Yang *et al.*
BMC Medical Research Methodology (2025) 25:70
https://doi.org/10.1186/s12874-025-02522-4

BMC Medical Research Methodology

RESEARCH

Open Access

Discretizing multiple continuous predictors with U-shaped relationships with ln*OR*: introducing the recursive gradient scanning method in clinical and epidemiological research

Shuo Yang[1], Huaan Su[1,2], Nanxiang Zhang[1], Yuduan Han[1], Yingfeng Ge[1], Yi Fei[1], Ying Liu[1], Abdullahi Hilowle[3], Peng Xu[4] and Jinxin Zhang[1*]

Abstract

Background Assuming a linear relationship between continuous predictors and outcomes in clinical prediction models is often inappropriate, as true linear relationships are rare, potentially resulting in biased estimates and inaccurate conclusions. Our research group addressed a single U-shaped independent variable before. Multiple U-shaped predictors can improve predictive accuracy by capturing nuanced relationships, but they also introduce challenges like increased complexity and potential overfitting. This study aims to extend the applicability of our previous research results to more common scenarios, thereby facilitating more comprehensive and practical investigations.

Methods In this study, we proposed a novel approach called the Recursive Gradient Scanning Method (RGS) for discretizing multiple continuous variables that exhibit U-shaped relationships with the natural logarithm of the odds ratio (ln*OR*). The RGS method involves a two-step approach: first, it conducts fine screening from the 2.5th to 97.5th percentiles of the ln*OR*. Then, it utilizes an iterative process that compares *AIC* metrics to identify optimal categorical variables. We conducted a Monte Carlo simulation study to investigate the performance of the RGS method. Different correlation levels, sample sizes, missing rates, and symmetry levels of U-shaped relationships were considered in the simulation process. To compare the RGS method with other common approaches (such as median, Q_1-Q_3, minimum *P*-value method), we assessed both the predictive ability (e.g., *AUC*) and goodness of fit (e.g., *AIC*) of logistic regression models with variables discretized at different cut-points using a real dataset.

Results Both simulation and empirical studies have consistently demonstrated the effectiveness of the RGS method. In simulation studies, the RGS method showed superior performance compared to other common discretization methods in discrimination ability and overall performance for logistic regression models across various U-shaped scenarios (with varying correlation levels, sample sizes, missing rates, and symmetry levels of U-shaped relationships). Similarly, empirical study showed that the optimal cut-points identified by RGS have superior clinical predictive power, as measured by metrics such as *AUC*, compared to other traditional methods.

Conclusions The simulation and empirical study demonstrated that the RGS method outperformed other common discretization methods in terms of goodness of fit and predictive ability. However, in the future, we will focus

*Correspondence:
Jinxin Zhang
zhjinx@mail.sysu.edu.cn
Full list of author information is available at the end of the article

on addressing challenges related to separation or missing binary responses, and we will require more data to validate our method.

Keywords The Recursive Gradient Scanning Method, Optimal cut-points, Discretization, U-shaped relationships, Ln*OR* (logarithm of Odds Ratio), Regression modelling

Background

In clinical and epidemiological research, elucidating the association between continuous exposures or covariates and disease outcomes is paramount for risk prediction and treatment stratagem [1–3]. In recent years, there has been increasing recognition of U-shaped dose–response relationships in medical research. Several physiological indicators, such as body weight and frailty [4], body weight, waist-to-height ratio, and depression [5], protein intake and muscle gain [6], blood uric acid content and risk of death [7], and uric acid and risk of chronic kidney disease [7, 8], all demonstrate a U-shaped pattern. These findings suggest that both low and high levels of these factors may be associated with negative health outcomes, while moderate levels are associated with better health. However, there is often a lack of consideration for the discretization method in the study of these relationships. This gap in practice may lead to oversimplification of the complex relationships between continuous exposures or covariates and disease outcomes, potentially affecting the accuracy and reliability of the findings.

Dealing with continuous predictors in model development poses significant challenges [9]. Researchers often simplify the interpretation by categorizing these predictors using reference ranges or based on data distribution. However, this strategy has been criticized for potentially introducing bias and losing valuable information present in the continuous nature of the variables [10–12]. Moreover, due to variations in positive outcomes, optimal cut-points determined by data also vary. There is consensus among statisticians that treating variables as continuous, using non-parametric or spline regressions is preferred when the latent trend is complex. Failing to model the functional form appropriately (i.e., the pattern of the relationship between a continuous predictor and the outcome) can lead to a substantial loss of statistical power to find and model the true underlying relationship. In turn, this may produce a prediction model with worse predictive performance and wrong predictions on clinical decisions, which can adversely influence patient care [13]. However, these methods are more complex either to apply or to interpret, which has led to the continued use of categorized data in epidemiological studies. How continuous predictors are discretized during model development will influence disease predictions for an individual, thus potentially impacting subsequent treatment strategies and further prognosis. Therefore, researchers should deliberately consider how continuous predictors are examined and discretized to ensure the development of a robust model that provides accurate predictions.

Potential approaches for discretizing continuous predictors are to (1) categorize them into two or more groups by arithmetic mean, median, or quartiles; and (2) select cut-points for discretizing continuous predictors based on the statistical significance of the predictor's coefficients in the model. Categorizing continuous predictors into groups by arithmetic mean, median, or quartiles, a widely discredited practice, may lead to weaker performance compared to a model where the functional form is appropriately modeled [13]. In the latter approach, each unique value of the predictor is considered a candidate boundary point, and the best boundary point is selected based on criteria such as the minimum *P*-value or the maximum Youden's index. These classification methods may result in individuals with similar risk levels being grouped differently, potentially leading to misclassification of high and low-risk individuals and reducing statistical significance of corresponding explanatory variables in the prediction model. These traditional discretization methods are based on a logistic regression framework that assumes a linear relationship between predictors and the ln*OR*. Consequently, they are unable to capture U-shaped relationships and do not yield appropriate cut-points. This limitation negatively impacts the accuracy of the models and leads to erroneous estimates.

Our research group has proposed the "two cut-points with maximum *OR* value method" [14] and "two cut-points with minimum *AIC* value method" [15], which offer better goodness of fit and interpretation compared to commonly used methods such as the direct method with continuous form, median classification method, and minimum *P*-value method. Our preliminary work in this area has drawn wide attention by the international biostatistical community [16–21], contributing to addressing challenges in the application of multivariate regression analysis and making different research results more comparable.

In medical research, multivariate regression models may involve a number of non-monotonic independent variables [22–24]. Discretizing these variables requires advanced statistical theory and algorithmic

implementation [25, 26]. Currently, research on discretizing multiple continuous independent variables, particularly when the outcome exhibits U-shaped relationships with one more independent variable simultaneously, are limited [27–29]. To address these limitations, we propose the Recursive Gradient Scanning (RGS) method. This approach uses a gradient scanning [30] principle for fine-tuning, enabling the generation of a discretization scheme with professional logic and interpretability. During the process of identifying optimal cut-points, it is crucial to consider the effects of ln*OR*. By iteratively sorting the ln*OR*, high-risk groups are defined based on consistent thresholds of log relative hazards, enhancing the clinical rationale for classification. Our method offers more accurate optimal cut-points and prevents individuals with similar risks from being placed in opposing risk groups. We compare the performance of the RGS method with other common discretization methods through simulation studies and demonstrate its application in a real dataset on depression.

This paper is structured as follows: Sect. "Methods" outlines the RGS method, Sect. "Results" compares its performance with other methods in simulation studies, Sect. "Discussion" presents the results from the application of the RGS method to a real dataset, and Sect. "Conclusions" provides discussion and conclusions.

Methods

Identifying optimal cut-points based on log(η) values and Modeling

During the process of determining optimal cut-points, it's crucial to consider the effects of ln*OR* or ln*HR* (natural logarithm of the odds ratio or hazard ratio). Therefore, we use log(η) (representing ln*OR* or ln*HR*) to identify two optimal cut-points (P_1, P_2) for the predictive variable X. Patients with X smaller than P_1 or larger than P_2 are classified into the high-risk group (or the low-risk group in an inverse U-shaped relationship) respectively. This approach ensures that the two high-risk groups are defined according to the same threshold of log(η), enhancing the clinical relevance of the classification. Two R packages, "CutpointsOEHR" and "TCPMOR", have been developed to assist investigators in easily implementing the optimal equal-*HR* (and *OR*) method.

Existing nonlinear methods can identify the shape of relationships between independent variables and ln*OR* but do not support discretization. In contrast, our study employs spline techniques to detect U-shaped relationships and then utilizes scanning methods for discretization, enabling us to categorize the population into low and high-exposure groups. Our method enhances the precision of determining optimal cut-points, ensuring that individuals with similar risks are not allocated to opposite risk groups. This process involves seeking two cut-points with nearly identical log(η) values, achieved through a series of algorithms or iterative methods. These aim to pinpoint two cut-points with almost same log(η) values, by which the individuals can be allocated into significantly distinguished subgroups. The process of identifying these optimal cut-points within the framework of the equal-*OR* (or *HR*) method consists of several steps. Integrating specific formulas into this framework can provide additional clarity. In addition, we applied gradient scanning for fine screening from the 2.5th to 97.5th percentiles of the ln*OR*, in order not to skip from the optimal cut-points. Below are the steps, along with integrated formulas, for identifying optimal cut-points based on log(η) values.

Data preparation

Data collection and preprocessing involve gathering the dataset relevant to the study and ensuring it contains all necessary variables and outcomes for analysis.

Calculation of log(η) for each potential cut-points

Compute log(η) values using a specific model or equation that relates to the context of the study, such as a case–control study or a cohort study. For instance, log(η) could be represented as follows:

$$\log(\eta) = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_\rho X_\rho \quad (1)$$

where log(η) is a function or model based on the predictor variables. $x_1, x_2 \cdots x_p$ denotes the predictor variables or features in the model.

Semiparametric regression models are flexible statistical techniques that allow the observed data to determine the functional form of the relationship between predictors and responses. They can summarize complex data structures such as nonlinearity within the linear model framework, making them particularly useful for graphical representation to assess the shape of continuous risk factor-response curves.

Splines are commonly used smoothing functions for quantitative risk assessments because they enable the estimation of risk for specific levels of a continuous predictor. Semiparametric regression models with smoothing splines use knots located at each unique value of the continuous predictor variable and incorporate a penalty term to avoid overfitting. They calculate a weighted average of the outcome variable across local regions of the continuous predictor, giving higher weight to center data points than outermost data points. This procedure plots a smoothed curve with minimal constraints on its shape.

Identification of optimal cut-points

Utilize the log(η) values to identify a pair of cut-points within the dataset. Define criteria or algorithms aiming to find these points where log(η) values demonstrate specific properties, such as approximate equality or meeting certain pre-conditions. One possible formula within this step might involve a condition for approximate equality.

$$Abs\left(\log\left(\eta_{\text{cut}-\text{points1}}\right) - \log\left(\eta_{\text{cut}-\text{points2}}\right)\right) \approx 0 \quad (2)$$

where Abs() denotes the absolute value function.$\log\left(\eta_{\text{cut}-\text{points1}}\right)$ and $\log\left(\eta_{\text{cut}-\text{points2}}\right)$ represent log(η) values at the final determined cut-points.

Graphical diagnostic plot

The semiparametric models with penalized B-splines (P-splines) were fitted using the R package "SemiPar" [31]. This approach balances the goodness of fit and variance to curve the relationship and assess the statistical significance of the non-linear term.

Select two optimal cut-points for each continuous explanatory variable as original cut-points

Deviation from linearity in these models can be assessed using the estimated degrees of freedom and corresponding *P*-value. Degrees of freedom (*df*) represent the total influence of all observations on the estimates of statistical parameters. In the semiparametric regression setting, estimated *df* for the nonparametric component provides evidence for a departure from linearity. Both the visual representation of the curve indicates a U-shaped relationship and $df > 2$ (to ensure nonlinearity), then the "two cut-points with maximum *OR* value method" was used to identify the original upper and lower cut-points of the continuous explanatory variable at which the *OR* reaches its maximum.

The illustration for discretizing one non-monotonic independent variable

In our study, the problem of discretization can be defined as follows. Assuming a dataset *S* consisting of *N* samples, *M* attributes, and *c* class labels, a discretization scheme D_A would exist on the continuous attribute $A \in M$, which partitions this attribute into *n* paired discrete cut-points with equal log(η):[(d_{1L}, d_{1R}), (d_{2L}, d_{2R}), ...(d_{nL}, d_{nR})], where d_1 and d_n are, respectively, the $P_{97.5}$ and $P_{2.5}$ value of log(η), and $P_A = \{ d_1, d_2, \ldots\ldots d_n\}$ represents the set of candidate gradient of A in descending order. The criterion for optimizing the discretization of log(η) for multiple variables typically involves minimizing *AIC* to achieve the best discretization effect.

We can associate a typical discretization with a univariate discretization. Although this property will be reviewed in the next section, it is necessary to introduce it here for the basic understanding of the basic discretization process. In the context of univariate scenarios, achieving this is already quite challenging, while multivariate discretization, which considers multiple features simultaneously, presents an even greater complexity. A typical discretization process generally consists of four steps (shown in Fig. 1): ① sorting observations by the values of the feature (i.e. log(η)) to be discretized, ② scanning candidate cut-points with equal log(η) iteratively, ③ dichotomizing observations into low-risk and high-risk groups by each candidate cut-points, and ④ evaluating model performance.

The RGS method for the discretization of multiple non-monotonic independent variables

The methodology for discretizing multiple non-monotonic independent variables is detailed as follows:

① For variables exhibiting U-shaped associations with ln*OR*, we used the "two cut-points with maximum *OR* value method" to search a pair of original cut-points for scanning.

② We determined the percentile rankings (Q_k, $k = 1, 2, \ldots, 100$) of the estimated ln*OR* values for each variable. Horizontal lines (i.e. gradients) were drawn at each percentile between the 2.5th and 97.5th percentiles, intersecting the U-shaped curve at two cut-points. The interval of the ln*OR* values is calculated as:

$$\Delta lnOR = lnOR_{P97.5} - lnOR_{P2.5} \quad (3)$$

③ Using the R-function spline interpolation technique, new cut-points were generated as candidate cut-points, ensuring a smooth curve with equal ln*OR* values in lower and higher cut-points ($|lnOR_{nL} - lnOR_{nR}| \leq \Delta lnOR \times 1/100$).

④ The RGS method refined boundary points by iteratively scanning from the original cut-points, moving vertically in each gradient step ($\Delta lnOR \times 1/100$). The scan continued upward or downward until model fit no longer improved, halting at $P_{97.5}$ or $P_{2.5}$ of ln*OR* respectively. With *k* variables, 2^k scanning methods were employed (e.g., for $k = 2$: upward x_1 combined with downward x_2, downward x_1 combined with upward x_2, both upward, both downward).

⑤ Optimal cut-points were selected based on model fitness measurements (*AIC*, Nagelkerke R^2, and −2Log-likelihood), with the final cut-points used in regression models.

The assumptions in the RGS approach

First, the visual representation of the curve indicates a U-shaped relationship and $df > 2$ to ensure nonlinearity. Second, it is essential to have sufficient values at both

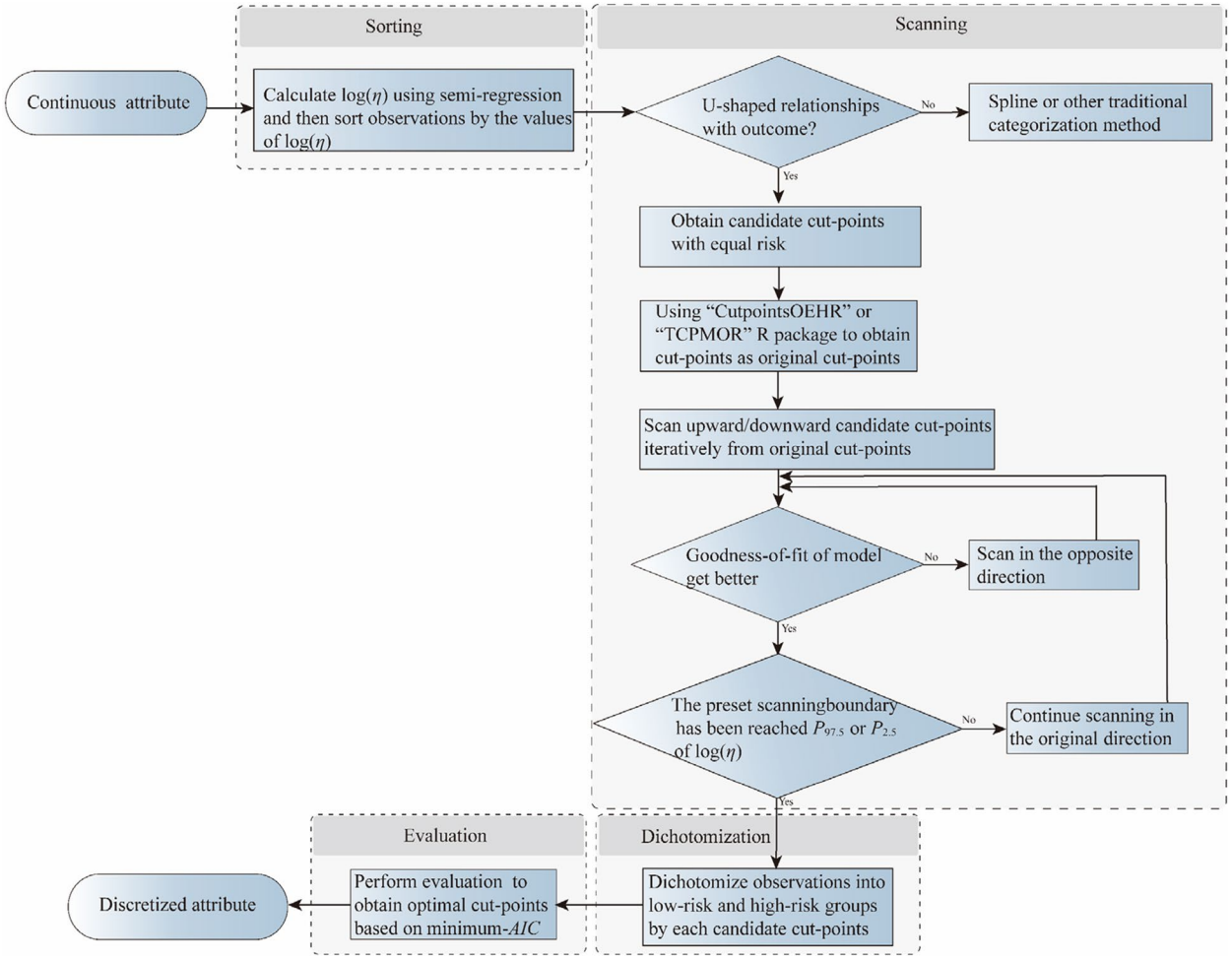

Fig. 1 Discretization process for the RGS

ends of the discretized variables or enough positive outcome events.

Be cautious that if the $\Delta \ln OR$ is too large, we should narrow the scanning strategy to the 0.1th percentiles of the $\ln OR$ to ensure the iterative process runs smoothly and to avoid missing the optimal cut-points.

Validation and assessment

Validate the identified cut-points using statistical measures. Evaluate the reliability and effectiveness of these discretization methods within the context of the study.

Measures of goodness-of-fit

The goodness of fit of models containing multiple predictors for binary responses was assessed using several metrics. Coefficients of determination (R^2) were calculated to evaluate the proportion of variance explained by the models. The Akaike Information Criterion (*AIC*) was used to compare the relative quality of different models, with lower *AIC* values indicating better fit. Additionally, the −2loglikelihood was computed as a measure of how well the model predicts the observed data, with lower values indicating better fit. These metrics were particularly useful for assessing the performance of the RGS method compared to other discretization methods in fitting models.

Measures of predictive ability

To validate a prediction model systematically, the predictive performance of the model is commonly addressed by discrimination (the ability of the model to distinguish between positive and negative outcomes) and calibration (the agreement between the observed outcomes and predicted probabilities).

In this study, we used several statistical techniques to assess the predictive performance of different classification models. *AUC* (the area under the receiver operating characteristic curve) was calculated to summarize the

discriminating power of each gradient (corresponding to candidate cut-points with same $\log(\eta)$). A higher *AUC* value indicates better predictive capability.

Decision Curve Analysis [32] (DCA) is usually applied in clinical effectiveness evaluation. It compares the net benefit of model intervention with the net benefit of default methods (full intervention and no intervention). Typically, a higher net benefit value indicates better performance of the model at a specified probability threshold.

Design of the simulation study

In this study, we recommend using the RGS method to discretize data in logistic regression models when there is a non-monotonic relationship between logit π and continuous predictors. Monte Carlo simulations were conducted to compare the performance of the RGS method with four traditional discretization methods (median, Q_1-Q_3, minimum *p*-value, and two cut-points with maximum *OR* value method) in terms of model fitting and prediction ability.

The simulation study included various scenarios, such as different levels of correlations, sample sizes, missing rates, and the symmetry of U-shaped curves. These scenarios were made to mirror real-world scenarios, ensuring that the findings are applicable and reliable. Furthermore, our previous research has demonstrated that these parameters significantly impact model performance [15]. For the simulations, two non-monotonic U-shaped variables, X_1 and X_2, were generated. Additionally, we included two covariates, a continuous variable, X_3, which is linearly related to $\ln OR$, and a binary variable, X_4. Specifically, X_1 has a mean of 0.0 and a standard deviation of 1.0 in the case group, while in the control group, it has a mean of 0.0 and a standard deviation of 0.5. For X_2, the case group exhibits a mean of 0.0 with a standard deviation of 0.8, compared to the control group, which has a mean of 0.0 and a standard deviation of 0.3. X_3 has a standard deviation of 1.0, with a mean of 1.0 for a "yes" outcome and 0.0 for a "no" outcome.

Correlation levels

We generated a correlation matrix with specified coefficients of 0.3 (low correlation), 0.6 (moderate correlation), and 0.9 (high correlation) for X_1, X_2, and X_3 to effectively explore the correlation relationships among these continuous predictors. Subsequently, we assessed the performance of the RGS method under these conditions.

Sample sizes

The sample size scenarios considered were 100, 200, 300, 400, 500, and 1000 observations. These scenarios were chosen to evaluate the performance of the discretization methods across a range of sample sizes.

Missing rates

The missing data scenarios were based on a missing completely at random (MCAR) mechanism and a missing completely at random (MAR) mechanism. MCAR means that the missingness of data was completely unrelated to any other variables or the data itself, occurring purely at random. In contrast, MAR indicates that the probability of missingness depends on the observed data but not on the missing values. Under MAR, the likelihood of a value being missing can be estimated based on the available non-missing data. The missing rate scenarios considered were 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50%. These scenarios were chosen to evaluate the performance of the discretization methods under varying rates of missing data. To avoid a reduction in sample size, we employed multiple imputation techniques to handle missing data. Specifically, we used the Predictive Mean Matching (PMM) and Random Forest (RF) methods for imputation [33]. Multiple imputation was conducted to ensure that missing data were adequately addressed.

Symmetry of u-shaped curves

Three scenarios were considered for the symmetry of U-shaped curves, symmetric (the curve is perfectly symmetric), partially symmetric (the curve is somewhat symmetric) and severely asymmetric (the curve shows significant asymmetry, deviating greatly from a symmetric shape). These scenarios were chosen to capture different degrees of symmetry in the U-shaped curves and to assess how the discretization methods perform under varying levels of symmetry.

Implementation in R

The statistical analyses were conducted using several R packages. The "maxstat" package was utilized to implement the minimum *P*-value method with log-rank statistics. For fitting logistic regression models with splines, the "SemiPar" package was employed. Additionally, the "CutpointsOEHR" and "TCPMOR" packages were used to obtain the optimal equal-*HR* (or *OR*) in univariate discretization. All statistical tests were performed at a two-sided significance level of 0.05, with *P*-values below this threshold considered statistically significant. The analyses were conducted using R version 4.2.2 (R Foundation for Statistical Computing, http://www.R-project.org).

Application to the behavioral risk factor surveillance system (BRFSS) data

We analyzed data from the Behavioral Risk Factor Surveillance System (BRFSS) [34] survey conducted by the Centers for Disease Control and Prevention through a telephone survey of individuals aged 18 years or older

across all 50 states in the USA between 2011 and 2019. The study was exempt from Institutional Review Board approval since it utilized de-identified data from a publicly available dataset in the public domain. In this state-based, random-digit-dialed telephone survey collecting self-reported health information, a total of 871,919 adults with complete data were included in the analysis. Previous studies have identified U-shaped patterns among monthly exercise frequency, weekly exercise sessions, and depression [35]. To identify the optimal cut-points for discretizing monthly exercise frequency and weekly exercise sessions to maximize improvements in mental health, we applied the RGS method.

Results

Results of the simulation study

Scenario1: Correlation levels

The *AIC* and *AUC* values for the correlation scenarios with coefficients of 0.3, 0.6, and 0.9 are presented in Tables 1, 2. These scenarios investigated the effects of varying correlation levels on model performance. It was observed that as the correlation increased, *AIC* values rose while *AUC* values declined. Notably, the RGS method exhibited superior performance across all correlation levels.

Scenario2: Sample size

In Tables 3, 4, the *AIC* and *AUC* values of the RGS method were compared to four traditional discretization methods (median, Q_1-Q_3, minimum *P*-value, and two cut-points with maximum *OR* value method) across various sample sizes. The results showed that as the sample size increased, the *AIC* and *AUC* increased for all discretization methods. Notably, the RGS method exhibited not only lower *AIC* but also higher *AUC* values compared to traditional discretization methods, indicating superior model fit, parsimony, and predictivity.

Table 1 Comparison of Discretization Methods in Terms of *AIC* under Various Correlation Levels

Discretization methods	Low correlation	Moderate correlation	High correlation
Min-*P*	1159.97$^{(5)}$	2392.18$^{(5)}$	3120.58$^{(5)}$
Median	1147.76$^{(4)}$	2357.96$^{(4)}$	3090.42$^{(4)}$
Q_1-Q_3	1078.21$^{(3)}$	2237.64$^{(3)}$	3021.35$^{(3)}$
Two cut-points with maximum *OR*	1051.36$^{(2)}$	2112.35$^{(2)}$	2935.22$^{(2)}$
RGS	963.95$^{(1)}$	1972.57$^{(1)}$	2832.17$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AIC* values of the model, with $^{(1)}$ representing the best

Table 2 Comparison of Discretization Methods in Terms of *AUC* under Various Correlation Levels

Discretization methods	Low correlation	Moderate correlation	High correlation
Min-*P*	0.721$^{(5)}$	0.652$^{(5)}$	0.598$^{(5)}$
Median	0.732$^{(4)}$	0.659$^{(4)}$	0.612$^{(4)}$
Q_1-Q_3	0.737$^{(3)}$	0.668$^{(3)}$	0.638$^{(3)}$
Two cut-points with maximum *OR*	0.747$^{(2)}$	0.672$^{(2)}$	0.642$^{(2)}$
RGS	0.752$^{(1)}$	0.674$^{(1)}$	0.654$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AUC* values of the model, with $^{(1)}$ representing the best

Scenario3: missing rates

As shown in Table 5, increasing missing rates have led to higher *AIC* values of the discretization methods after imputation using the PMM method, indicating poorer fit. However, regardless of the missing rate scenarios, the RGS method consistently exhibited excellent performance, as indicated by its consistently the smallest *AIC* value, suggesting it as the optimal discretization approach.

To further compare different missing value imputation methods, this study also utilized RF multiple imputation method. The results, as shown in Table 6, indicated that the model *AIC* after RF imputation was smaller than that after PMM imputation. This finding suggested that the RF imputation method offered a better model fitting effect compared to PMM.

In terms of model prediction ability, as shown in Table 7, the *AUC* of the RGS method proposed in this study achieved consistently the highest *AUC* values in each scenario of missing rate parameters, indicating its superior predictive ability.

For better visualization, we conducted a decision curve analysis at a 5% missing rate. The results, depicted in Fig. 2, revealed the predictive performance rankings of the discretization methods as follows: RGS method, Min-*P*, Q_1-Q_3, and Median method. These rankings were consistent with the results in Table 8, where the RGS method consistently outperformed other traditional discretization methods.

The results for the MAR mechanism were presented in the supplementary files (Tables S1-S3) and showed more favorable findings after PMM and RF imputation. Additionally, we provided the results for the MCAR mechanism in a complete case analysis (Tables S4-S5), which are similar to those of the imputation data.

Table 3 Comparison of Discretization Methods in Terms of *AIC* under Various Sample Size Scenarios

Discretization methods	Sample size = 100	Sample size = 200	Sample size = 300	Sample size = 400	Sample size = 500	Sample size = 1000
Min-*P*	557.12$^{(5)}$	839.16$^{(5)}$	1115.55$^{(5)}$	1329.40$^{(3)}$	2787.46$^{(5)}$	3557.27$^{(4)}$
Median	554.85$^{(4)}$	835.74$^{(4)}$	1111.00$^{(4)}$	1385.80$^{(4)}$	2776.10$^{(4)}$	3562.69$^{(5)}$
Q_1-Q_3	532.27$^{(3)}$	801.73$^{(3)}$	1065.79$^{(3)}$	1391.47$^{(5)}$	2663.12$^{(3)}$	3417.70$^{(3)}$
Two cut-points with maximum *OR*	529.77$^{(2)}$	797.96$^{(2)}$	1060.78$^{(2)}$	1323.16$^{(2)}$	2650.62$^{(2)}$	3401.65$^{(2)}$
RGS	513.58$^{(1)}$	722.95$^{(1)}$	961.07$^{(1)}$	1198.78$^{(1)}$	2401.45$^{(1)}$	3281.89$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AIC* values of the model, with$^{(1)}$ representing the best

Table 4 Comparison of Discretization Methods in Terms of *AUC* under Various Sample Size Scenarios

Discretization methods	Sample size = 100	Sample size = 200	Sample size = 300	Sample size = 400	Sample size = 500	Sample size = 1000
Median	0.559$^{(5)}$	0.651$^{(5)}$	0.657$^{(5)}$	0.688$^{(5)}$	0.692$^{(5)}$	0.716$^{(5)}$
Min-*P*	0.644$^{(4)}$	0.652$^{(4)}$	0.661$^{(4)}$	0.692$^{(4)}$	0.695$^{(4)}$	0.732$^{(4)}$
Q_1-Q_3	0.648$^{(3)}$	0.656$^{(3)}$	0.664$^{(3)}$	0.693$^{(3)}$	0.712$^{(3)}$	0.739$^{(3)}$
Two cut-points with maximum *OR*	0.652$^{(2)}$	0.667$^{(2)}$	0.670$^{(2)}$	0.697$^{(2)}$	0.717$^{(2)}$	0.746$^{(2)}$
RGS	0.658$^{(1)}$	0.668$^{(1)}$	0.684$^{(1)}$	0.708$^{(1)}$	0.754$^{(1)}$	0.766$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AUC* values of the model, with $^{(1)}$ representing the best

Scenario4: symmetric extents

The *AIC* and *AUC* values for symmetry, partial symmetry, and completely asymmetric scenarios were presented in Tables 9, 10. These scenarios explored the effects of symmetric extents on model performance. It was found that as the symmetry decreased, *AIC* values increased, whereas *AUC* values decreased. The RGS method demonstrated superior performance across all different levels of symmetry.

Results of the application on a real dataset

Figure 3 showed the non-linear associations of metabolic equivalents (METs) and exercise frequency with depression using penalized cubic regression splines. Table 11 illustrated the performance of different estimated cut-points in logistic regression models. ① The original cut-points for physical activity frequency were 16.5–23.7 times/month, while those for MET were 751.3–1847.6 METs-min/week. ② The model using the optimal cut-points showed better performance in discrimination ability and overall performance than the model using the original cut-points. Physical activity frequency between about 12.6–26.5 times per month was associated with a lower mental health burden. For METs-min/week, better mental health was observed when the METs-min/week was about 610.9–2203.7.

The cut-points estimated by the RGS method demonstrated superior performance in discrimination ability and overall performance (explained variance) in terms of R^2, *AIC*, and *AUC* measures (as shown in Table 12).

Discussion

The relationship between the predictors and outcomes may be monotonic or non-monotonic, with U-shape patterns being common in non-monotonic dose–response relationships in medical research [36–39]. In our study, we explored how to discretize multiple continuous predictors using a pair of cut-points for U-shaped relationships. When identifying these optimal cut-points, it is crucial to first confirm whether the validity of the U-shaped association [40, 41]. Our research demonstrated that the RGS method outperforms other discretization approaches in logistic regression modelling, particularly in complicated scenarios. This suggests that RGS is well-suited for handling U-shaped relationships by iteratively optimizing cut-points within the limited range of the ln*OR*. By addressing the challenge of discretizing multiple variables, we successfully identified the optimal cut-points, enhancing the model's predictive capability and goodness of fit. *AIC* provided an unbiased assessment of model complexity [42], and *AUC* calculations confirmed that RGS excelled in class discrimination. This

Table 5 Comparison of Discretization Methods in Terms of *AIC* under Different Missing Rate Scenarios (PMM Imputation)

Discretization methods	Missing rate = 5%	Missing rate = 10%	Missing rate = 15%	Missing rate = 20%	Missing rate = 25%	Missing rate = 30%	Missing rate = 35%	Missing rate = 40%	Missing rate = 45%	Missing rate = 50%
Median	16,659.48$^{(5)}$	14,850.27$^{(5)}$	16,199.15$^{(5)}$	14,998.76$^{(5)}$	14,848.78$^{(5)}$	16,523.14$^{(5)}$	15,294.25$^{(5)}$	15,713.18$^{(5)}$	15,448.73$^{(5)}$	18,898.45$^{(5)}$
Min-*P*	16,590.45$^{(4)}$	14,779.44$^{(4)}$	16,130.69$^{(4)}$	14,927.24$^{(4)}$	14,777.97$^{(4)}$	16,453.29$^{(4)}$	15,221.31$^{(4)}$	15,646.77$^{(4)}$	15,375.05$^{(4)}$	18,808.32$^{(4)}$
Q_1-Q_3	16,072.52$^{(3)}$	14,348.81$^{(3)}$	15,628.81$^{(3)}$	14,492.30$^{(3)}$	14,347.37$^{(3)}$	15,941.39$^{(3)}$	14,777.79$^{(3)}$	15,159.94$^{(3)}$	14,927.06$^{(3)}$	18,260.03$^{(3)}$
Two cut-points with maximum *OR*	15,516.96$^{(2)}$	13,836.89$^{(2)}$	15,088.19$^{(2)}$	13,975.25$^{(2)}$	13,835.50$^{(2)}$	15,389.96$^{(2)}$	14,250.57$^{(2)}$	14,635.54$^{(2)}$	14,394.51$^{(2)}$	17,608.82$^{(2)}$
RGS	14,979.25$^{(1)}$	13,383.41$^{(1)}$	14,565.78$^{(1)}$	13,517.24$^{(1)}$	13,382.07$^{(1)}$	14,857.09$^{(1)}$	13,783.53$^{(1)}$	14,128.80$^{(1)}$	13,625.92$^{(1)}$	15,794.52$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AIC* values of the model, with $^{(1)}$ representing the best

Table 6 Comparison of Discretization Methods in Terms of *AIC* under Different Missing Rate Scenarios (RF Imputation)

Discretization methods	Missing rate = 5%	Missing rate = 10%	Missing rate = 15%	Missing rate = 20%	Missing rate = 25%	Missing rate = 30%	Missing rate = 35%	Missing rate = 40%	Missing rate = 45%	Missing rate = 50%
Median	13,068.49$^{(4)}$	11,641.94$^{(5)}$	12,706.33$^{(5)}$	11,758.36$^{(5)}$	11,640.78$^{(5)}$	12,960.45$^{(5)}$	11,990.00$^{(5)}$	12,325.14$^{(5)}$	12,111.11$^{(5)}$	14,815.53$^{(5)}$
Min-*P*	12,560.27$^{(3)}$	11,200.33$^{(4)}$	12,213.20$^{(4)}$	11,312.33$^{(4)}$	11,199.21$^{(4)}$	12,457.47$^{(4)}$	11,535.19$^{(4)}$	11,846.80$^{(4)}$	11,651.70$^{(4)}$	14,253.54$^{(4)}$
Q_1-Q_3	12,558.22$^{(5)}$	11,194.40$^{(3)}$	12,211.21$^{(3)}$	11,306.34$^{(3)}$	11,193.28$^{(3)}$	12,455.44$^{(3)}$	11,529.08$^{(3)}$	11,844.88$^{(3)}$	11,645.53$^{(3)}$	14,245.99$^{(3)}$
Two cut-points with maximum *OR*	12,407.38$^{(2)}$	11,076.74$^{(2)}$	12,064.85$^{(2)}$	11,187.51$^{(2)}$	11,075.63$^{(2)}$	12,306.15$^{(2)}$	11,407.90$^{(2)}$	11,702.90$^{(2)}$	11,523.13$^{(2)}$	14,096.26$^{(2)}$
RGS	12,384.66$^{(1)}$	11,065.24$^{(1)}$	12,042.81$^{(1)}$	11,175.89$^{(1)}$	11,064.13$^{(1)}$	12,283.66$^{(1)}$	11,396.06$^{(1)}$	11,681.52$^{(1)}$	11,265.75$^{(1)}$	13,058.72$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AIC* values of the model, with $^{(1)}$ representing the best

Table 7 Comparison of Discretization Methods in Terms of *AUC* under Different Missing Rate Scenarios (RF Imputation)

Discretization methods	Missing rate = 5%	Missing rate = 10%	Missing rate = 15%	Missing rate = 20%	Missing rate = 25%	Missing rate = 30%	Missing rate = 35%	Missing rate = 40%	Missing rate = 45%	Missing rate = 50%
Min-*P*	$0.667^{(2)}$	$0.559^{(5)}$	$0.716^{(5)}$	$0.601^{(5)}$	$0.712^{(3)}$	$0.657^{(5)}$	$0.614^{(5)}$	$0.633^{(5)}$	$0.657^{(5)}$	$0.652^{(4)}$
Median	$0.652^{(4)}$	$0.652^{(2)}$	$0.739^{(3)}$	$0.692^{(4)}$	$0.675^{(4)}$	$0.664^{(3)}$	$0.639^{(4)}$	$0.689^{(2)}$	$0.695^{(3)}$	$0.667^{(2)}$
Q_1-Q_3	$0.656^{(3)}$	$0.648^{(3)}$	$0.732^{(4)}$	$0.697^{(2)}$	$0.632^{(5)}$	$0.661^{(4)}$	$0.680^{(2)}$	$0.686^{(3)}$	$0.685^{(4)}$	$0.656^{(3)}$
Two cut-points with maximum *OR*	$0.651^{(5)}$	$0.644^{(4)}$	$0.746^{(2)}$	$0.693^{(3)}$	$0.717^{(2)}$	$0.670^{(2)}$	$0.677^{(3)}$	$0.684^{(4)}$	$0.698^{(2)}$	$0.651^{(5)}$
RGS	$0.668^{(1)}$	$0.658^{(1)}$	$0.766^{(1)}$	$0.708^{(1)}$	$0.754^{(1)}$	$0.684^{(1)}$	$0.688^{(1)}$	$0.693^{(1)}$	$0.711^{(1)}$	$0.668^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AUC* values of the model, with$^{(1)}$ representing the best

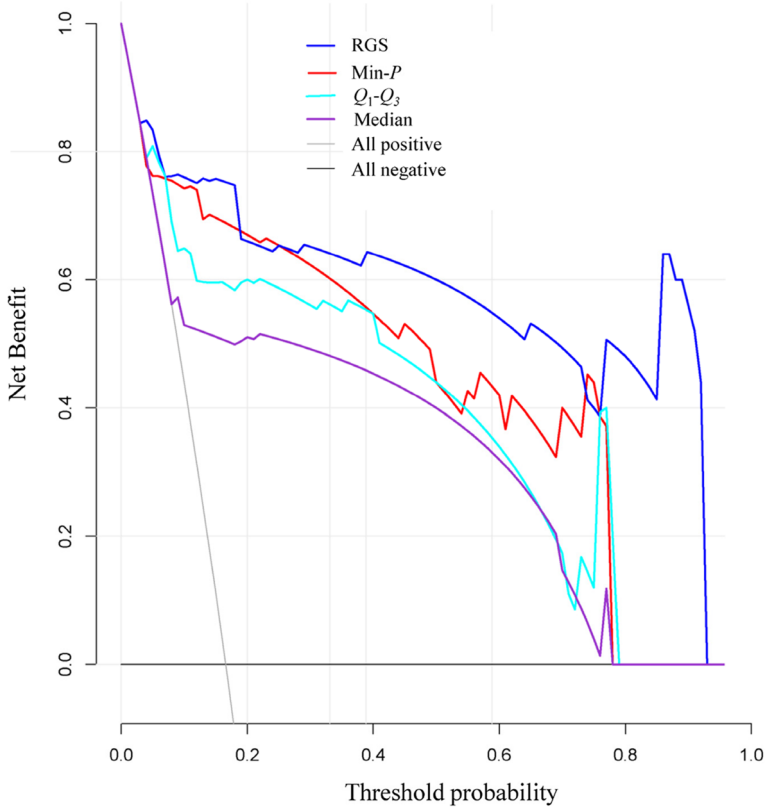

Fig. 2 Comparison of Decision Curves for Different Discretization Methods (Missing Rate = 5%)

Table 8 Advantages of the RGS Method over Three Traditional Discretization Methods

Methods being compared	Net Benefit Curve Area Below Threshold	Number of Patients Avoiding Intervention per 100 Patients (Mean)
Min-*P*	0.169	28.2
Q_1-Q_3	0.195	31.3
Median	0.219	35.4

The horizontal axis of the decision curve represents the threshold probability. When the risk probability of patient *i*, denoted as P_i, reaches a certain value P_t for various evaluation methods, it is defined as positive, indicating that intervention is taken. The calculation method for reducing unnecessary interventions per 100 patients is as follows: (Net benefit of the model-net benefit of all treatments) / (P_t / (1-P_t)) × 100

Table 9 Comparison of Discretization Methods in Terms of *AIC* under Various Symmetric Levels

Discretization methods	Symmetric	Partially symmetric	Completely asymmetric
Min-*P*	1234.31$^{(5)}$	2472.62$^{(5)}$	3173.23$^{(5)}$
Median	1229.28$^{(4)}$	2462.55$^{(4)}$	3160.30$^{(4)}$
Q_1-Q_3	1179.25$^{(3)}$	2362.33$^{(3)}$	3031.23$^{(3)}$
Two cut-points with maximum *OR*	1173.72$^{(2)}$	2351.24$^{(2)}$	3017.44$^{(2)}$
RGS	1063.38$^{(1)}$	2130.21$^{(1)}$	2911.21$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the *AIC* values of the model, with$^{(1)}$ representing the best

underscores RGS's superiority over other discretization methods and highlights the importance of methodological rigor in optimizing predictive accuracy in logistic regression. Despite minor concerns, such as the risk of local optima [43] and reduced computational efficiency compared to simpler methods, our research shows no significant drawbacks. Future work will focus on addressing these limitations and improving model performance. Notably, our efforts are culminating in the release of an R package named "CutpointsRGS", as a user-friendly tool for researchers across various domains.

In our simulation study, RF imputation achieved the best model fit compared to PMM for handling missing data. This efficiency gain can be attributed to RF's capacity to utilize available information more effectively by accommodating nonlinear relationships among predictors. Furthermore, RF simplifies the imputation process by reducing the need to explore associations between predictor variables

Table 10 Comparison of Discretization Methods in Terms of *AUC* under Various Symmetric Levels

Discretization methods	Symmetric	Partially symmetric	Completely asymmetric
Min-*P*	0.716$^{(5)}$	0.651$^{(5)}$	0.559$^{(5)}$
Median	0.732$^{(4)}$	0.652$^{(4)}$	0.643$^{(4)}$
Q_1-Q_3	0.739$^{(3)}$	0.656$^{(3)}$	0.648$^{(3)}$
Two cut-points with maximum *OR*	0.746$^{(2)}$	0.667$^{(2)}$	0.652$^{(2)}$
RGS	0.749$^{(1)}$	0.668$^{(1)}$	0.658$^{(1)}$

$^{(>1\,2\,3\,4\,5)}$ indicates the ranking based on the *AUC* values of the model, with$^{(1)}$ representing the best

and eliminates the necessity of specifying how the outcome should be conditioned in the imputation models for covariates [44]. Other studies [45] also support the notion that RF imputation is particularly valuable for complex epidemiological datasets with missing values."

The symmetry of U-shaped relationships significantly affects the difficulty of identifying cut points. As symmetry increases, the process of determining these cut points becomes less complex [15]. In symmetric U-shaped scenarios, the RGS method consistently outperforms other approaches. Importantly, even in cases where the independent variable exhibits an asymmetric U-shaped relationship with log(η), RGS maintains superior performance. Moreover, as the degree of asymmetry increases, the advantages of the RGS method remain pronounced. Thus, the RGS method demonstrates broader applicability and stability across varying symmetry levels in U-shaped relationships, underscoring its robustness in real-world applications.

U-shaped effects arise from complex interactions among various factors, and understanding these dynamics is essential for both theoretical exploration and practical applications in public health and psychology [46, 47]. In our study, we identified U-shaped relationships among exercise frequency, METs, and depression. Supporting our findings, a large cohort study by Harvey et al. [48] demonstrated that the protective effect of exercise against depression is associated with low levels of

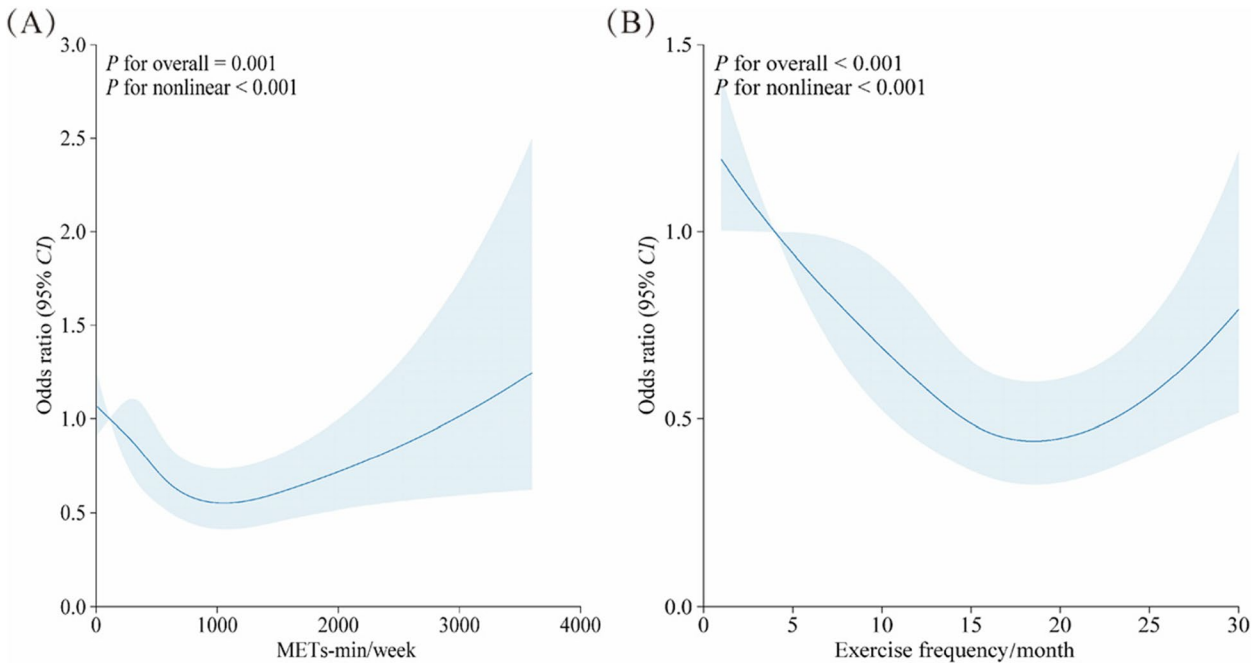

Figure3 Describing the relationships between METs, exercise frequency, and depression using a nonlinear regression model

Table 11 Performance of different estimated cut-points by RGS method

Cut-points	MET/week$_L$	MET/week$_U$	Freq/month$_L$	Freq/month$_U$	OR_{MET}(95%*CI*)	OR_{Freq}(95%*CI*)	*AIC*	R^2	−2Loglikelihood
Original cut-points	751.3	1847.6	16.5	23.7	1.18(1.10–1.30)	1.58(1.44–1.74)	17,042	0.031	13,636.40
Optimal cut-points	610.9	2203.7	12.6	26.5	1.59(1.24–2.06)	1.36(1.15–1.58)	16,903	0.037	13,579.28

L represents the lower cut-point, and footnote *U* represents the upper cut-point

Table 12 The predictive capacity and goodness-of-fit among different methods

Methods	*AUC*	*AIC*	Adjusted R^2
Min-*P*	0.731$^{(3)}$	17,968$^{(4)}$	0.029$^{(4)}$
Median	0.729$^{(4)}$	17,034$^{(3)}$	0.031$^{(3)}$
Q_1-Q_3	0.737$^{(2)}$	16,975$^{(2)}$	0.034$^{(2)}$
RGS	0.751$^{(1)}$	16,903$^{(1)}$	0.037$^{(1)}$

$^{(1\,2\,3\,4\,5)}$ indicates the ranking based on the model evaluation values, with$^{(1)}$ representing the best

activity, showing no additional benefits beyond one hour per week. Furthermore, previous research indicates that vigorous exercise can increase oxidative stress and reduce antioxidant defenses, particularly through rapid vitamin E turnover [49]. Zunszain et al. [50] also found that cortisol responses are positively correlated with the intensity and duration of physical activity, with elevated adrenocorticotropic hormone levels observed after intense exercise. Taken together, these studies suggest that high levels of physical activity may negatively impact the immune system by increasing oxidative stress and cortisol levels, potentially leading to depression. Additionally, extreme exercise has been linked to cardiotoxicity [51–53] and various musculoskeletal complications [54], both of which are known to elevate the risk of depression.

Much literature has criticized categorization for inevitably resulting in information loss [10–12, 55]. In logistic regression applications, discretization is necessary for certain purposes; however, this study does not encourage researchers to blindly discretize continuous independent variables. The results of this study demonstrate that discretization leads to superior overall performance from a statistical perspective. Meanwhile, categorization lends itself to a visual presentation of the data that is more natural and familiar to most readers or an outcome could be based on a particular variable falling within a prescribed window [56]. Especially when an appropriate parametric model is absent, visual presentation provides the optimal presentation of the exposure–response relation. Potentially, it may reduce the arbitrariness of the choice of cut-points.

Our study addresses several challenges in logistic regression modeling, particularly in the context of discretizing continuous predictors. One major challenge is the improper handling of continuous predictors, often leading to the loss of information and oversimplification of relationships [9]. By comparing different discretization methods, we were able to identify a more effective approach for capturing the non-linear relationships between predictors and outcomes. This addresses the challenge of properly modeling complex relationships, mainly in scenarios with U-shaped relationships and varying levels of missing values. The models proposed in this study incorporate both categorical covariates that influence the outcome and continuous covariates in linear form. We intend to face directly the complexities that real research scenarios may present [57, 58].

Discretizing predictor variables is a computational operation that involves information loss, whether the relationship between the independent variable and the outcome is monotonic or U-shaped. However, in medical practice, obtaining a reasonable estimate of the *OR* is necessary for clearly prioritizing the intervenable predictive factors to formulate intervention strategies [59, 60]. The discretization method we propose sacrifices the original pattern of the U-shaped relationship in the predictive model [61], and may encounter challenges related to separation or missing binary responses[62]. However, it provides direct guidance for further intervention. More importantly, the statistical evaluation measures are optimal. Both simulation and empirical studies make the further extension of this method feasible.

Our study exhibited the importance of methodological considerations in improving the predictive performance of logistic regression models, providing valuable insights for researchers and practitioners in the medical field. The use of appropriate discretization methods, such as RGS, can significantly improve the predictive performance of logistic regression models, particularly in scenarios where predictors possess U-shaped relationships with the outcome. We look forward to medical colleagues analyzing their research data using the method proposed in this study, to further investigate the generalizability of the findings obtained in this study. However, a limitation of our study is that it focused solely on two U-shaped predictors. Further research is needed to explore the applicability of RGS in more complicated scenarios involving multiple predictors or diverse relationships. Additionally, the method may not perform optimally with small sample sizes, a limitation that should be considered in future applications of the model. In our future work, we plan to conduct these additional validations to further strengthen our conclusions and demonstrate the method's efficacy across diverse datasets.

Conclusions

In summary, our study introduces the RGS method as an effective approach for identifying optimal cut-points among multiple continuous predictors exhibiting U-shaped relationships with $\log(\eta)$ in analysis. This method serves as a valuable tool for researchers grappling with the challenge of discretizing continuous predictors within logistic regression models. We strongly advocate for a preliminary exploration of the relationships between continuous predictors and outcomes to ascertain potential U-shaped patterns. Real-world data

will inevitably exhibit non-typical U-shaped patterns, and the sample size and symmetry may also become worse than ideal. Thus, our simulation scenarios have fully considered the variety of data. There is enough confidence to believe that the RGS method can contribute to establishing an ideal model.

Abbreviations

RGS	The recursive gradient scanning method
CI	Confidence interval
AIC	Akaike information criterion
OR	Odds ratio
ln*OR*	The natural logarithm of odds ratio
HR	Hazard ratio
log(*η*)	Ln*OR* or ln*HR*
Min-*P*	The minimum *P*-value approach
SE	Standard error
Q_1-Q_3	Lower and upper quantile values
TCPMOR	Two cut-points with maximum *OR* value method
CutpointsOEHR	The optimal equal-HR method to estimate cut-points
PMM	Predictive mean matching
RF	Random forest
BRFSS	Behavioral risk factor surveillance system

Acknowledgements
The authors acknowledge the selfless work of individuals who assisted with the design of the research and provided guidance for writing the paper.

Authors' contributions
SY was involved in the conception of the article, analyzing the data, interpreting the outcomes, and preparing the initial manuscript. HS, NZ, YH, YG, YF, and YL assisted in the critical review of the manuscript and interpretation of the outcomes. PX collected the data and performed data cleaning. A Hilowle polished the manuscript in language and grammar. JZ contributed to the conception of the article and critically reviewed the manuscript. All authors have reviewed and approved the manuscript's content.

Funding
Open access funding provided by the Natural Science Foundation of Guangdong Province, China (2022A1515011237, 2023A1515011951). The Foundation had no influence on the study design, data collection, analysis, interpretation, writing, or decision to submit the manuscript for publication.

Data availability
The BRFSS datasets are publicly available online at https://www.cdc.gov/brfss/index.html (accessed on 16 December 2023). The specific data cleaning rules are provided in the Methods section. The cleaned dataset used in the present study can be obtained by contacting the corresponding author, Jinxin Zhang.

Declarations

Ethics approval and consent to participate
This study did not involve human participants, human data, or human tissue; therefore, no ethical approval or consent to participate was necessary.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Supplementary Information
Supplementary 1.

Author details
[1]Department of Medical Statistics, School of Public Health, Sun Yat-Sen University, Guangzhou 510080, China. [2]The People's Hospital of Jiangmen, No. 172 Gaodi Li, Pengjiang District, Jiangmen, Guangdong 529000, China. [3]Department of Cardiology, The Third Affiliated Hospital of Sun Yat-Sen University, Guangzhou 510630, China. [4]The First Affiliated Hospital of Bengbu Medical University, Bengbuaq, Anhui 233004, China.

Received: 13 July 2024 Accepted: 25 February 2025
Published online: 12 March 2025

References

1. Sheng M, Yang J, Bao M, Chen T, Cai R, Zhang N, et al. The relationships between step count and all-cause mortality and cardiovascular events: A dose-response meta-analysis. J Sport Health Sci. 2021;10(6):620–8.
2. Baralić K, Javorac D, Marić Đ, Đukić-Ćosić D, Bulat Z, Miljaković EA, et al. Benchmark dose approach in investigating the relationship between blood metal levels and reproductive hormones: Data set from human study. Environ Int. 2022;165.
3. Chiu HY, Lee HC, Chen PY, Lai YF, Tu YK. Associations between sleep duration and suicidality in adolescents: A systematic review and dose-response meta-analysis. Sleep Med Rev. 2018;42:119–26.
4. Watanabe D, Yoshida T, Watanabe Y, Yamada Y, Kimura . A U-Shaped Relationship Between the Prevalence of Frailty and Body Mass Index in Community-Dwelling Japanese Older Adults: The Kyoto-Kameoka Study. Journal of Clinical Medicine 2020; 9 (5): 1367.
5. Ma W, Yan Z, Wu W, Li D, Zheng S, Lyu J. Dose-Response Association of Waist-to-Height Ratio Plus BMI and Risk of Depression: Evidence from the NHANES 05–16. Int J Gen Med. 2021;14:1283–91.
6. Tagawa R, Watanabe D, Ito K, Ueda K, Nakayama K, Sanbongi C, et al. Dose-response relationship between protein intake and muscle mass increase: a systematic review and meta-analysis of randomized controlled trials. Nutr Rev. 2020;79(1):66–75.
7. Hu L, Hu G, Xu BP, Zhu L, Zhou W, Wang T, et al. U-Shaped Association of Serum Uric Acid With All-Cause and Cause-Specific Mortality in US Adults: A Cohort Study. J Clin Endocrinol Metab. 2020;105(3):e597–609.
8. Nakayama S, Satoh M, Tatsumi Y, Murakami T, Muroya T, Hirose T, et al. Detailed association between serum uric acid levels and the incidence of chronic kidney disease stratified by sex in middle-aged adults. Atherosclerosis. 2021;330:107–13.
9. Ma J, Dhiman P, Qi C, Bullock G, van Smeden M, Riley RD, et al. Poor handling of continuous predictors in clinical prediction models using logistic regression: a systematic review. J Clin Epidemiol. 2023;161:140–51.
10. Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006;332(7549):1080.
11. Cumsille J. Effect of dichotomizing continuous variables in regression models [dissertation]. Department of Biostatistics, University of North Carolina at Chapel Hill; 1998.
12. Fernandes A, Malaquias C, Figueiredo D, Da Rocha E, Lins R. Why Quantitative Variables Should Not Be Recoded as Categorical. J App Math Phys. 2019;7:1519–30.
13. 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(23):4124–35.
14. He XY, Yang S, Fei Y, Zhang JX. TCPMOR 1.0. R package, https://cran.r-project.org/web/packages/TCPMOR/index.html; 2024 [accessed 13 June 2024].
15. Chen YM, Huang JL, He XY, Gao YX, Mahara G, Lin ZC, et al. A novel approach to determine two optimal cut-points of a continuous predictor with a U-shaped relationship to hazard ratio in survival data: simulation and application. BMC Med Res Methodol. 2019;19(1):96.
16. Fleischer J, Bezes P, James O, Yesilkagit K. The politics of government reorganization in Western Europe. Governance. 2023;36(1):255–74.x
17. Sim EJ, Ko KP, Ahn C, Park SM, Surh YJ, An S, et al. Isoflavone intake on the risk of overall breast cancer and molecular subtypes in women at high risk for hereditary breast cancer. Breast Cancer Res Treatment. 2020;184(2):615–26.
18. Uehara Y, Hakozaki T, Kitadai R, Narita K, Watanabe K, Hashimoto K, et al. Association between the baseline tumor size and outcomes of patients with non-small cell lung cancer treated with first-line immune checkpoint inhibitor monotherapy or in combination with chemotherapy. Transl Lung Cancer Res. 2022;11(2):135–49.
19. Mauck DE, Fennie KP, Ibañez GE, Fenkl EA, Sheehan DM, Maddox LM, et al. Gay Neighborhoods: Can They Be Identified in a Systematic Way Using Latent Class Analysis?.Archives of Sexual Behavior 2022; 51(7): 3395–3401.

20. Li D, Li S, Xia Z, Cao J, Zhang J, Chen B, et al. Prognostic significance of pretreatment red blood cell distribution width in primary diffuse large B-cell lymphoma of the central nervous system for 3P medical approaches in multiple cohorts.Epma j 2022; 13 (3): 499–517.
21. Starzer AM, Steindl A, Mair MJ, Deischinger C, Simonovska A, Widhalm G, et al. Systemic inflammation scores correlate with survival prognosis in patients with newly diagnosed brain metastases. Br J Cancer. 2021;124(7):1294–300.
22. Cui Y, Zhou HL, Wei MH, Song WJ, Di DS, Zhang RY, et al. Multiple vitamin co-exposure and mortality risk: A prospective study. Clin Nutr. 2022;41(2):337–47.
23. EFSA Scientific Committee, More S, Benford D, Hougaard Bennekou S, Bampidis V, et al. Opinion on the impact of non-monotonic dose responses on EFSA's human health risk assessments. EFSA Journal. European Food Safety Authority 2021; 19 (10): e06877.x
24. Park C S, Choi Y-J, Rhee T-M, Lee HJ, Lee HS, Park JB, et al. U-Shaped Associations Between Body Weight Changes and Major Cardiovascular Events in Type 2 Diabetes Mellitus: A Longitudinal Follow-up Study of a Nationwide Cohort of Over 1.5 Million. Diabetes Care 2022; 45 (5): 1239–1246.
25. Liu F, Walters S J, Julious S A. Design considerations and analysis planning of a phase 2a proof of concept study in rheumatoid arthritis in the presence of possible non-monotonicity[J].BMC Medical Research Methodology,2017, 17 (1): 149.
26. Wu S, Zhang Q, Li Y, Liang H. Assessment of nonlinear dose-response relationships via nonparametric regression. J Biopharm Stat. 2024;34(1):136–45.
27. Wu S, Li X, Xia Y, Liang H. A novel model-checking approach for dose-response relationships. Stat Methods Med Res. 2021;30(9):2119–29.
28. Yuan Y, Yin G. Dose-response curve estimation: a semiparametric mixture approach. Biometrics. 2011;67(4):1543–54.
29. Ducharme GR, Fontez B. A smooth test of goodness-of-fit for growth curves and monotonic nonlinear regression models. Biometrics. 2004;60(4):977–86.
30. Lei Y, Tang K. Learning Rates for Stochastic Gradient Descent With Nonconvex Objectives. IEEE Trans Pattern Anal Mach Intell. 2021;43(12):4505–11.
31. Wand MP, Coull BA, French JL, Ganguli B. SemiPar 1.0. R package, http://cran.r-project.org. r-project. org,2005; 2024 [accessed 13 June 2024].
32. Van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med. 2000;19(24):3401–15.
33. Harel O, Mitchell EM, Perkins NJ, Cole SR, Tchetgen Tchetgen EJ, Sun B, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. Am J Epidemiol. 2018;187(3):576–84.
34. Remington PL, Smith MY, Williamson DF, Anda RF, Gentry EM, Hogelin GC. Design, characteristics, and usefulness of state-based behavioral risk factor surveillance: 1981–87. Public Health Rep. 1988;103(4):366–75.
35. Xu P, Huang Y, Hou Q, Cheng J, Ren Z, Ye R, et al. Relationship between physical activity and mental health in a national representative cross-section study: Its variations according to obesity and comorbidity. J Affect Disord. 2022;308:484–93.
36. Ma Y, Liang L, Zheng F, Shi L, Zhong B, Xie W. Association Between Sleep Duration and Cognitive Decline. JAMA Netw Open. 2020;3(9): e2013573.
37. Zhou C, Wu Q, Ye Z, Liu M, Zhang Z, Zhang Y, et al. Inverse Association Between Variety of Proteins With Appropriate Quantity From Different Food Sources and New-Onset Hypertension. Hypertension. 2022;79(5):1017–27.
38. Zhou C, Zhang Z, Liu M, Zhang Y, Li H, He P, et al. Dietary carbohydrate intake and new-onset diabetes: A nationwide cohort study in China. Metabolism. 2021;123: 154865.
39. Kubota Y, Iso H, Yamagishi K, Sawada N, Tsugane S. Daily Total Physical Activity and Incident Cardiovascular Disease in Japanese Men and Women: Japan Public Health Center-Based Prospective Study. Circulation. 2017;135(15):1471–3.
40. Haans RFJ, Pieters C, He ZL. Thinking about U: Theorizing and testing U- and inverted U-shaped relationships in strategy research. Strateg Manag J. 2016;37(7):1177–95.
41. Lind JT, Mehlum H. With or Without U? The Appropriate Test for a U-Shaped Relationship. Oxford Bull Econ Stat. 2010;72(1):109–18.
42. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). Psychol Methods. 2012;17(2):228–43.
43. Shireman EM, Steinley D, Brusco MJ. Local Optima in Mixture Modeling. Multivariate Behav Res. 2016;51(4):466–81.
44. Hong S, Lynn HS. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Med Res Methodol. 2020;20(1):199.
45. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. Am J Epidemiol. 2014;179(6):764–74.
46. Kim SY, Jeon SW, Shin DW, Oh KS, Shin YC, Lim SW. Association between physical activity and depressive symptoms in general adult populations: An analysis of the dose-response relationship. Psychiatry Res. 2018;269:258–63.
47. Kim SY, Park JH, Lee MY, Oh KS, Shin DW, Shin YC. Physical activity and the prevention of depression: A cohort study. Gen Hosp Psychiatry. 2019;60:90–7.
48. Harvey SB, Øverland S, Hatch SL, Wessely S, Mykletun A, Hotopf M. Exercise and the Prevention of Depression: Results of the HUNT Cohort Study. Am J Psychiatry. 2018;175(1):28–36.
49. Pingitore A, Lima GP, Mastorci F, Quinones A, Iervasi G, Vassalle C. Exercise and oxidative stress: potential effects of antioxidant dietary strategies in sports. Nutrition. 2015;31(7–8):916–22.
50. Zunszain PA, Anacker C, Cattaneo A, Carvalho LA, Pariante CM. Glucocorticoids, cytokines and brain abnormalities in depression. Prog Neuropsychopharmacol Biol Psychiatry. 2011;35(3):722–9.
51. O'Keefe JH, Franklin B, Lavie CJ. Exercising for health and longevity vs peak performance: different regimens for different goals. Mayo Clin Proc. 2014;89(9):1171–5.
52. Trivax JE, Franklin BA, Goldstein JA, Chinnaiyan KM, Gallagher MJ, deJong AT, et al. Acute cardiac effects of marathon running. J Appl Physiol (1985) 2010; 108(5): 1148–1153.
53. Wilson M, O'Hanlon R, Prasad S, Deighan A, Macmillan P, Oxborough D, et al. Diverse patterns of myocardial fibrosis in lifelong, veteran endurance athletes. J Appl Physiol (1985) 2011; 110(6): 1622–1626.
54. Franklin BA, Billecke S. Putting the benefits and risks of aerobic exercise in perspective. Curr Sports Med Rep. 2012;11(4):201–8.
55. Polley MYC, Dignam JJ. Statistical considerations in the evaluation of continuous biomarkers. J Nucl Med. 2021;62(5):605–11.
56. Zaidi NA, Du Y, Webb GI. On the effectiveness of discretizing quantitative attributes in linear classifiers. IEEE Access. 2020;8:198856–71.
57. Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biom J. 2018;60(3):431–49.
58. Barrio I, Arostegui I, Rodríguez-Álvarez MX, Quintana JM. A new approach to categorising continuous variables in prediction models: Proposal and validation. Stat Methods Med Res. 2017;26(6):2586–602.
59. 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(4):675–80.
60. Perperoglou A, Sauerbrei W, Abrahamowicz M, Schmid M. A review of spline function procedures in R[J].BMC Medical Research Methodology 2019; 19: 1–46.
61. Yang S, Zhang N, Liang Z, Han Y, Luo H, Ge Y, et al. Examining the U-shaped relationship of sleep duration and systolic blood pressure with risk of cardiovascular events using a novel recursive gradient scanning model. Front Cardiovasc Med. 2023;10:1210171.
62. Maity AK, Pradhan V, Das U. Bias reduction in logistic regression with missing responses when the missing data mechanism is nonignorable. Am Stat. 2018;73(4):340–9.

Publisher's Note