patients. *Medical Care.* (1990) 28:338–48. doi: 10.1097/00005650-199004000-00005

6. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Med Res Methodol.* (2012) 12:21. doi: 10.1186/1471-2288-12-21

7. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge: Cambridge University Press (2003).

8. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology.* (1995) 6:356–65. doi: 10.1097/00001648-199507000-00005

9. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Danger of using "optimal" cut points in the evaluation of prognostic factors. *J Natl Cancer Inst.* (1994) 86:829–35. doi: 10.1093/jnci/86.11.829

10. Gauthier J, Wu QV, Gooley TA. Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians. *Bone Marrow Transplant.* (2020) 55:675–80. doi: 10.1038/s41409-019-0679-x

11. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology.* (1995) 6:450–4. doi: 10.1097/00001648-199507000-00025

12. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* (2006) 25:127–41. doi: 10.1002/sim.2331

13. Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K. Statistical methods in epidemiology: a comparison of statistical methods to analyze dose–response and trend analysis in epidemiologic studies. *J Clin Epidemiol.* (1998) 51:1223–33. doi: 10.1016/S0895-4356(98)00129-2

14. Filardo G, Hamilton C, Hamman B, Ng HKT, Grayburn P. Categorizing BMI may lead to biased results in studies investigating in-hospital mortality after isolated CABG. *J Clin Epidemiol.* (2007) 60:1132–9. doi: 10.1016/j.jclinepi.2007.01.008

15. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med.* (1989) 8:551–61. doi: 10.1002/sim.4780080504

16. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer (2013).

17. Marsh LC, Cormier DR. *Spline Regression Models*. Thousand Oaks, CA: Sage Publications, Inc. (2002).

18. O'Brien SM. Cutpoint selection for categorizing a continuous predictor. *Biometrics.* (2004) 60:504–9. doi: 10.1111/j.0006-341X.2004.00196.x

19. de Boor CR. *A Practical Guide to Splines*. New York, NY: Springer-Verlag New York (1978).

20. Smith PL. Splines as a useful and convenient statistical tool. *Am Stat.* (1979) 33:57–62. doi: 10.1080/00031305.1979.10482661

21. Wijnstok NJ, Hoekstra T, van Mechelen W, Kemper HCG, Twisk JWR. Cohort profile: the Amsterdam growth and health longitudinal study. *Int J Epidemiol.* (2013) 42:422–9. doi: 10.1093/ije/dys028

22. Wijnstok NJ, Serné EH, Hoekstra T, Schouten F, Smulders YM, Twisk JWR. The relationship between 30-year developmental patterns of body fat and body fat distribution and its vascular properties: the Amsterdam Growth and Health Longitudinal Study. *Nutr Diabetes.* (2013) 3:e90. doi: 10.1038/nutd.2013.31

23. Harrell FE. *rms: Regression Modeling Strategies*. R package version 6.0-1 (2020). Available online at: https://cran.r-project.org/package=rms (accessed July 31, 2022).

24. Korn EL, Graubard bI. *Analysis of Health Surveys*. 1 ed. New York, NY: Wiley-Interscience (1999).

25. Stone CJ, Koo C. Additive splines in statistics. In: *American Statistical Association Proceedings of the Statistical Computing Setting*. Washington, DC: American Statistical Association (1985). p. 45–8.

26. Wand MP. A comparison of regression spline smoothing procedures. *Comput Stat.* (2000) 15:443–62. doi: 10.1007/s001800000047

27. Lambert P. *Spline Continuity*. Available online at: https://pclambert.net/interactivegraphs/spline_continuity/spline_continuity (accessed July 31, 2022).

28. Lambert P. *The Number and Location of Knots*. Available online at: https://pclambert.net/interactivegraphs/spline_eg/spline_eg (accessed July 31, 2022).

29. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. *BMC Med Res Methodol.* (2019) 19:46. doi: 10.1186/s12874-019-0666-3

30. Ezekiel M. *Methods of Correlation Analysis*. New York, NY: John Wiley and Sons (1930).

31. Howe CJ, Cole SR, Westreich DJ, Greenland S, Napravnik S, Eron JJ Jr. Splines for trend analysis and continuous confounder control. *Epidemiology*. (2011) 22:874–5. doi: 10.1097/EDE.0b013e31823029dd

32. Eisen EA, Agalliu I, Thurston SW, Coull BA, Checkoway H. Smoothing in occupational cohort studies: an illustration based on penalised splines. *Occup Environ Med.* (2004) 61:854–60. doi: 10.1136/oem.2004.013136

33. Binder H, Sauerbrei W, Royston P. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Stat Med.* (2013) 32:2262–77. doi: 10.1002/sim.5639

34. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM, et al. Adjustment for continuous confounders: an example of how to prevent residual confounding. *CMAJ.* (2013) 185:401–6. doi: 10.1503/cmaj.120592

35. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2019).

36. Gu C. Smoothing spline ANOVA models: R package gss. *J Stat Softw*. (2014) 58:1–25. doi: 10.18637/jss.v058.i05

37. Kooperberg C. *Polspline: Polynomial Spline Routines*. R package version 1.1.20 (2022). Available online at: https://CRAN.R-project.org/package=polspline (accessed July 31, 2022).

38. Cox NJ. RCSPLINE: Stata Module for restriced cubic spline smoothing. Statistical Software Components S456884 (2007).

39. IBM SPSS Statistics. *Spline Regression With Estimated Knots in SPSS*. (2020). Available online at: https://www.ibm.com/support/pages/spline-regression-estimated-knots-spss#:~:text=Regression%20models%20in%20which%20the,with%20the%20SPSS%20REGRESSION%20procedure

40. SAS Programming Documentation. *The TRANSREG Procedure*. (2019). Available online at: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_transreg_details03.htm