Dealing with continuous variables and modelling non-linear associations in healthcare data: practical guide

Pedro Lopez-Ayala,[1] Richard D Riley,[2,3] Gary S Collins,[4] Tobias Zimmermann[1,5,6]

[1]Cardiovascular Research Institute Basel (CRIB) and Department of Cardiology, University Hospital Basel, University of Basel, Basel, Switzerland

[2]School of Health Sciences, College of Medicine and Health, University of Birmingham, Birmingham, UK

[3]National Institute for Health and Care Research (NIHR) Birmingham Biomedical Research Centre, UK

[4]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, UK

[5]Bloomsbury Institute of Intensive Care Medicine, University College London, London, UK

[6]Intensive Care Unit, Department of Acute Medicine, University Hospital Basel, CH-4031 Basel, Switzerland

Correspondence to:
T Zimmermann
tobias.zimmermann@usb.ch
(or @tzimmermann@fediscience.org on Mastodon;
ORCID 0000-0002-6862-5205)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2025;390:e082440
http://dx.doi.org/10.1136/bmj-2024-082440

Accepted: 31 January 2025

Proper handling of continuous variables is crucial in healthcare research, for example, within regression modelling for descriptive, explanatory, or predictive purposes. However, inadequate methods are commonly used. This article highlights the importance of appropriately handling continuous variables, and illustrates the consequences of categorisation. This article also explains why assuming a linear relationship between the independent and dependent variable might be inappropriate, and describes how to use splines or fractional polynomials to model non-linear relationships.

Continuous variables such as age, vital parameters, or biomarker concentrations are abundant in healthcare research. Whether the research aim is to describe (eg, whether age is associated with six month mortality after a diagnosis of covid-19), explain (eg, does the effect of a new cancer drug vary according to the value of a continuous biomarker), or predict (eg, does adding blood pressure to the model improve the prediction accuracy of risk for cardiovascular disease),[1] researchers should appropriately model the association between independent and dependent variables. Researchers frequently encounter this challenge, for example, when fitting a regression model. But too often, the approaches used are inadequate, including submissions to *The BMJ*.[2] In this article, we provide an overview of the current state of handling continuous variables in healthcare research. We discuss the drawbacks of categorising a continuous variable, and the potential limitations of assuming a linear relationship between independent and dependent variables. We discuss existing reviews of current practice and then outline two recommended approaches that allow for non-linear relationships: fractional polynomials[3-5] and splines,[6-8] with a particular focus on restricted cubic splines. Box 1 provides a list of key terms, and the key messages are illustrated throughout using the publicly available acute bacterial meningitis dataset,[9] where we examine the association between levels of glucose in cerebrospinal fluid (CSF) and a diagnosis of acute bacterial meningitis. Special emphasis is also placed on the interpretation, graphical presentation, and reporting of models using non-linear transformations of continuous variables. To facilitate broader adoption of a flexible modelling approach, we provide R and Stata code for reproducing the case study.

SUMMARY POINTS

Continuous variables (eg, age, blood pressure, biomarker levels, birthweight) frequently need to be analysed in healthcare research studies

When modelling the association between a continuous independent variable and a dependent variable, many researchers categorise the independent variable or assume (often incorrectly) a linear relationship. Such approaches could lead to fallacious results

Non-linear relationships are common in nature and require flexible modelling techniques to be adequately captured

Splines and (fractional) polynomials are two widely used tools to allow non-linear relationships between independent and dependent variables

R and Stata code accompanying this article will enable readers to more easily make use of splines or other functions to model non-linear relationships in their own healthcare research studies

Case study

The acute bacterial meningitis dataset contains 501 patients with acute meningitis who were admitted to hospital at Duke University Medical Center (Durham, NC, USA) between January 1969 and July 1980.[9] We use this dataset to demonstrate the impact of different methods of handling continuous variables and how to visualise and report results when using restricted cubic splines (rcs). We focus on the diagnostic ability of glucose in CSF for differentiating between acute bacterial and viral meningitis. Logistic regression is used to build diagnostic models that predict acute bacterial meningitis (binary dependent variable, Y). Based on previous medical knowledge, glucose in CSF (csf_gl), age, sex, and CSF leucocytes (csf_leuk) are considered important diagnostic factors (independent variable, x). CSF glucose, age, and CSF leucocytes are treated as continuous variables, sex as a categorical variable with two levels (male/female). Three different logistic regression models are fitted with varying transformations of the continuous variables: linear (model 1), categorical (model 2), and non-linear using restricted cubic splines with three knots (model 3). To keep the focus on flexible modelling approaches, we perform a full case analysis for the case study despite the presence of missing values in the dataset (supplemental table 1). For completeness, we briefly address flexible multiple imputation allowing for non-linearity during the imputation process in the supplemental material. Throughout the article, we refer to this case study and present examples based on this dataset.

Box 1: Key terms

Linearity

A relationship between two variables is linear if every unit change in the independent continuous variable leads to a constant change in the dependent variable (potentially on a transformed scale, such as logit-risk or log-hazard). The graphical representation of such an association is a straight line.

Categorisation

Continuous variables such as age or biomarker levels are split into distinct groups based on arbitrarily chosen cut-off points. An example of categorisation would be separating individuals into three age groups: <30, 30-60, and >60 years. Dichotomisation is an extreme form of categorisation in which only two groups are created on the basis of values above or below the chosen cut-off point.

Non-linearity

A relationship between two variables is non-linear if every unit change in the independent continuous variable does not lead to a constant proportional change in the dependent variable. The graphical representation of such a relationship is a curved line. Non-linear relationships are very common in nature but require flexible modelling approaches to be adequately represented in statistical models. The association between heart rate and adverse events is an example of a non-linear (U shaped) relationship, since neither a heart rate too low, nor one too high is particularly healthy.

Polynomials

Polynomials involve coefficients raised to whole number exponents that are used for fitting non-linear relationships in regression models, with their exponent determining the shape of the relationship. Fractional polynomials differ from that by allowing coefficients raised to powers not restricted to whole numbers.

Splines

Splines facilitate modelling non-linear relationships by splitting a continuous independent variable into intervals and fitting piecewise polynomial functions in each interval, which are smoothly joined at connection points called knots and ensure continuity of the relationship. When polynomials with a power of x^3 are used, they are called cubic splines.

Restricted cubic splines

Restricted cubic splines are a particular kind of splines, in which cubic polynomials are fitted to the inner intervals, while linearity is forced in the outermost intervals (before the first knot and after the last knot). This approach can improve stability if there are few datapoints and outliers at the tail ends of the distribution of the modelled variable and reduces the number of degrees of freedom used.

Degrees of freedom

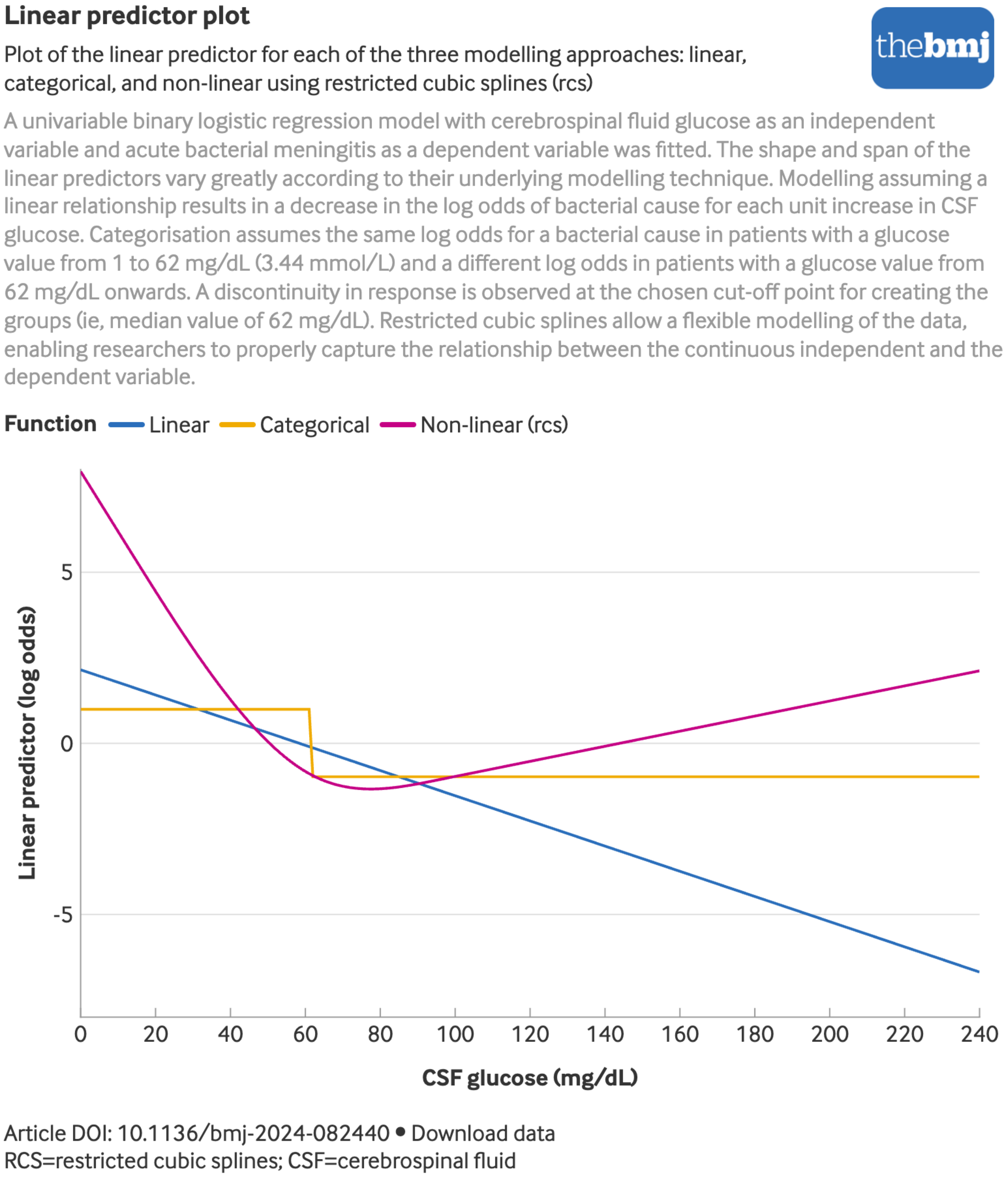
Available degrees of freedom refer to the number of independent pieces of information that can be used to estimate the parameters of a statistical model. Degrees of freedom are determined by the effective sample size (eg, number of participants and events). A higher number of available degrees of freedom allows for more flexible modelling.

Methods for modelling continuous variables

When modelling the association between a continuous independent variable and a dependent variable, researchers often choose between one of the following three techniques: assume and model a linear relationship between the independent and dependent variable; categorise the independent variable (ie, grouping values into two or more categories); or allow for a non-linear relationship by using flexible modelling techniques. We now introduce these different approaches and discuss their limitations.

Linear relationships

Imposing a linear relationship means that every unit change in the independent continuous variable will lead to a constant proportional change in the dependent variable. For example, a change in age from 30 to 31 years will have the same effect as a change from 90 to 91 years. Linear relationships are simple to model and are relatively easy to interpret, which contributes to their popularity—along with it being the default approach—that is, if a model is fit with the help of popular statistical software without otherwise specifying its formulation. However, genuine linear relationships are rare in nature. Assuming linearity when the relationship is in fact non-linear could lead to misleading results. In prediction modelling, forcing a linear relationship between an independent and dependent variable can lead to a reduction in model performance, the extent of which depends on the degree of deviation from linearity of the true relationship and the strength of the independent variable in the model.[6 10] If an independent variable is not a particularly strong predictor, using a linear term for modelling might not substantially affect the model's performance. However, if the variable is known to have an important role (eg, age in a model predicting mortality), the effect of incorrectly specifying the true relationship will be more pronounced. Likewise, departures from linearity in the true underlying relationship between a continuous independent variable and dependent variable can substantially reduce the statistical power to detect clinically important differences between treatment groups in randomised controlled trials.[11]



Linear predictor plot

Plot of the linear predictor for each of the three modelling approaches: linear, categorical, and non-linear using restricted cubic splines (rcs)

A univariable binary logistic regression model with cerebrospinal fluid glucose as an independent variable and acute bacterial meningitis as a dependent variable was fitted. The shape and span of the linear predictors vary greatly according to their underlying modelling technique. Modelling assuming a linear relationship results in a decrease in the log odds of bacterial cause for each unit increase in CSF glucose. Categorisation assumes the same log odds for a bacterial cause in patients with a glucose value from 1 to 62 mg/dL (3.44 mmol/L) and a different log odds in patients with a glucose value from 62 mg/dL onwards. A discontinuity in response is observed at the chosen cut-off point for creating the groups (ie, median value of 62 mg/dL). Restricted cubic splines allow a flexible modelling of the data, enabling researchers to properly capture the relationship between the continuous independent and the dependent variable.

Article DOI: 10.1136/bmj-2024-082440 • Download data

RCS=restricted cubic splines; CSF=cerebrospinal fluid

Fig 1 | Linear predictor plot for three modelling approaches to analyse continuous variables in a case study of cerebrospinal fluid glucose and acute bacterial meningitis. An interactive version of this graphic is available at https://public.flourish.studio/visualisation/22342870/

Table 1 | Model specifications of three logistic regression models fitted with different transformations of continuous variables as part of a case study to predict acute bacterial meningitis

Variables	Coefficients
Model 1: linear	
Intercept	0.484
csf_gl	−0.026
Age	0.007
sex=male	−0.218
csf_leuk	0.002
Model 2: categorical	
Intercept	0.783
csf_gl_cat=≥62 mg/dL	−1.701
age_cat=≥ 3 years	−1.451
sex=male	−0.159
csf_leuk_cat=≥295.5 count/mm^3	1.938
Model 3: non-linear (restricted cubic splines)	
Intercept	10.057
csf_gl	−0.212
csf_gl’	0.133
age	−0.405
age’	2.916
sex=male	−0.308
csf_leuk	0.002
csf_leuk’	−0.005

CSF=cerebrospinal fluid; csf_gl=CSF glucose; csf_leuk=CSF leucocytes; *_cat=categorised versions of continuous variables. 62 mg/dL=3.44 mmol/L.

Case study

Figure 1 shows the univariable linear predictor of CSF glucose as an independent variable and acute bacterial meningitis as a dependent variable for each of the three modelling approaches (linear, categorical, and non-linear). A straight line relationship can be observed with a constant (in this case negative) proportional change in log odds per unit increase in glucose (blue line). Consequently, under the assumption of linearity, a continuous variable has one coefficient (table 1, model 1).

In model 3, each continuous variable modelled with a three knot restricted cubic spline (rcs) has two coefficients—one for each inner interval. In a model using rcs, coefficients alone cannot be interpreted sensibly. In model 3, names of continuous variables (eg, csf_gl and csf_gl’) represent the different spline segments. The intercept of a model is the expected value of Y when all x=0.

Categorisation

Based on the mistaken belief that grouping aids interpretation, continuous variables are often categorised. However, categorisation creates implausible step-functions and has faced widespread criticism owing to multiple issues.[10 12-15] Firstly, by categorising a continuous variable, information is lost and statistical power to detect an association between the independent and dependent variable is reduced.[12] Equally, the performance of the derived prediction model will be severely impaired, leading to suboptimal predictions and ultimately flawed decision making. This affects both regression methods [10] and machine learning classifiers.[16] Different degrees of misclassification exist when it comes to categorisation; fewer categories—that is, less granularity in grouping—leads to a greater loss of information. Therefore, dichotomisation—that is, creating only two groups (eg, by splitting at the sample median)—introduces the greatest loss.[12]

Secondly, categorisation assumes a discontinuity in response as interval boundaries are crossed. Such discontinuities can be found in instances where time has a crucial role (such as stock price drops) or in cases governed by legal regulations (such as voting eligibility at 18 years of age). However, they do generally not occur in nature. Additionally, categorisation assumes that the association between the independent and dependent variable is constant within categories, including either end of the same category. This assumption makes little sense from a physiological perspective: modelling age as a categorical independent variable with two levels, ≥55 and <55 years, with death as a dependent variable assumes that a 55-year-old has the same association with death as an 80-year-old but not a 54-year-old. Thirdly, cut-off points for creating groups are often chosen arbitrarily and lack rationale.[17] On the other hand, data-driven approaches for categorising a continuous variable (eg, the median value) are cohort dependent. Hence different cut-off points will be used for the same variable across different studies, making comparisons difficult. Among all data-driven methods, the so-called optimal cut-off point method (also known as the minimum P value approach) is arguably the worst. It has been shown to inflate type I error rates and results do not replicate across different studies. This method has long been a major issue in prognostic research and biomarker literature, and its flaws were already pointed out 30 years ago in the field of oncology.[18]

Furthermore, categorisation of continuous variables fails to use all available information, leading to problems with residual confounding that are often neglected.[12 13 19 20] Finally, as previously mentioned, the issues with categorisation are exacerbated in dichotomisation. The creation of more categories, however, only partly alleviates the issue. A continuous variable categorised into six groups will spend five degrees of freedom whereas modelling it with a restricted cubic spline with three knots spends two degrees of freedom, and in most cases captures the underlying association more appropriately. This point is particularly relevant for prediction modelling and sample size calculation.[21]

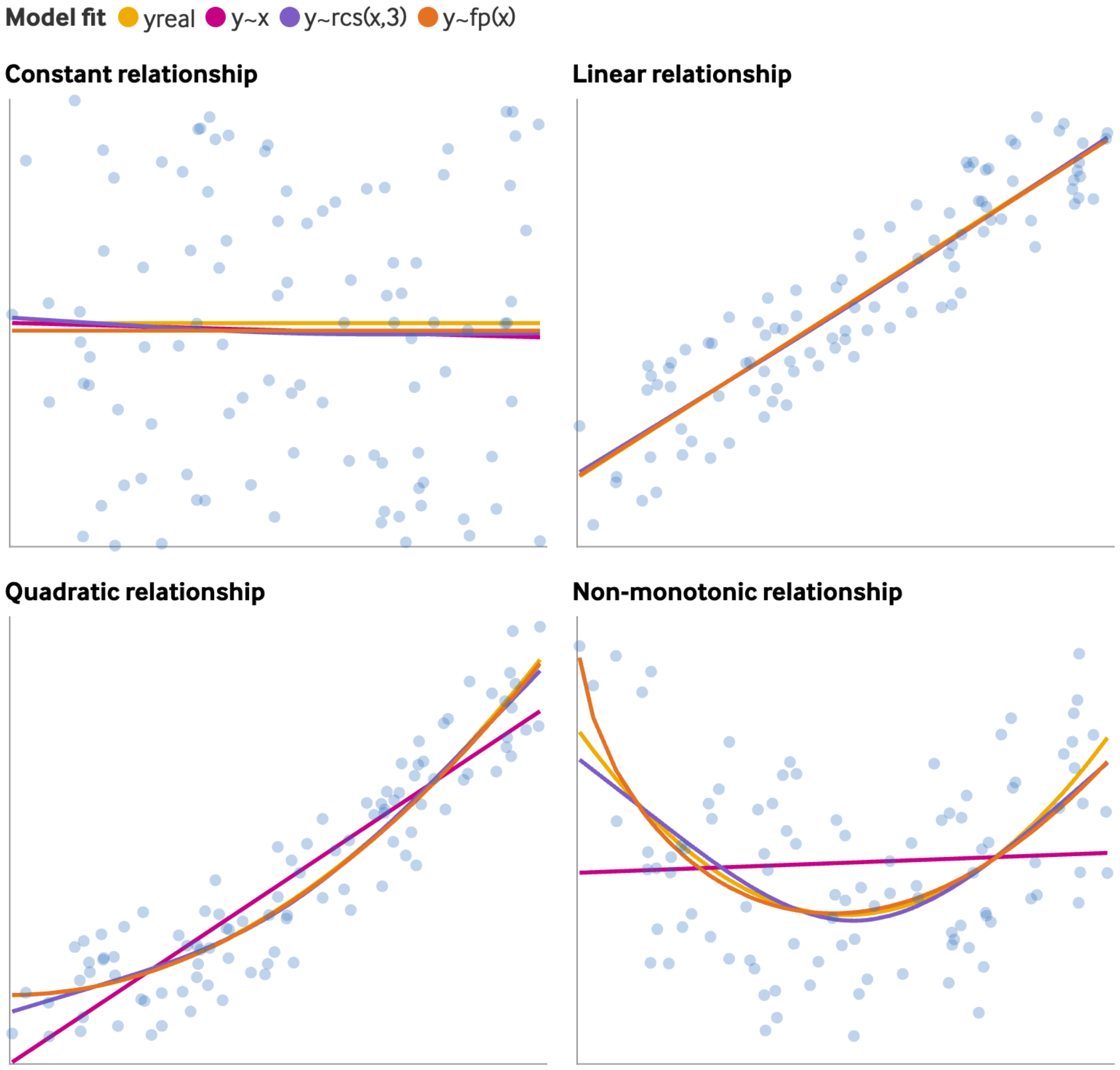
Case study

When a continuous variable is categorised into *n* groups, it is often represented as a set of dummy variables, also known as indicator variables. Each group of the categorical variable is assigned its own coefficient, except for one reference group which is not explicitly estimated. Therefore, the categorised variable has (*n*−1) coefficients associated with it. The yellow line in figure 1 shows the step-function created by categorisation (more specifically dichotomisation) of CSF glucose at the median (62 mg/dL (3.44

Different functional forms fitted on simulated data

Illustration of fitting errors for simulated relationships between a continuous independent and the dependent variable.

Data were sampled from a uniform distribution and four different relationships (constant, linear, quadratic, and non-monotonic) were fitted (yellow line, ie, true relationship). For each, a linear function (y~x), a non-linear function using restricted cubic splines (y~rcs(x,3)), and a non-linear function using fractional polynomials (y~fp(x)) are shown for comparison. The closer the other three lines are to the yellow line, the better the fit (ie, the function used for fitting the data more accurately describes the real underlying relationship). Modelling a constant or linear relationship (first two panels) in a non-linear way will only lead to a mild error owing to overfitting. However, fitting a linear model when the true relationship is non-linear (last two panels) can lead to severe errors.



Constant relationship

Linear relationship

Quadratic relationship

Non-monotonic relationship

Article DOI: 10.1136/bmj-2024-082440 • Download data

rcs=restricted cubic splines; fp=fractional polynomials; yreal=true relationship; y=dependent variable; x=independent continuous variable

Fig 2 | Fitting errors for simulated relationships between a continuous independent and the dependent variable. An interactive version of this graphic is available at https://public.flourish.studio/visualisation/22356795/

mmol/L)), indicating lower log odds for patients with CSF glucose levels above the chosen cut-off point. All patients below the cut-off point share the same log odds for a diagnosis of bacterial meningitis, just as all patients above the cut-off point share the same log odds; independent of what exactly their CSF glucose level is. Dichotomisation (n=2) of a variable results in one coefficient (table 1, model 2)

Non-linear relationships

As an alternative to categorisation or imposing linearity, flexible modelling strategies using fractional polynomials or splines (see Approaches for non-linear modelling), for example, allow the modelling of non-linear relationships between a continuous independent variable and a dependent variable.[22] A non-linear relationship allows the effect of a one-unit change in the independent variable to vary across its spectrum of values. For example, a change in age from 30 to 31 years might have little impact on risk, whereas a change in age from 90 to 91 years might have a substantial impact on risk. Truly linear relationships are rare in nature, so anticipating non-linearity as a default is generally a sensible approach. Multiple clinical studies have investigated natural associations such as those between body mass index and mortality[23] or between high density lipoprotein levels and stroke[24] and have demonstrated important non-linear relationships between independent and dependent variables that would otherwise have been lost. Assuming non-linearity when the relationship is constant or linear will only lead to slight overfitting in most cases (fig 2). In scenarios with a very low number of cases, this error might increase (see Differences and limitations).

Case study

To display non-linear relationships, CSF glucose was modelled using restricted cubic splines with three knots. Default knot positions, as chosen by the software package, are based on quantiles of the variable distribution. For CSF glucose in this dataset, knots are placed at 11.8, 62.0, and 95.0 mg/dL (the 10th, 50th, and 90th quantiles, respectively). The purple line in figure 1 clearly reveals a non-linear relationship between CSF glucose and a diagnosis of bacterial meningitis. It seems that patients towards the lower levels and towards the higher levels of CSF glucose have increased odds for a diagnosis of bacterial meningitis. The lowest odds for bacterial meningitis can be found in patients with a CSF glucose level of about 75 mg/dL, in contrast to what the models built on linearity assumption and categorisation would suggest. Continuous variables modelled with restricted cubic splines with three knots (k=3) have two (k−1) coefficients each (table 1, model 3), which are non-interpretable by themselves but can instead be nicely interpreted graphically (see Graphical display of non-linear relationships).

Table 2 | Prediction table based on model 3 in figure 4, using restricted cubic splines

CSF glucose level (mg/dL)	Odds ratio (95% CI)
20	200.45 (25.42 to 1580.69)
30	26.70 (7.41 to 96.29)
40	4.39 (2.46 to 7.84)
50	1.00 (1.00 to 1.00)
60	0.35 (0.23 to 0.54)
100	0.33 (0.15 to 0.69)
120	0.58 (0.24 to 1.44)
150	1.41 (0.39 to 5.06)

Data are odds ratios for arbitrarily chosen values across the range of CSF glucose levels in relation to the chosen reference value of 50 mg/dL (2.77 mmol/L), for which the odds ratio is per definition 1 and the confidence intervals collapse at 1.

Current practice of handling continuous variables

The flawed handling of continuous variables is a longstanding problem evident from reviews done in various fields, including prediction modelling, prognostic factor research, and causal inference. Categorisation (especially dichotomisation) is common but linear relationships are also often assumed without thought. For example, in prediction modelling, a 2011 systematic review of 43 risk prediction models for type 2 diabetes showed that 21 (49%) were developed by categorising all continuous independent variables.[25] Similarly, of 14 risk prediction models for chronic kidney disease, only six (43%) were found to use continuous independent variables.[26] A recent systematic review found that among 118 studies developing clinical prediction models using logistic regression, only 18 (15%) assessed linearity or used methods to handle non-linearity. In contrast, categorisation was used in 67 (57%) of the studies.[14] Likewise, in another systematic review evaluating 62 clinical prediction models for blood transfusion in patients undergoing elective surgery, 40 (65%) categorised their continuous variables.[27] Only a fifth of studies examined non-linearity and only three (8%) studies included non-linear relationships in their final model. Issues are also evident in the field of machine learning. For example, among 62 studies in oncology developing 152 prognostic models based on machine learning, 24 (39%) categorised all or some continuous independent variables before modelling.[28 29]

Prognostic factor research can also be subject to the mishandling of continuous variables.[30-32] Many prognostic variables, including physiological and biomarker measurements, are continuous. Too often, these variables are categorised into low and high risk groups on the basis of an arbitrarily chosen cut-off point, often driven by statistical significance. This method has detrimental consequences, such as reducing the power to detect the genuine relationship with the dependent variable, leading to inconclusive or misleading results. This problem has been repeatedly shown by systematic reviews in different fields of medicine, such as cardiology, oncology, and psychology.[33-35]

Similar issues can be found in explanatory research, based on either observational studies or randomised controlled trials. Reporting practices regarding adjustment for continuous confounders are generally poor,[36 37] which is all the more critical because categorising continuous confounders is known to lead to residual confounding.[13 20] In a 2022 cross sectional survey, the authors assessed the frequency of reporting on methods used to adjust for confounding in 537 original research articles. In 45% of the studies, it was not clear how multicategory or continuous variables were treated in the analysis. A 2013 study, Groenwold and colleagues assessed current practice in the reporting of confounding adjustment in 53 papers and found that in 68% of the studies, it was unclear how age as a confounder variable was modelled.[19]

In randomised controlled trials, continuous variables are relevant in various settings: when conditioning on prognostic baseline covariates to improve statistical power and to model conditional treatment effects[38]; when used within stratified randomisation; or when assessing heterogeneity of treatment effects (ie, subgroup analysis, treatment-covariate interactions). In all three scenarios, simulation studies have shown that dichotomisation, categorisation, and assuming a linear relationship lead to large reductions in power when the true relationship is non-linear.[11 39 40] Sullivan and colleagues reviewed 32 trials that used stratified randomisation based on a continuous variable and none had adjusted for continuous values in the primary analysis. The current standard of handling continuous variables in subgroup analyses is even more deficient. In a systematic review of randomised controlled trials in 2016-21, Williamson et al[41] found that among 178 studies with a subgroup formed on the basis of a continuous variable, 169 (95%) dichotomised the continuous variable. Instead, the continuous variable could have been better modelled as a smooth non-linear function and an interaction with the treatment specified.[42 43] Results could have been displayed by showing how the estimated treatment effect from the model varied with the level of the continuous variable (R code provided, see Data and code sharing). Riley et al provide an example where non-linear treatment-covariate interactions would have been missed if an individual patient data meta-analysis had just assumed a linear relationship.[44]

Box 2: Recommendations for modelling multiple continuous independent variables

- Do not categorise continuous variables or automatically assume linearity, because important non-linear relationships might not be fully captured or even completely missed; this recommendation applies independently of whether the aim is to predict, explain, or describe
- Make use of matter based knowledge when fitting the regression model
- Allow flexible non-linearity for variables known to have a non-linear relationship and for variables not known to have an approximately linear relationship
- If degrees of freedom are limited, the adequate modelling of independent variables that are known or expected to be most important should be prioritised; however, whenever possible, other covariates should also be modelled to allow non-linear relationships, to avoid suboptimal predictive performance or residual confounding
- The choice between fractional polynomials and splines can be based on model purpose or personal preference, since results in most cases can be expected to be similar

Modelling multiple continuous independent variables

When working with a model containing multiple continuous independent variables, researchers might wonder which ones should be allowed to have a non-linear relationship with the dependent variable. To answer this, we differentiate between whether the goal is to predict, explain, or describe.[1] There is no one-size-fits-all approach, however the following strategies provide an evidence based approach to modelling multiple continuous independent variables for different purposes (box 2).

Since prediction models aim for the most accurate predictions and interpretability is often not the main consideration, specifying a complex model allowing non-linear relationships for all continuous variables might be the best option; provided a sufficient effective sample size (eg, number of participants and events) is available for the number of model parameters to be fitted (ie, the so-called degrees of freedom is appropriate; see Sample size calculation, supplemental materials). This is relevant especially in the light of so-called black box machine learning algorithms (see Machine learning, supplemental materials). When dealing with a limited number of available degrees of freedom, continuous independent variables that are known or expected to be more important for predicting a dependent variable should be allowed greater flexibility (eg, more knots in spline functions or higher-order polynomials). Assuming linearity in a less important independent variable might only lead to slight underfitting and thus a slight decrease in predictive performance in most cases.

In lieu of relevant subject knowledge, measures of association, such as partial χ^2 tests or marginal generalised rank correlations, can be used to evaluate the predictive importance of each continuous independent variable in the model.[6] This approach provides a foundation for determining which variables should be allowed greater flexibility and effectively allocating the available degrees of freedom. Importantly, assessing the strength of association between independent and dependent variables does not provide any insights about the functional form of the association. In addition, this assessment is separate from variable selection and independent variables should not be removed from the model on the basis of a low strength of association.

A commonly used approach to decide how to best model a particular continuous independent variable is to assess its association with the outcome using scatter plots or descriptive statistics. However, analysing the relationship separately (ie, outside of the modelling process) can lead to an inflated type 1 error.[45] Because researchers are not masked to the dependent variable, so-called phantom degrees of freedom are being spent during the subjective assessment. If these phantom degrees of freedom that represent complexity in the model are simply dismissed, resulting confidence intervals will be too narrow, P values too small, and R^2 too large. In addition, when using a resampling procedure (such as bootstrapping) for internal validation of the prediction model, incorporating the functional selection process among the steps to properly calculate optimism corrected performance can be challenging.[46] Therefore, if assumptions about relationships are made before the modelling process, there should be no post hoc revisions.

Explanatory models often have one independent variable of primary interest and additional variables representing potential confounders. In such a model, the variable of primary interest should be modelled flexibly to best capture its underlying relationship with the dependent variable. Equally, confounders should be modelled to allow non-linearity whenever possible, because imposing a linear relationship between the confounder and dependent variable or categorisation of the confounder can lead to residual confounding.[13 20]

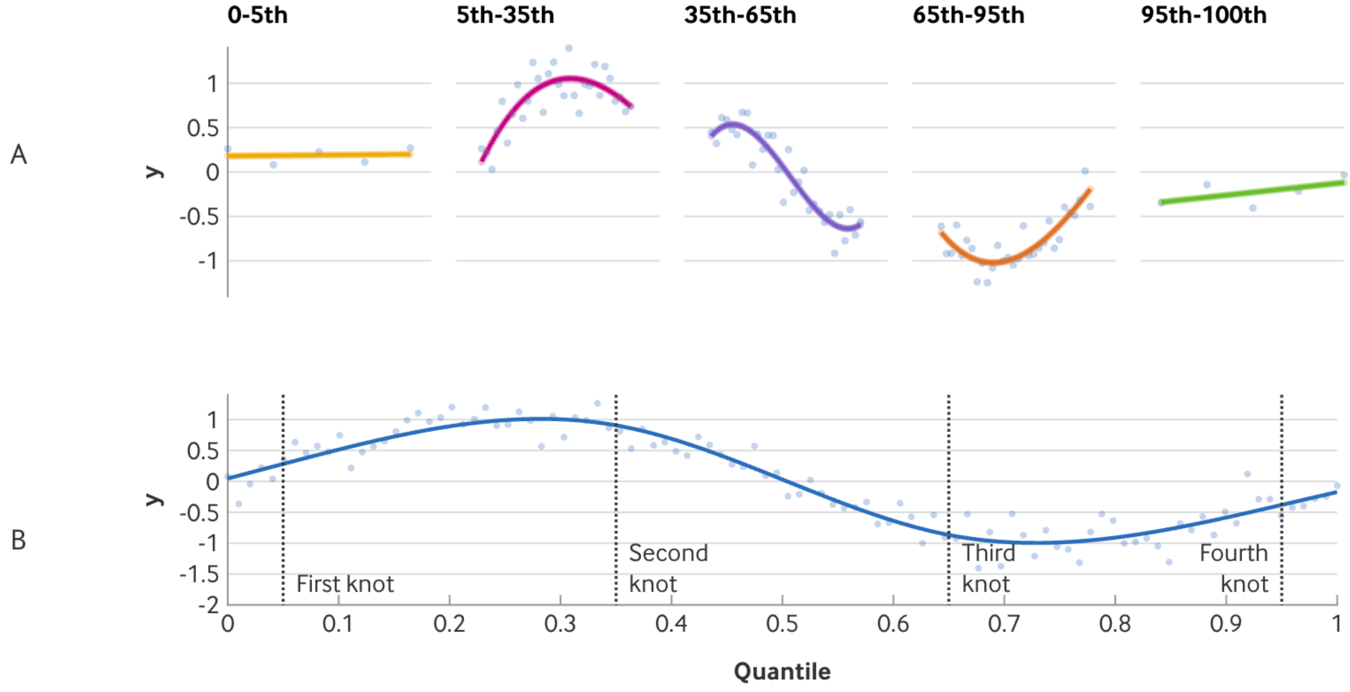
Descriptive modelling is aimed at summarising or representing the data structure in a compact manner. Appropriately capturing the investigated relationship is key. The true underlying relationship between independent and dependent variable is unknown most of the time, so flexible transformations (eg, splines) are favoured over approaches that assume a specific distribution.

Subject knowledge can inform the specifications of the modelling approach (eg, polynomial degree, or a specific parametric distribution); however, there is often no theoretical foundation for assuming a particular relationship between independent and dependent variable. Splines and (fractional) polynomials offer great flexibility independent of

Illustration of restricted cubic splines

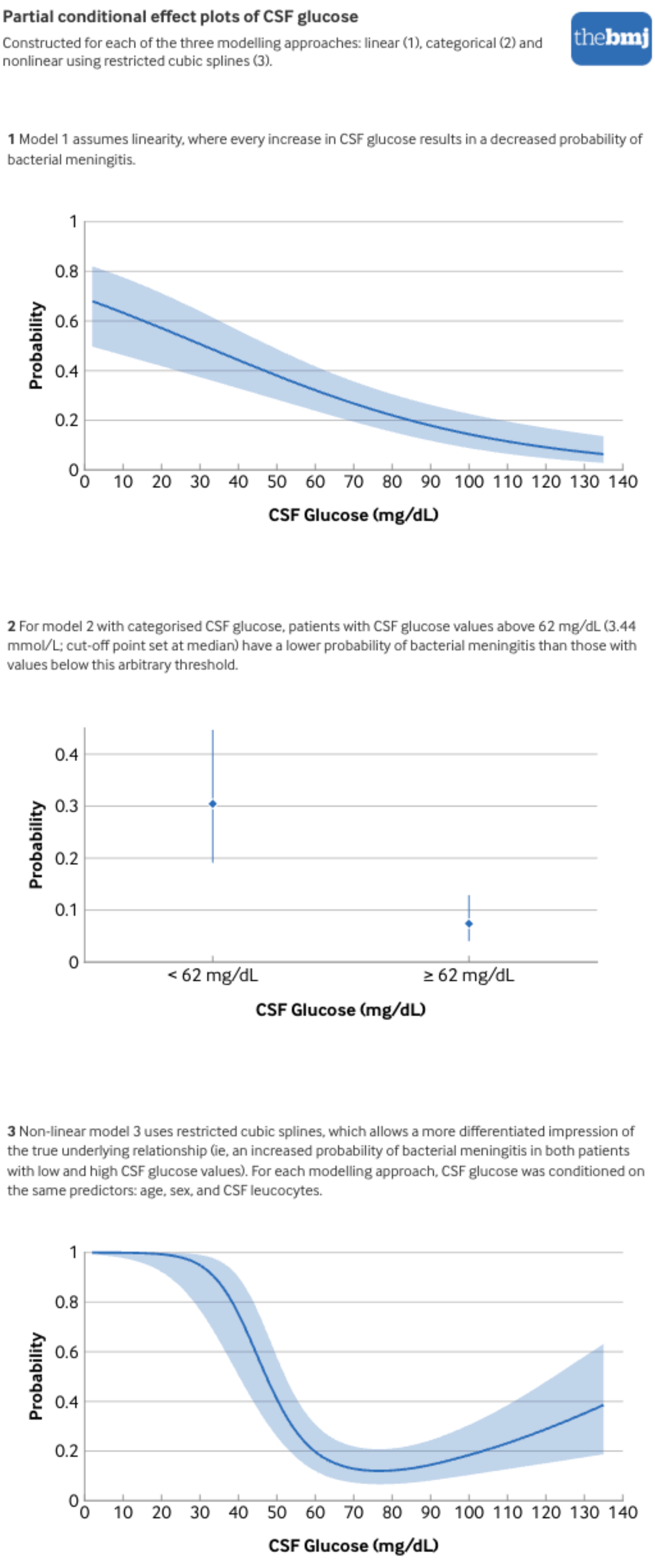
(A) Different elements of a four knot restricted cubic spline: yellow line (0 to 5th quantile) and green line (95 to 100th quantile) of the plot represent the tails that are restricted to a linear function, while the pink, purple, and orange lines in the inner intervals are piecewise cubic polynomials fitted to the data.
(B) End product with smooth connections between the individual functions joining at the defined knot positions.

Dots are simulated data points sampled from a continuous probability distribution



Article DOI: 10.1136/bmj-2024-082440 • Download data

Fig 3 | Restricted cubic splines. An interactive version of this graphic is available at https://public.flourish.studio/visualisation/22357822/



Partial conditional effect plots of CSF glucose

Constructed for each of the three modelling approaches: linear (1), categorical (2) and nonlinear using restricted cubic splines (3).

1 Model 1 assumes linearity, where every increase in CSF glucose results in a decreased probability of bacterial meningitis.

2 For model 2 with categorised CSF glucose, patients with CSF glucose values above 62 mg/dL (3.44 mmol/L; cut-off point set at median) have a lower probability of bacterial meningitis than those with values below this arbitrary threshold.

3 Non-linear model 3 uses restricted cubic splines, which allows a more differentiated impression of the true underlying relationship (ie, an increased probability of bacterial meningitis in both patients with low and high CSF glucose values). For each modelling approach, CSF glucose was conditioned on the same predictors: age, sex, and CSF leucocytes.

Article DOI: 10.1136/bmj-2024-082440 • Download data
CSF=cerebrospinal fluid

Fig 4 | Partial conditional effect plots of cerebrospinal fluid (CSF) glucose for three modelling approaches: (1) linear, (2) categorical, and (3) non-linear, using restricted cubic splines. An interactive version of this graphic is available at https://public.flourish.studio/visualisation/22399040/

the underlying relationship, which avoids making potentially harmful assumptions.

In very small cohorts, the flexible approximation of the relationship between independent and dependent variable can be severely limited by the lack of available degrees of freedom. If the effect size is large or the relationship between independent and dependent variable is firmly established, a parsimonious model can be built. Otherwise, this issue is best dealt with before study commencement by performing adequate sample size calculations.[21]

Approaches for non-linear modelling

Fractional polynomials

As a default in regression models, a linear relationship is being assumed between continuous independent variables and the dependent variable. Historically, when a non-linear relationship was suspected, the continuous variable was transformed using, for example, a quadratic (x^2) or cubic (x^3) polynomial. These low-order polynomials, however, often do not fit the data particularly well and higher-order polynomials tend to have poor fit in the tails of the distribution of the continuous variable. In 1994, Royston and Altman formally introduced fractional polynomials, providing a more flexible parameterisation for continuous variables.[3] As with regular polynomials, power functions involved in the construction of fractional polynomials are defined over the entire range of the variable. Application to a wide variety of models found polynomial functions of the first order (β_1x^{p1}) or second order ($\beta_1x^{p1} + \beta_2x^{p2}$) to be sufficient in most cases. The best suited power transformation x^p for a continuous variable is chosen on the basis of a series of tests. The default starting point is a linear function. A more complex, non-linear function with a set of S={−2, −1, −0.5, 0, 0.5, 1, 2, 3} (x^0 denotes log(x)) as candidates for the exponent p is only chosen if supported by the underlying data, favouring less complex models. Even though the set of exponents is limited, it allows a wide variety of possible forms (eight in the first order and 32 in the second order). Comparison of fit between models with different exponents is based on maximising likelihood. The significance level α defaults to 0.05 but can be changed depending on the application of the model (eg, prediction (relaxed α) *v* explanation (intermediate α)).[47] The function selection procedure uses the principle of a closed test procedure that ensures that the overall type 1 error is close to the nominal significance level.[5]

Splines

With splines, a continuous variable is split into different intervals; within each interval, a polynomial function of the same degree (eg, cubic) is fitted to the data. These individual functions are constrained to join smoothly at the knots that define the intervals (fig 3). The important differentiation between splines and categorisation is that the resulting function remains continuous and the variation of the effect of a change in the independent variable on the dependent variable is not only possible between intervals but also within intervals. The functions that are being fit piecewise to each interval depend on the spline used. The simplest spline function is a linear spline. However, because they usually do not fit non-linear relationships well, non-linear polynomials can be used instead. The most frequently used splines are cubic and different variations exist—for example, smoothing splines, penalised splines, B splines, or restricted cubic splines. All of them have been shown to adequately model non-linear relationships, making any of them a good choice for flexible modelling. An exhaustive description of the different types of splines is provided by Perperoglou and colleagues in their comprehensive review.[7] In this article, we will focus on restricted cubic splines, which offer relatively straightforward interpretation, are widely used in the medical literature, and are well implemented in R, Stata, and SAS.[6 48 49]

Restricted cubic splines, also called natural splines, are cubic piecewise polynomials with a constraint to linearity in the tails (ie, before the first knot and after the last knot) making them more robust.[6 50] Unconstrained cubic spline functions, just like fractional polynomials, can behave poorly in the tails where only few data points are available and outliers have a big impact (fig 2). To work with restricted cubic splines, the number and position of knots need to be specified. The exact position of the knots usually does not have a major impact on results, so putting them at fixed quantiles of the variable's marginal distribution (eg, 10th, 50th, and 90th quantiles for three knots) is appropriate in most cases.[6] The number of knots should be picked with sample size in mind and lies usually between three and five. With an increasing number of knots, the flexibility of the modelled relationship is increased, but at the cost of adding model complexity (increase in degrees of freedom used), as well as an increased risk of overfitting. However, by spending fewer degrees of freedom (number of knots−1) than other cubic spline variants or fractional polynomials, restricted cubic splines generally reduce overfitting. Estimators of model quality, such as the Akaike information criterion, can be used to determine the optimal number of knots since the metric favours simpler models in case of similar model fit.

Case study

For continuous variables modelled with restricted cubic splines, coefficients alone do not provide interpretable information. Instead, the model and its variables can be visualised using partial conditional effect plots. Figure 4 shows, for each of the three fitted models, the predicted probability of having acute bacterial meningitis across the range of CSF glucose values conditioned on the remaining predictors in the model being held constant at their median (continuous variables), their mode (categorical variables), or the lowest value (binary variables). The plot for model 1 (linear, fig 4A) shows a decrease in probability of acute bacterial meningitis with increasing CSF glucose levels, and model 2 (categorical,[22] fig 4B) shows a higher probability of acute bacterial meningitis in patients with CSF glucose <62 mg/dL (3.44 mmol/L). Model 3 (non-linear, fig 4C) shows that lower CSF glucose levels are associated with the highest probability of acute bacterial meningitis with a steep decline in probability towards intermediate levels. Importantly, however, it additionally reveals another increase in probability towards higher levels of CSF glucose.

Alternative approaches

This article focuses on fractional polynomials and splines for non-linear modelling, although alternative approaches exist, such as generalised additive models and locally estimated scatterplot smoothing (loess). Further alternative approaches can be found in the field of machine learning (eg, k-nearest neighbours algorithms, and artificial neural networks). These alternatives are introduced briefly in the supplemental material.

Differences and limitations

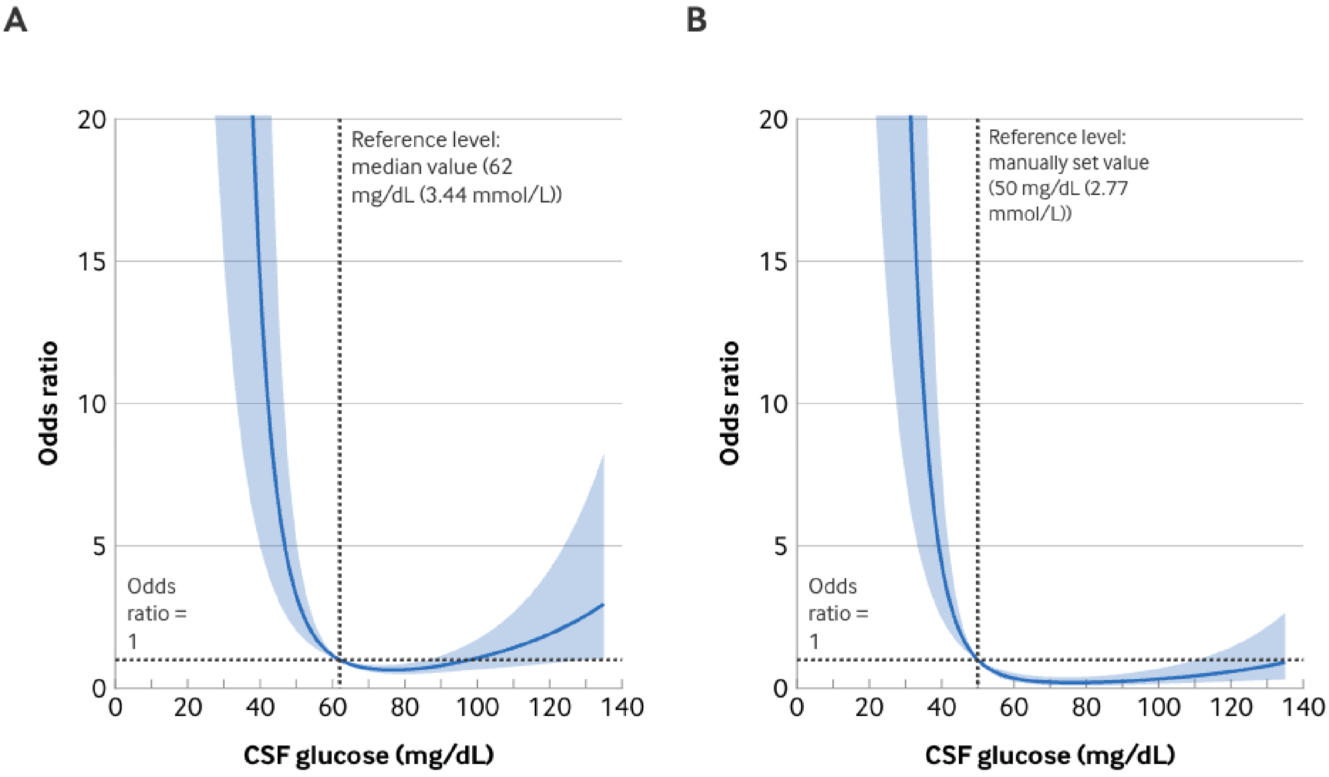
Fractional polynomials and splines both allow modelling non-linear relationships, but important differences exist. Splines are constructed from local piecewise polynomial functions between knots. Fractional polynomials are defined over the entire range of the continuous variable. This approach can make fractional polynomials vulnerable to distortions in their overall shape caused by values in the tails of the variable distribution (eg, a high number of cases with values close to 0). Splines are powerful tools with a high degree of flexibility. However, the functional form of splines can be complex in its expression, because for each continuous variable in a model that is transformed using splines, a specific coefficient gets determined for each interval. When dealing with many continuous variables, for example, when developing and reporting a prediction model, this complexity can become burdensome. The combination of non-interpretability and number of coefficients of a variable transformed with splines, compared to, for example, categorisation which provides easily graspable coefficients, may deter users from applying this technique. In contrast to splines, fractional polynomials offer simpler expression, interpretation, and reporting, making them a more accessible alternative. For both splines and fractional polynomials, graphical representation using partial conditional effect plots and dose-response plots are well suited for the inspection and interpretation of the relationship between an independent and dependent variable.

Previous investigations have revealed that in many settings, fractional polynomials and splines produce similar results.[10 11 39 47] A recent commentary by Sauerbrei et al concluded that there is not enough evidence to unequivocally recommend one transformation over the other.[22] The ultimate decision of which flexible modelling technique to use is left to

Dose-response plot of CSF glucose

Conditional dose-response odds ratio plot for the variable of interest, CSF glucose. The reference value should be chosen based on clinical knowledge.

Conditional values for age (median), sex (male), and CSF leucocytes (median) do not differ between panels A and B. Note that y axis has been truncated to show odds ratios up to a maximum of 20 for visual reasons.



Article DOI: 10.1136/bmj-2024-082440 • Download data

CSF=cerebrospinal fluid

Fig 5 | Conditional dose-response odds ratio plot for the variable of interest, cerebrospinal fluid (CSF) glucose, to determine odds of having a bacterial cause of meningitis. An interactive version of this graphic is available at https://public.flourish.studio/visualisation/22395598/

the preference of analysts and depends on the purpose of the model.

Both polynomials and splines can perform poorly in terms of extrapolation. Restricted cubic splines, forcing linearity in the tail ends, do better in some cases. If a reliable relationship between the independent and dependent variable has been established either from previous investigations or based on theoretical frameworks (more frequently encountered in other areas of research such as physics), then a fully parametric parsimonious model can be built without the use of splines or polynomials.

Graphical display of non-linear relationships

In addition to partial conditional effect plots, dose-response plots are another graphical representation of the modelled relationship between the continuous independent and the dependent variable. They are a good way to report and interpret effect measures of a non-linear continuous variable across its distribution.[51] Dose-response plots are based on predictions conditioned on the remaining variables in the model and, depending on the conducted analysis, allow not only the inspection of the non-linear relationship between independent and dependent variable, but

Table 3 | Specific R packages and functions for fitting non-linear functions

Package	Function	Purpose
{rms}	lsp	Linear spline
{rms}	rcs	Restricted cubic spline
{rms}	gTrans	To specify new polynomial bases, smooth relationships with a discontinuity at one or more values of *x* or grouped categorical variables
{rms}	pol	Non-orthogonal polynomials
{splines}	ns	Generate the B spline basis matrix for a natural cubic spline
{splines}	bs	Generate the B spline basis matrix for a polynomial spline
{pspline}	—	Fit a penalised smoothing spline
{polspline}	—	Comprehensive toolkit for polynomial spline fitting
{mfp}	fp	Defines a fractional polynomial object for a single input variable
{mfp2}	fp2	To state which continuous variables should be modelled using fractional polynomials

Several R packages can be used to fit regression models using splines or fractional polynomials (table 4). All packages come with vignettes, online books, or website supplementals. The {rms} package provides the ability to fit almost any regression model and a wide variety of post-estimation and plotting functions. The package does not fit mixed effect models. The {gam} package and {mgcv} package allow users to fit generalised additive models and generalised mixed models, respectively. The {VGAM} package allows fitting vector generalised additive and linear models. The {mfp2} package implements the multivariable fractional polynomials (MFP) approach—it combines variable (backward elimination approach) and function (fractional polynomials) selection simultaneously in multivariable regression modelling. At the time of writing, the package supports linear, logistic, Poisson, and Cox regression models.

also the determination of, for example, levels with the highest and lowest risk, the optimal dose of a drug, or treatment efficacy. Effect measures (eg, odds ratios) are calculated in relation to a reference point, at which the confidence interval collapses. This reference point defaults to the median of the investigated continuous variable. While this approach may serve as a reasonable default, the choice of reference point should be guided by prior clinical knowledge. In the absence of such information, a clinically meaningful threshold, such as the lower or upper reference limit in the case of biomarkers, can be used as an alternative. The reference point should not be chosen through data-driven methods—that is, trying to achieve the highest effect measure. Whatever reference point is chosen, authors should always provide their rational. Several prominent examples can be found in the literature where dose-response plots were used as the primary reporting measure.[52-54]

In our case study, we determine the lower limit of normal for cerebrospinal fluid (CSF) glucose (50 mg/dL; 2.77 mmol/L) to be important for comparison, so we set the reference value (odds ratio 1) for the dose-response plot accordingly (fig 5). Based on this graph, the odds of different CSF glucose levels can be determined on a continuous scale and sensible interpretations can be made. For example, a patient with a CSF glucose of 40 mg/dL has higher odds (150%) of having a bacterial cause of meningitis than a patient with a CSF glucose of 50 mg/dL (reference), conditioned on the same age, sex, and CSF leucocytes. If visual inspection of the dose-response plot is not sufficient for the purpose and a single measure rather than a continuum is required, an odds ratio can be calculated between two relevant values—for example, the 25th and the 75th percentile. However, the chosen percentile will inevitably involve subjectivity. Prediction tables with odds ratios in relation to the reference point can be created for arbitrary values of the range of the independent variable (table 2).

Summary statistics for best fit

Although we would recommend graphical methods, as described above (see Graphical display of non-linear relationships), for the visualisation and interpretation of models using flexible modelling approaches, there might be situations in which a single test or summary statistic is required. Common metrics for model comparison include the Akaike information criterion, bayesian information criterion, Mallow's C_p criterion, and likelihood ratio test. The Akaike and bayesian information criterions are used for comparison of nested and non-nested models, while the likelihood ratio test is used to compare nested models. For further insights, we encourage readers to consult the extensive literature on these methods.[1 6 22 55] However, these measures should not lead to so-called criterion hacking and post hoc variable selection, since this will not lead to a parsimonious model (ie, a model that achieves an optimal balance between accuracy and simplicity) but overfitting.[56] In contrast, if two different models were prespecified in the analysis plan (eg, a base model and one including interactions, or a model using three knots and one using five knots for the independent variable of interest), these metrics can be used to determine which model is best.

Final comments on case study

The finding that low CSF glucose levels are associated with a higher probability of acute bacterial meningitis is certainly clinically relevant and all three models arrive at this conclusion. However, this finding is only part of the whole picture. Compared with model 3, model 1 does not capture the non-monotonic relationship between CSF glucose and acute bacterial meningitis and therefore produces inaccurate predictions. Model 1 fails to capture the lowest risk for acute bacterial meningitis at a CSF glucose level of around 75 mg/dL (4.16 mmol/L) and a following increase in risk for higher values. Owing to a lack of flexibility, model 1 also fails to accurately predict the magnitude of probability in patients with low CSF glucose and

Table 4 | Specific R packages for multivariable regression modelling incorporating splines or fractional polynomials

Package	Description and reference
{rms}	Regression modelling strategies[6]
{gam}	Generalised additive models6[4]
{mgcv}	Generalised additive models[65]
{VGAM}	Vector generalised linear and additive models[66]
{mfp2}	Multivariable fractional polynomials with extensions[5]
{rstanarm}	Bayesian applied regression modelling via Stan[67]
{brms}	Bayesian regression models using Stan[68]

For Stata users, the {makespline} command (as of Stata 18) allows users to fit B splines, piecewise polynomials, and restricted cubic spline basis functions. For previous Stata versions, the {mkspline} command allows users to construct linear and restricted cubic splines. The {fp} command allows users to fit regression models with fractional polynomials, and implements the MFP approach described above. Examples with the {mkspline} and {fp} commands as well as other user-written commands for reproducing the case study can be found in the accompanying repository.

gives a maximum probability point estimate for the lowest CSF glucose values of around 65-70%. Model 2 arguably does worse, because it attributes the same probability of acute bacterial meningitis to a patient with a CSF glucose level of, for example, 10 mg/dL and 60 mg/dL. According to model 3, the more likely approximation of the true relationship between CSF glucose and acute bacterial meningitis, those two patients would have vastly different probabilities—that is, 100% for 10 mg/dL and around 20% for 60 mg/dL.

Despite pointing towards an increased odds of acute bacterial meningitis with low CSF glucose levels, both model 1 and model 2 yield highly inaccurate predictions across the range of CSF glucose and leave us misinformed because they cannot capture the underlying U shaped (ie, non-monotonic; fig 1) relationship between CSF glucose and acute bacterial meningitis. Failure to adequately approximate the true association between the continuous independent variable and the dependent variable can also result in models with poor diagnostic discrimination. A substantial difference in model discrimination (assessed using the area under the receiver operating characteristic curve (AUC)) can be found between model 1 (0.87 (95% CI 0.83 to 0.91)) and model 2 (0.87 (0.83 to 0.90)) compared with model 3 (0.96 (0.94 to 0.98)). Further performance measures such as optimism-corrected AUC, calibration, and clinical usefulness (ie, net benefit) are beyond the scope of this article, but further information can be found elsewhere.[10] Regardless, the code for conducting these analyses using the acute bacterial meningitis dataset can be found in the online repository accompanying this article (see Data and code sharing).

Software

Many excellent and validated software packages are available for commonly used statistical tools such as R[57] and Stata.[58] These packages make it relatively easy even for non-statisticians to successfully apply non-linear modelling of continuous variables.

For R users, several packages provide functions to fit a variety of splines (table 3). The {rms} package contains non-linear transformation functions for fitting linear splines (lsp) and restricted cubic splines (rcs).[59] The base R {splines} package provides functions for fitting splines either using the B spline basis (bs) or the natural cubic spline basis (ns). The {pspline} package fits penalised smoothing splines[60] and the {polspline} package fits polynomial splines.[61] Examples of fitting different splines can be found in the accompanying online repository. Perperoglou et al provide an in-depth review of multiple kinds of splines.[7] The concept of fractional polynomials is implemented in the {mfp}[62] and {mfp2}[63] packages.

Conclusions

The flawed handling of continuous variables remains prevalent in the medical literature. Assuming linearity and categorising continuous variables are statistical practices that should be avoided. Instead, non-linearity should be incorporated as part of the model building process. Established techniques, such as restricted cubic splines or fractional polynomials, are relatively easy to use and provide important flexibility to the modelling process.

We thank Ahmed Al-Hindawi (University College London and Imperial College London, UK), Ivo Strebel, Matthias Diebold, Luca Koechlin, and Jasper Boeddinghaus (all University Hospital Basel, Switzerland) for providing valuable insights and critical feedback during the drafting and revision of this article.

Contributors: TZ and PLA conceived the paper and wrote the first draft. RDR and GSC revised the manuscript and provided critical feedback. TZ is guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: No specific funding was available for this work. GSC and RDR are National Institute for Health and Care Research (NIHR) senior investigators and supported by a grant from the Engineering and Physical Sciences Research Council for "Artificial intelligence innovation to accelerate health research" (EP/Y018516/1). RDR is supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. GSC is supported by Cancer Research UK (programme grant C49297/A27294). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Data sharing: The acute bacterial meningitis data are freely available at https://hbiostat.org/data courtesy of the Vanderbilt University Department of Biostatistics. The R and Stata code accompanying this article can be found at https://gitlab.com/tobiaszimmermann/modelling-nonlinear-associations.

Dissemination to participants and related patient and public communities This article and its contents will be disseminated to the public and interested parties through presentations and lectures, social media, and word of mouth.

Provenance and peer review: Not commissioned; externally peer reviewed.

1 Shmueli G. To Explain or to Predict? *Stat Sci* 2010;25:289-310. doi:10.1214/10-STS330.

2 Riley RD, Cole TJ, Deeks J, et al. On the 12th Day of Christmas, a Statistician Sent to Me. *BMJ* 2022;379:e072883. doi:10.1136/bmj-2022-072883

3 Royston P, Altman DG. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. *Appl Stat* 1994;43:429-67. doi:10.2307/2986270.

4 Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999;28:964-74. doi:10.1093/ije/28.5.964

5 Royston P, Sauerbrei W. *Multivariable Model-Building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.* Wiley, 2008. doi:10.1002/9780470770771.

6 Harrell FE. *Regression Modelling Strategies.* 2nd ed. Springer, 2015. doi:10.1007/978-3-319-19425-7.

7 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

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

9 Spanos A, Harrell FEJr, Durack DT. Differential diagnosis of acute meningitis. An analysis of the predictive value of initial observations. *JAMA* 1989;262:2700-7. doi:10.1001/jama.1989.03430190084036

10 Collins GS, Ogundimu EO, Cook JA, Manach YL, Altman DG. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Stat Med* 2016;35:4124-35. doi:10.1002/sim.6986

11 Kahan BC, Rushton H, Morris TP, Daniel RM. A comparison of methods to adjust for continuous covariates in the analysis of randomised trials. *BMC Med Res Methodol* 2016;16:42. doi:10.1186/s12874-016-0141-3

12 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:1080.1. doi:10.1136/bmj.332.7549.1080

13 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

14 Ma J, Dhiman P, Qi C, et al. Poor handling of continuous predictors in clinical prediction models using logistic regression: a systematic review. *J Clin Epidemiol* 2023;161:140-51. doi:10.1016/j.jclinepi.2023.07.017

15 Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *AJNR Am J Neuroradiol* 2011;32:437-40. doi:10.3174/ajnr.A2425

16 Vrudhula A, Hughes JW, Yuan N, et alThe Impact of Task Set-up in Algorithm Design: Regression versus Classification. *NEJM AI* 2024;1:AIcs2300176. doi:10.1056/AIcs2300176.

17 Wainer H, Gessaroli M, Verdi M. Visual Revelations: Finding What Is Not There through the Unfortunate Binning of Results: The Mendel Effect. *Chance* 2006;19:49-52. doi:10.1080/09332480.2006.10722771.

18 Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;86:829-35. doi:10.1093/jnci/86.11.829

19 Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KG, PROTECT WP2 (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium, Work Programme 2 [Framework for pharmacoepidemiology studies]). Adjustment for continuous confounders: an example of how to prevent residual confounding. *CMAJ* 2013;185:401-6. doi:10.1503/cmaj.120592

20 Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 1997;8:429-34. doi:10.1097/00001648-199707000-00014

21 Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441. doi:10.1136/bmj.m441

22 Sauerbrei W, Perperoglou A, Schmid M, et al, for TG2 of the STRATOS initiative. State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. *Diagn Progn Res* 2020;4:3. doi:10.1186/s41512-020-00074-3

23 Bhaskaran K, Dos-Santos-Silva I, Leon DA, Douglas IJ, Smeeth L. Association of BMI with overall and cause-specific mortality: a population-based cohort study of 3·6 million adults in the UK. *Lancet Diabetes Endocrinol* 2018;6:944-53. doi:10.1016/S2213-8587(18)30288-2

24 Li H, Qian F, Zuo Y, et al. U-Shaped Relationship of High-Density Lipoprotein Cholesterol and Incidence of Total, Ischemic and Hemorrhagic Stroke: A Prospective Cohort Study. *Stroke* 2022;53:1624-32. doi:10.1161/STROKEAHA.121.034393

25 Collins GS, Mallett S, Omar O, Yu LM. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med* 2011;9:103. doi:10.1186/1741-7015-9-103

26 Collins GS, Omar O, Shanyinde M, Yu LM. A systematic review finds prediction models for chronic kidney disease were poorly reported and often developed using inappropriate methods. *J Clin Epidemiol* 2013;66:268-77. doi:10.1016/j.jclinepi.2012.06.020

27 Dhiman P, Ma J, Gibbs VN, et al. Systematic review highlights high risk of bias of clinical prediction models for blood transfusion in patients undergoing elective surgery. *J Clin Epidemiol* 2023;159:10-30. doi:10.1016/j.jclinepi.2023.05.002

28 Dhiman P, Ma J, Andaur Navarro CL, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. *BMC Med Res Methodol* 2022;22:101. doi:10.1186/s12874-022-01577-x

29 Andaur Navarro CL, Damen JAA, van Smeden M, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. *J Clin Epidemiol* 2023;154:8-22. doi:10.1016/j.jclinepi.2022.11.015

30 Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. *Br J Cancer* 1994;69:979-85. doi:10.1038/bjc.1994.192

31 Riley RD, Hayden JA, Steyerberg EW, et al, PROGRESS Group. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. *PLoS Med* 2013;10:e1001380. doi:10.1371/journal.pmed.1001380

32 Hilsenbeck SG, Clark GM, McGuire WL. Why do so many prognostic factors fail to pan out?*Breast Cancer Res Treat* 1992;22:197-206. doi:10.1007/BF01840833

33 Malats N, Bustos A, Nascimento CM, et al. P53 as a prognostic marker for bladder cancer: a meta-analysis and review. *Lancet Oncol* 2005;6:678-86. doi:10.1016/S1470-2045(05)70315-6

34 Nicholson A, Kuper H, Hemingway H. Depression as an aetiologic and prognostic factor in coronary heart disease: a meta-analysis of 6362 events among 146 538 participants in 54 observational studies. *Eur Heart J* 2006;27:2763-74. doi:10.1093/eurheartj/ehl338

35 Whiteley W, Chong WL, Sengupta A, Sandercock P. Blood markers for the prognosis of ischemic stroke: a systematic review. *Stroke* 2009;40:e380-9. doi:10.1161/STROKEAHA.108.528752

36 Groenwold RHH, Van Deursen AMM, Hoes AW, Hak E. Poor quality of reporting confounding bias in observational intervention studies: a systematic review. *Ann Epidemiol* 2008;18:746-51. doi:10.1016/j.annepidem.2008.05.007

37 Müllner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Ann Intern Med* 2002;136:122-6. doi:10.7326/0003-4819-136-2-200201150-00009

38 Kahan BC, Jairath V, Doré CJ, Morris TP. The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. *Trials* 2014;15:139. doi:10.1186/1745-6215-15-139

39 Sullivan TR, Morris TP, Kahan BC, Cuthbert AR, Yelland LN. Categorisation of continuous covariates for stratified randomisation: How should we adjust?*Stat Med* 2024;43:2083-95. doi:10.1002/sim.10060

40 Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med* 2004;23:2509-25. doi:10.1002/sim.1815

41 Williamson SF, Grayling MJ, Mander AP, et al. Subgroup analyses in randomized controlled trials frequently categorized continuous subgroup information. *J Clin Epidemiol* 2022;150:72-9. doi:10.1016/j.jclinepi.2022.06.017

42 Brankovic M, Kardys I, Steyerberg EW, et al. Understanding of interaction (subgroup) analysis in clinical trials. *Eur J Clin Invest* 2019;49:e13145. doi:10.1111/eci.13145

43 Vaduganathan M, Claggett BL, Inciardi RM, Fonarow GC, McMurray JJV, Solomon SD. Estimating the Benefits of Combination Medical Therapy in Heart Failure With Mildly Reduced and Preserved Ejection Fraction. *Circulation* 2022;145:1741-3. doi:10.1161/CIRCULATIONAHA.121.058929

44 Riley RD, Debray TPA, Fisher D, et al. Individual participant data meta-analysis to examine interactions between treatment effect and participant-level covariates: Statistical recommendations for conduct and planning. *Stat Med* 2020;39:2115-37. doi:10.1002/sim.8516

45 Grambsch PM, O'Brien PC. The effects of transformations and preliminary tests for non-linearity in regression. *Stat Med* 1991;10:697-709. doi:10.1002/sim.4780100504

46 Collins GS, Dhiman P, Ma J, et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* 2024;384:e074819. doi:10.1136/bmj-2023-074819

47 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

48 Royston P, Sauerbrei W. Multivariable Modeling with Cubic Regression Splines: A Principled Approach. *Stata J* 2007;7:45-70. doi:10.1177/1536867X0700700103.

49 Desquilbet L, Mariotti F. Dose-response analyses using restricted cubic spline functions in public health research. *Stat Med* 2010;29:1037-57. doi:10.1002/sim.3841

50 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

51 Zimmermann T, Lopez-Ayala P, Singer M. Serial assessments of cardiac output and mixed venous oxygen saturation in comatose patients after out-of-hospital cardiac arrest. *Crit Care* 2023;27:451. doi:10.1186/s13054-023-04734-w

52 O'Donnell M, Mente A, Rangarajan S, et al, PURE Investigators. Urinary sodium and potassium excretion, mortality, and cardiovascular events. *N Engl J Med* 2014;371:612-23. doi:10.1056/NEJMoa1311889

53 Crane PK, Walker R, Hubbard RA, et al. Glucose levels and risk of dementia. *N Engl J Med* 2013;369:540-8. doi:10.1056/NEJMoa1215740

54 Madsen CM, Varbo A, Tybjærg-Hansen A, Frikke-Schmidt R, Nordestgaard BG. U-shaped relationship of HDL and risk of infectious disease: two prospective population-based cohort studies. *Eur Heart J* 2018;39:1181-90. doi:10.1093/eurheartj/ehx665

55 Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. *Biom J* 2018;60:431-49. doi:10.1002/bimj.201700067

56 Chatfield C. Model uncertainty, data mining and statistical inference. *J R Stat Soc Ser A Stat Soc* 1995;158:419-66. doi:10.2307/2983440.

57 R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, 2024.

58 StataCorp. Stata Statistical Software. 2023.

59 Harrell Jr FE. *Rms: Regression modeling strategies*. 2024.

60 Ripley B. *Pspline: Penalized smoothing splines*. 2022.

61 Kooperberg C. *Polspline: Polynomial spline routines*. 2023.

62 Ambler G, Benner A. *Mfp: Multivariable fractional polynomials*. 2023.

63 Kipruto E, Kammer M, Royston P, *et al. Mfp2: Multivariable fractional polynomial models with extensions*. 2023.

64 Hastie T, Tibshirani R. *Generalized additive models*. Routledge, 2017.

65 Wood SN. *Generalized additive models: An introduction with R.* 2nd ed. CRC Press/Taylor & Francis Group, 2017. doi:10.1201/9781315370279.

66 Yee TW. *Vector generalized linear and additive models: With an implementation in R.* New York, Heidelberg, Dordrecht, London: Springer, 2016.

67 Goodrich B, Gabry J, Ali I, *et al*. Rstanarm: Bayesian applied regression modeling via Stan. 2020.

68 Bürkner P-C. Brms: An *R* Package for Bayesian Multilevel Models Using *Stan*. *J Stat Softw* 2017;80:1-28. doi:10.18637/jss.v080.i01.

Web appendix: Supplemental material