# Appendix

For readability we adopt the following notation throughout the manuscript:

* The trials as j = 1,2, …, n
* Trial participants as i = 1,2, …,,
* Continuous effect modifier: X
* Binary treatment indicator: T with value 0 for the control group and 1 for the experimental group.
* The true association of X with the outcome: f(X)
* Smoothed estimated function:
* the boundaries of X and the boundaries of X per trial j
* The number of (inner) knots:
* w є [1,…, κ +1] the intervals defined by the knots
* g: a link function and its inverse function
* B(X;d) a basis function of dth degree

## Polynomial regression and truncated power series

In general, to model the association between an independent variable X and an outcome Y, generalised linear models (GLMs) are used. In case of non-linear associations, transformed versions of X can be used instead of X. For instance, the statistical model for a GLM with link function g and with a d-degree polynomial of X is:

However, a global function over the full range of X may have poor fit near the boundaries due to instability of the estimated polynomial in this area. To avoid these issues, piecewise polynomials may be preferred to global functions. The model for a d-degree polynomial for interval *w*, between knot *tw* and *tw+1*, would be:

These piece-wise polynomials, when fitted in two consecutive intervals, will show different predicted values at the boundaries of the intervals (*i.e.* at the knots), thus their functional shape will be discontinuous. For this reason, we may use restrictions to “connect” interval-specific polynomials. One convenient solution is to fit a global polynomial, and model the deviations from this globally defined shape within truncated parts of X. Thereto, each basis function is a polynomial with one term. Given a non-decreasing sequence of knots a truncated power series basis is defined by the following basis functions:

and the statistical model for the association between X and Y is:

The + subscript denotes that for a given z

The first term in equation (4) generates the global polynomial, often called the “basic” polynomial, whereas the second term, often called the “secondary” polynomial, is modelling the deviations from it. The resulting splines, using truncated power series basis functions, are often called polynomial ‘regression’ splines [1]. The term truncated reflects the fact that the intervals for the power series in the secondary polynomial are shortened to produce estimates only in sub-domains of X. A disadvantage of truncated power series is that they can still show erratic behaviour near the ranges of X.

# Regression splines

## **Natural** or restricted splines

A solution to this erratic behavior near the boundaries is to restrict the truncated power series to be linear near the boundaries of X [2]. These splines are often called natural or restricted (polynomial) splines. Given a non-decreasing sequence of knots the statistical model is given as:

where

and for *w* є [2, κ-1]

Harrell shows that restricted cubic splines can also be written as truncated power series with a linear “basic” polynomial, by dividing the basis functions by [3]. Therefore, an equivalent statistical model to (5) may be written as follows:

The number and location of the knots may be based on clinical knowledge or on descriptive statistics. For instance, Harrell suggests the use of quantiles and advocates that four knots in most cases are adequate [2, 4], see Table 1. Depending on the available sample size and required complexity of the functional shape we may use a different number of knots.

In Figure 20(a) we show the basis functions - scaled by for the restricted cubic splines approach with 5 knots placed at the 5%, 27.5%,50%, 72.5%, 95% quantiles. In our single study example, we used restricted cubic spline transformations of X both as main effects and as interactions with the treatment. Following Harrell’s suggestion, we placed 5 knots at values corresponding to 5%, 27.5%, 50%, 72.5% and 95% quantiles of X.

.

## B-splines

B-splines are another commonly applied spline approach. They are based on a parametrisation of polynomial cubic splines. Given a non-decreasing knot sequence and X є , the dth degree B-splines basis functions are calculated by the following algorithm proposed by De Boor [5].

First, d additional knots are generated before and d additional knots after . These are often called outer knots and their choice is arbitrary. We can set them to be equidistant or even equal to the boundary values and of X. A new knot sequence is generated, where:

are the left outer knots or endpoints,

,

, , …, the inner knots

,

the right outer knots

Within each interval *w* a zero-degree B-spline is calculated. Zero-degree B-splines are step functions equal to 1 within an interval and 0 otherwise.

All succeeding basis functions, with degree >1, are calculated using the following formula:

where [1, 2, …, d]. For example, the first degree is calculated using the zero degree B-splines, and the second degree from and so on, using formula 7.

Three variations of B-splines based on the inner knot positionings have been proposed. B-splines with a uniform knot vector use equidistant knots and are the most typically applied B-splines [6]. B-splines with an open uniform knot vector also use equidistant knots but they allow analysis of closed curves. Non-uniform B-splines use non-equidistant knots, placed at positions of the researcher’s choice. To our topic, uniform and non-uniform B-splines are the most relevant. Non-uniform B-splines may reflect the a-priori knowledge of a researcher over the underlying complexity of the functional form and/or distribution of the continuous variable.

In Figure 5(b) we show the basis functions of a 2nd degree B-spline with 4 equidistant knots; 2 inner knots plus the boundaries [α, β], placed at values α=0, 0.33, 0.66 and 1=β.

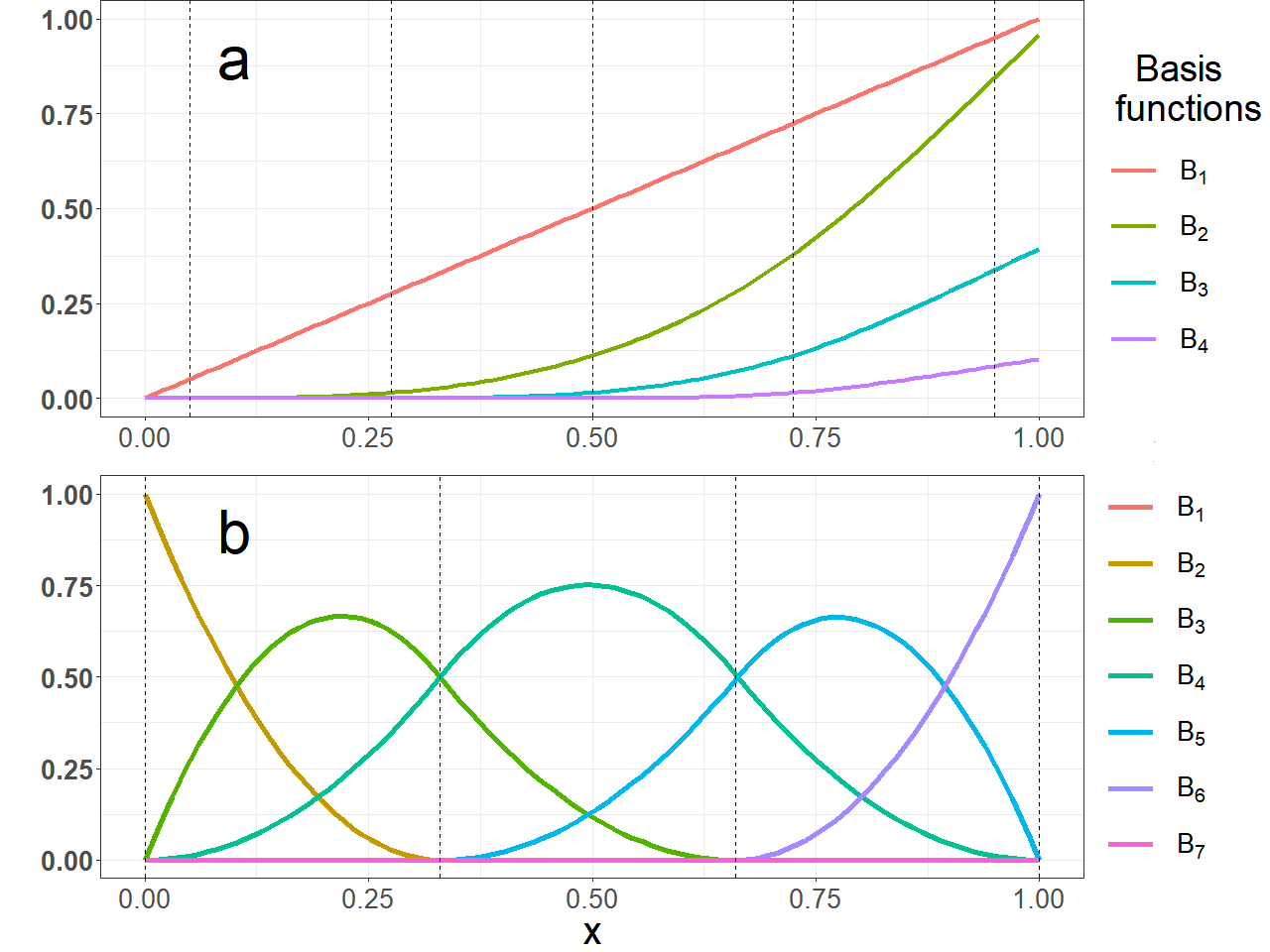


Figure 5. approximately here

# Penalised splines

The two commonly applied penalised splines that we discuss, P-splines and Smoothing splines, increase the number of knots to a large set (usually, 10-40) or even to be equal to the number of observations. This way they circumvent the problem of choosing the number and positions of the knots. Since estimating one parameter for each observation would clearly lead to a perfect fit and thus generate functional shapes with extreme variability, penalised splines introduce in their optimisation functions a penalty term () multiplied by a non-negative , often called a tuning parameter. As the term “tuning” implies, changing the value of changes the magnitude of the penalisation.

In GLMs the estimation of the regression coefficients is accomplished through optimisation of functions of . For Gaussian outcomes the least squares optimisation is estimating the that minimise the squared distance of the predicted and the observed values of the outcome, while for outcomes belonging to the exponential family (Gaussian, Binary, Poisson etc) we estimate the maximising the likelihood function of . Adding a penalty term () results in the following optimisation equations:

* Least squares approach

* Maximum Likelihood approach

Penalised splines circumvent the problem of knot selection, but at a cost. By using a penalty in their optimisation function, they introduce bias in their estimate in order to obtain a more stable solution. Further, in both P-splines and Smoothing splines the tuning parameter must be specified. Too high or too low values of may lead to over- or undersmoothing respectively. Several approaches have been proposed in order to determine the “optimal” , such as Akaike’s information criterion AIC [7], “leave one out” generalised cross-validation (GCV) [8] or mixed-effects modelling [9]. These processes are automated in most of the statistical packages. Briefly, when using the AIC, a series of models fitted with different values is compared and the one with the lowest AIC is selected. “Leave one out” GCV is an iterative process, the algorithm goes as follows: 1) one observation is omitted 2) a model is fitted 3) using the model a prediction of the omitted value is generated and 4) the distance between the observed and predicted value is calculated. This procedure is repeated for each observation and for a series of values. The that minimizes the GCV minimizes the sum of the squared distances, i.e. the GCV score, is selected. In Bayesian/mixed effects modelling approach the penalty term is estimated in a similar way as random effects parameters.

## P-splines

A specific type of penalised splines, P-splines, proposed by Eilers and Marx [10], is a penalised version of B-splines, using a specific penalty term based on the sum of p-order differences between the coefficients of two consecutive intervals . The first order differences are defined as follows: , but Eilers and Marx propose the use of second order differences, which are the first order differences of the first order differences

.

Note that the degree of the underlying B-splines may be different from the order of the differences. A common combination is that of a third-degree B-spline with a second order difference. Using a penalty based on a zero-degree order difference results in the ridge penalty [11]. Note that in some occasions penalised splines and penalised B-splines are misinterpreted as P-splines, but not all penalised B-splines or penalised splines are P-splines. For instance, ordinary B-splines may be fitted using a Smoothing splines approach, but this does not make them P-splines, unless they are penalised using the approach suggested by Eilers and Marx.

P-splines are based on equidistant knots. It is possible to use a knot sequence that is not evenly spaced; but in this case, weights need to be introduced [6, 9]. As P-splines with non-equidistant knots are rarely used in practice we don’t consider them in this article.

## Smoothing splines

Smoothing splinesare another member of the family of penalised spline methods. Similar to P-splines the idea is to increase the number of knots, but this time to be equal or approximately equal to the number of observations. O’ Sullivan [41] suggested that a penalty based on Reinsch’s integral of the second derivative of , where is a cubic spline, multiplied by a tuning parameter, has good smoothing properties. This results in the following penalty term for Smoothing splines: .

## Single study artificial data-set simulation functions

The risk of mortality per participant in the single study data-set was generated using the following formulas:

* For the control group

* For the treated group

Equivalently equations (1) and (2) can also be combined into a single equation:

where would be the association of BMI with mortality risk for the control and the additive effect of the treatment.

## Multiple studies artificial data-set simulation functions

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | True underlying function forms | | BMI Ranges | | | | |
| Scenarios | Control | Treated | 1st Study | 2nd Study | 3rd Study | 4th Study | 5th Study |
| First scenario |  |  | [18.5,40] | [18.5,40] | [18.5,40] | [18.5,40] | [18.5,40] |
| Second scenario |  |  | [18.5,27] | [21.2,30.2] | [24.5,33.5] | [27.8,36.7] | [31.40] |
| Third scenario |  |  | [18.5,27] | [21.2,30.2] | [24.5,33.5] | [27.8,36.7] | [31.40] |

The risk of mortality per participant and per study j in the three multiple studies scenarios was generated using the following formulas:

# Appendix references

1. Hastie T, Friedman J, Tibshirani R. Basis Expansions and Regularization. In: The Elements of Statistical Learning. New York, NY: Springer New York; 2001. p. 115–63. doi:10.1007/978-0-387-21606-5\_5.

2. Harrell FE. General Aspects of Fitting Regression Models. In: Regression Modeling Strategies. Cham: Springer International Publishing; 2015. p. 24–6. doi:10.1007/978-3-319-19425-7\_2.

3. Harrell , FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Cham: Springer International Publishing; 2015. doi:10.1007/978-3-319-19425-7.

4. Stone, C. J. Comment: Generalized additive models.

5. de Boor C. Package for Calculating with B-Splines. SIAM J Numer Anal. 1977;14:441–72.

6. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R. BMC Med Res Methodol. 2019;19:46.

7. Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika. 1973;60:255–65.

8. Craven P, Wahba G. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math. 1978;31:377–403.

9. Wood SN. Introducing GAMs. In: Generalized Additive Models. 2nd edition. Chapman and Hall/CRC; 2017. p. 161–94. doi:10.1201/9781315370279-4.

10. Eilers PHC, Marx BD. Flexible smoothing with B -splines and penalties. Stat Sci. 1996;11:89–121.

11. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. 1st edition. Cambridge University Press; 2003. doi:10.1017/CBO9780511755453.