Hypothesis test of arbitrary parametric structure in a generalized additive model

Yihe Yang[1*] and Xiaofeng Zhu[1]

[1*]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, 10900 Euclid Ave, Cleveland, 44106, OH, USA.

*Corresponding author(s). E-mail(s): yxy1234@case.edu;

Abstract

Before applying a nonparametric model such as a generalized additive model (GAM), it is natural to ask whether a simpler parametric model suffices to capture the variation in the data. To address this fundamental question, we propose a new methodology named Test for Arbitrary Parametric Structure (TAPS), which provides estimation and inference tools to determine whether a parametric structure sufficiently describes a target function in a GAM. The key strategy of TAPS is to translate the test of parametric structure into the test of variance of the random effect, and the novelty of TAPS lies in showing how to construct this translation for an arbitrary parametric structure, including the linearity, piecewise linearity with slope changes, and linearity discontinuity with jumps. We illustrate the utility of TAPS across diverse scientific domains by using the UK Biobank data. Specifically, we applied TAPS to reveal widespread nonlinearity in polygenic risk score effects, though prediction improvement over the linear model was limited for most traits. Alternatively, we employed TAPS to identify the causal effects of retirement that change various health and lifestyle traits, using regression discontinuity and kink designs.

Keywords: Generalized Additive Model; Hypothesis Test; Polygenic Risk Score; Regression Discontinuity Design; Regression Kink Design.

1 Introduction

Generalized additive model (GAM) [1], an extension of the generalized linear model (GLM) [2], is used to model the non-linear effects of covariates on an outcome. Since

its first introduction by Hastie and Tibshirani [3], GAM has undergone substantial methodological and computational development. For instance, the R package `mgcv` exemplifies these advances, which provides a broad class of smoothers [4–6] and automatic smoothing parameter selection using information criteria [7, 8]. Furthermore, GAMs have demonstrated wide applicability across scientific disciplines. In epidemiology, they have been used to smooth daily incidence data and to reveal trends in childhood type 1 diabetes during the COVID-19 pandemic [9]. In ecology, GAMs have been employed to model the nonlinear relationships between environmental factors and biological responses in chaotic or near-chaotic ecological dynamic systems [10]. Furthermore, in spatial transcriptomics, GAMs are used to characterize the spatial distribution of cell types across tissues [11].

Before applying a nonparametric model such as a GAM, it is natural to ask whether a simpler parametric model suffices to capture the variation in the data. If correctly specified, a parametric model is often of greater scientific interest due to its interpretability and simplicity. For example, the genetic effect of an exposure is often assumed to be in a linear relationship with an outcome [12], yet growing evidence suggests the presence of non-linear genetic effects [13]. There is an ongoing discussion as to whether linear additive genetic models are sufficient to capture the majority of heritability of human complex traits [14, 15]. In addition, beyond interpretability and simplicity, some statistical models such as causal inference often require the presence of specific parametric forms. For example, in regression discontinuity design (RDD), the effect of treatment is modeled as a sudden jump at a cutoff point [16]. Regression kink design (RKD) is similar, but instead assumes a change in the slope at the cutoff. In both two designs, the causal interpretation relies on a specific parametric form, either a jump or a change in slope. Nonparametric models may better capture complex associations but do not, on their own, enable causal inference of treatment assignment in RDD or RKD.

Compared to the wide variety of estimation methods available for GAM, the tools for hypothesis testing remain relatively limited. To the best of our knowledge, among all parametric function forms with potential practical interpretations, current methods only allow testing whether the effect of a covariate follows a polynomial structure [17–20]. However, for other parametric structures, such as piecewise linearity or linear discontinuities, there are no established approaches for conducting corresponding hypothesis tests. Moreover, existing hypothesis tests are restricted to univariable nonparametric models, where only one nonlinear term is allowed to exist in the model. How to perform hypothesis testing for arbitrary parametric structures within a GAM remains an open question.

To address the lack of hypothesis testing approaches, we propose the Test for Arbitrary Parametric Structure (TAPS), a new method for assessing whether a parametric form can adequately capture a target function within a GAM. TAPS applies a key strategy of translating the test of a parametric structure into a variance component test, and introduces a general approach to construct such a translation for arbitrary parametric forms. In addition, we develop an R package, `mgcv.taps`, which integrates seamlessly with `mgcv`, one of the most powerful GAM software. This integration enables TAPS to be easily adopted without additional learning cost, to support

a wide range of outcome types including survival and ordinal responses [21], and to scale efficiently to biobank-scale datasets [22].

To demonstrate the practical utility of TAPS, we conducted two empirical analyses using data from the UK Biobank [23]. First, we found that the relationship between a trait and its polygenic risk score (PRS) was often non-linear, but the prediction improvements were generally limited for most traits. Second, we performed RDD and RKD around the retirement age of 65 to evaluate causal effects on health outcomes, where TAPS was used to assess whether the assumed parametric structure held, which is essential for the validity of causal interpretation. To facilitate practical adoption, we present three examples with analysis codes in the **Supplementary Materials**, showcasing the end-to-end application of TAPS across different scenarios.

2 Result

2.1 Overview of Method

A standard GAM takes the following representation

$$y_i \sim \mathcal{Y}(\mu_i, \phi_0), \quad g(\mu_i) = \eta_i = \boldsymbol{Z}_i^\top \boldsymbol{\gamma} + f_1(x_{i1}) + \cdots + f_J(x_{iJ}), \tag{1}$$

where y_i is a response variable belonging to a certain distribution $\mathcal{Y}(\mu_i, \phi_0)$ with mean μ_i and dispersion parameter ϕ_0, g is a known link function, η_i is a linear predictor, $\boldsymbol{Z}_i$ is the ith row of a parametric model matrix $\mathbf{Z}$, $\boldsymbol{\gamma}$ is a regression coefficient vector, and f_j is a smooth function of covariate x_{ij}.

TAPS performs a hypothesis test on whether a target function in (1), such as f_1, adheres to a parametric structure. The null and alternative hypotheses that TAPS addresses are

$$\mathrm{H}_0 : f_1 \in \mathcal{H}_A \qquad \text{v.s.} \qquad \mathrm{H}_1 : f_1 \in \mathcal{H} \text{ and } f_1 \notin \mathcal{H}_A, \tag{2}$$

where $\mathcal{H}_A$ is the space of functions with the target parametric structure and $\mathcal{H}$ is a general functional space. In particular, $\mathcal{H}$ should be sufficiently expressive to accommodate the nonsmooth functions in $\mathcal{H}_A$, but also capable of approximating smooth functions when $\mathcal{H}_A$ does not reflect the true structure. To achieve this, we propose a new mixed-effects representation of $f_1(x)$:

$$f_1(x) = \boldsymbol{A}_1(x)^\top \boldsymbol{\alpha}_1 + b_1(x), \tag{3}$$

where $\boldsymbol{A}_1(x) = (A_{11}(x), \ldots, A_{1q}(x))^\top$ is a vector of the basis functions of the parametric structure, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1q})^\top$ is a fixed vector, $b_1(x)$ is a non-parametric term orthogonal to $\boldsymbol{A}_1(x)^\top \boldsymbol{\alpha}_1$. According to the equivalence between RKHS and Gaussian processes [24], $b_1(x)$ can be regarded as a Gaussian process with covariance function

$$\mathrm{cov}(b_1(x_{i1}), b_1(x_{j1})) = \phi_1 K_1(x_{i1}, x_{j1}), \tag{4}$$

$K_1(x_i, x_j)$ is a known bivariate function, and $\phi_1 \geq 0$ is the variance of this Gaussian process. We call the equation (3) a mixed-effect representation of a function, as it shares the same form of the generalized linear mixed model (GLMM). As a result, the hypothesis (2) reduces to

$$\mathrm{H}_0 : \phi_1 = 0 \qquad \text{v.s.} \qquad \mathrm{H}_1 : \phi_1 > 0. \tag{5}$$

The techniques of GLMM [25], including the Wald test [26] and score test [18], can yield the p-value of the hypothesis test (5).

Methodologically, we address two key challenges in the implementation of TAPS. First, we show how to construct the random effect $b_1(x)$ that is adaptively orthogonal to $\boldsymbol{A}_1(x)$, guaranteeing an identifiable mixed-effect representation. We have developed the R package `mgcv.taps`, which enables `mgcv` to recognize the new mixed-effect representation (3), allowing an easy and computationally efficient implementation of TAPS in practice. Second, we explore the methodology to perform the hypothesis test (5) within a GAM model. Specifically, we design two testing procedures: a Wald test [26] and a score test [18]. Both tests exhibit well-controlled type I error rates, and the score test tends to be more powerful. However, the Wald test can be applied to all types of outcomes supported by `mgcv`, whereas the score test is currently restricted to the exponential family of distributions. We provide both testing options so that users can choose the most appropriate method depending on their analysis goals.

Fig. 1A shows a classic example used in literature, the motorcycle acceleration data, which records head acceleration during a simulated crash [21]. In this example, our goal is to test whether the time trend can be segmented into four phases, with changepoints marked by vertical dashed lines. The null hypothesis assumes that the function lies within a piecewise linear space $\mathcal{H}_A$, whose basis functions are shown in Fig. 1C. To test for potential deviations from this structure, TAPS constructed $b_1(x)$ from a space $\mathcal{H}_B = \mathcal{H}/\mathcal{H}_A$, with basis functions adaptively learned from the data, as shown in Fig. 1D. Fig. 1B displays the fitted trend under the mixed representation. The first three phases aligned with the piecewise linear form, while the fourth showed a slightly non-linear effect. Fig. 1E summarizes the overall TAPS workflow. Details of TAPS are shown in **Method** and **Supplementary Materials**.

2.2 Simulation

In the simulation study, we consider the following GAM:

$$g(\mathrm{E}(y_i)) = \eta_i = \alpha_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}),$$

where y_i is the response variable following a Gaussian, Poisson, or binary distribution; α_0 is an intercept; and f_1 is the target function to be tested. We evaluate four structural scenarios for f_1:

- Linearity: $f_1(x_1) = \alpha_1 x_1$, representing a simple linear structure.
- Piecewise Linearity: $f_1(x_1) = \alpha_1 x_1 + \alpha_2(x_1 - 0.5)\mathrm{I}(x_1 > 0.5)$, a piecewise linear function with a change of slope at $x_1 = \nu_1 = 0.5$.

- Linearity Discontinuity: $f_1(x_1) = \alpha_1 \mathrm{I}(x_1 > 0.5) + \alpha_2(x_1 - 0.5) + \alpha_3(x_1 - 0.5)\mathrm{I}(x_1 > 0.5)$, a discontinuous linear structure with a jump at $x_1 = \nu_0 = 0.5$.
- Linear Interaction: $f_{12}(x_1, x_2) = \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2$. TAPS can test whether a multivariate function follows a specific parametric structure like linear interaction.

In all scenarios, the functions f_3 and f_4 remain fixed, while f_2 is set to zero in the first three settings. The kernel function used to construct $b_1(x)$ is Matérn correlation function [27]. Full expressions for f_1 through f_4 (and f_{12}), as well as additional simulation settings, and codes, are provided in the **Supplementary Materials**. Each scenario was replicated 1,000 times.

We first evaluate whether the null distributions of the Wald and score test statistics are correctly specified. If so, the associated p-values should follow a uniform distribution on the interval $[0, 1]$. When the sorted p-values systematically lie above the diagonal in a quantile-quantile (QQ) plot, it indicates inflated Type II error rates (i.e., the test is overly conservative). Conversely, when the sorted p-values lie below the diagonal, it suggests inflated Type I error rates (i.e., the test is overly liberal). Fig. 2A-2B shows the QQ plots of p-values under the null across the four testing scenarios. We observe that the empirical quantiles of the score test closely follow the uniform distribution, regardless of the sample size or whether the response distribution is continuous or discrete. These results confirm that the score test controls both Type I and Type II error rates well. In contrast, the p-values from the Wald test tend to deviate upward from the uniform quantiles, indicating that the Wald test is more conservative and prone to inflated Type II error. This conservativeness arises from the penalization of nonparametric terms in `mgcv`, which affects the estimation of their effective degrees of freedom (edf). Since the null distribution of the Wald statistic depends on numerical approximation [26], it may become inaccurate when the nonparametric component has near-zero edf, thus leading to inflated Type II errors.

We then investigated the power of the two testing procedures. To simulate deviations from the null hypothesis, we introduce a deviation parameter d, where larger values of d correspond to stronger departures of f_1 from the parametric structure under the null. As shown in Fig. 2C-2D. The score test consistently achieves high power in detecting deviations from the null across all scenarios under Gaussian and Poisson outcomes, with slightly lower power under the binary distribution. As the deviation parameter d increases, power quickly approaches 1, especially when the sample size exceeds 1500. The Wald test is generally less powerful than the score test but still shows a clear upward trend and converges to high power with moderate sample sizes. Across the four scenarios, the linearity and interaction tests are particularly sensitive to deviations, with both tests attaining high power rapidly. In contrast, the piecewise linearity and discontinuity scenarios require larger sample sizes or stronger deviations (larger d) to achieve comparable power. Overall, both tests become increasingly effective as the deviation from the null grows and the available data become more informative.

Despite its strong performance, the score test is limited to outcomes from the exponential family, whereas the Wald test, though slightly less powerful, is not subject to this restriction. In addition, the Wald test can be applied to quantile generalized additive models [28], such as the median GAM (quantile = 0.5). In the **Supplementary**

Materials, we further examine the performance of the Wald test when the outcome is a survival time or an ordinal categorical variable, as well as its robustness to outliers in GAM and median GAM (Figure S6-S8).

2.3 Linearity Test of Polygenic Risk Scores

PRS is a score of an individual's genetic liability to a trait or disease, calculated based on genotype profiles and genome-wide association study (GWAS) results [12]. PRS has long been assumed to have a linear relationship with phenotypic outcomes [14]. However, recent studies suggest that this assumption may be overly simplistic, and PRS may exert nonlinear effects. For example, He et al. [29] developed a deep learning method called DeLIVR, which identified additional genes associated with HDL cholesterol (HDL-C) and LDL cholesterol (LDL-C) via nonlinear patterns that traditional linear approaches would miss. In addition, by using deep-learning-based approaches, Kelemen et al. [30] found that the gene-gene (GxG) and gene-environment (GxE) interactions contribute 7% and 4% gains in the proportion of variance explained (PVE), respectively.

Motivated by these emerging observations of potential nonlinearity in PRS effects, we applied TAPS to UK Biobank data to formally test whether linear models of PRS are sufficient to capture the genetic variation underlying several complex traits. We analyzed 37 continuous traits and 8 binary traits, including markers of blood pressure, lipids, glucose, liver function, kidney function, type 2 diabetes, and cardiovascular diseases. Our analysis is based on the genetically inferred European population in the UK Biobank [31]. The sample sizes range from 340,864 (peripheral artery disease) to 411,463 (hypertension) with a median of 392,102. We observed that many traits exhibited skewed distributions. Therefore, we performed analyses using both the raw trait values and their inverse-rank normalized (IRNT) counterparts, and used the corresponding GWAS summary statistics from the Neale Lab [32] for each setting. In addition, PRSs were estimated using SBayesRC [33] and predicted for individuals of European ancestry in the UK Biobank. Additional details are provided in **Method** and **Supplementary Materials**.

We used the score test to evaluate whether the effect of a PRS on each trait could be adequately modeled as linear. As shown in Fig. 3A, the negative $\log_{10}(p)$ values from the score tests varied substantially across traits, yet the vast majority are far below 0.05/48, a threshold based on Bonferroni correction, indicating that the linearity assumption rarely holds. To visualize these effects, Fig.3B presents example continuous traits under raw and IRNT traits, and Fig.3C shows example binary traits with estimated nonlinear PRS effects. For instance, total bilirubin (TBL) and C-reactive protein (CRP) exhibit strong statistical evidence of nonlinearity, while height and body mass index (BMI) showed weaker but still significantly non-linear effects. However, when analyzing the IRNT traits, we observed a systematic reduction in the effective degrees of freedom (edf) for the estimated smooth functions, suggesting weaker nonlinearity after transformation (Table S2). For example, the edf for gamma-glutamyltransferase (GGT) decreased from 8.18 in the raw trait to 3.44 after IRNT, while the score test p-value increased from effectively zero to 1.87×10^{-5}, indicating a substantial attenuation of the nonlinear signal. Similar patterns were observed

for fasting glucose (FPG) (edf: 7.61 to 2.24, p-values: 0 to 7.92×10^{-3}) and glycated hemoglobin (HBA1C) (edf: 7.78 to 6.32, p-values: 0 to 6.49×10^{-38}). These comparisons highlight that data transformation can significantly alter the apparent degree of nonlinearity, potentially obscuring interpretable biological signals in some cases.

We compared the PVE between linear and nonlinear PRS models to quantify how nonlinear modeling improves predictive performance. Table 1 summarizes the results for traits with PVE gains greater than 1%, while Table S3 presents the full results for raw traits and Table S4 for IRNT traits. Across all 45 traits, the median gain in PVE under raw modeling was 0.34%, with the largest improvement observed for TBL (21.6%). Other traits with notable increases include CRP (12.4%) and gamma-glutamyl transferase (GGT) (6.3%). Consistent with the score tests of linearity, the analysis-of-variance (ANOVA) (i.e., F or likelihood-ratio tests) supports the presence of nonlinear effects, even though the predictive gain appears modest. Given the large sample size of the UK Biobank, even subtle nonlinear patterns can lead to statistically significant results. Moreover, when we repeated the comparison under IRNT outcomes to reduce the influence of outliers and skewness, we found that the nonlinear signal became much weaker, with a median of gain in PVE being 0.03%. Only TBL and vitamin D (VTD) retained PVE gains above 1% after transformation, suggesting that for most traits, deviations from linearity are sensitive to distributional features. This underscores the importance of modeling scale and transformation when assessing nonlinear genetic effects.

2.4 Causal Inference of Retirement on Health and Lifestyle

Retirement is a significant life transition that has been linked to a range of health outcomes in older adults. For instance, a meta-analysis by Odone et al. [34] found that retirement is associated with nearly a 20% reduction in the risk of depression. In contrast, Furuya and Fletcher [35] applied an RDD to UK Biobank data and reported that retirement accelerates biological aging. Motivated by these contrasting findings, we conducted both RDD and RKD analyses to investigate whether retirement at age 65 in the UK is associated with (1) discontinuous jumps or (2) changes in the slope of health and lifestyle trajectories. These two structures represent alternative forms of causal effects under the regression discontinuity framework.

Specifically, focusing on individuals aged 60 to 70, we used age 65, the statutory retirement age in the UK, as the cutoff. Also, our analysis is based on the genetically inferred European population in the UK Biobank. The sample sizes range from 102,288 (vigorous physical activity time) to 183,216 (driving time) with a median of 174,370. Our analysis expanded upon previous work by including 37 continuous health-related variables and 24 additional lifestyle-related traits. All traits were normalized using IRNT. To identify potential causal effects of retirement on health and lifestyle traits, we applied both RDD (jump) and RKD (slope change) analyses centered at age = 65, the statutory retirement age in the UK. Importantly, we only interpreted these effects as causal when the underlying parametric structure, either a linear discontinuity or a piecewise linear form, was deemed adequate by TAPS. In other words, two criteria had to be satisfied: (1) the score test for nonlinearity was not significant ($p > 0.05$), indicating that a simplified parametric form sufficiently captured the age effect; and (2)

Table 1: Comparison of linear and non-linear PRS models. PVE (proportion of variance explained) is defined as the R^2 from `mgcv` models (adjusted for logistic regression where appropriate). Models labeled "Linear" assume additive PRS effects, while "Non-Linear" models allow for flexibility. Gain is calculated as $(\text{PVE}_{\text{NonLin}} - \text{PVE}_{\text{Linear}})/\text{PVE}_{\text{Linear}}$. Only traits with Gain > 1% are shown. Statistics correspond to ANOVA comparing the two models (F-statistics for continuous outcomes, χ^2-statistics for binary outcomes).

Trait	Type	Transformation	PVE Linear	PVE Non-Linear	Gain	Statistics	P-Value
Alanine Aminotransferase	Continuous	Raw	20.90%	21.47%	2.73%	360.07	0
Body Mass Index	Continuous	Raw	32.69%	33.21%	1.59%	385.34	0
C-reactive Protein	Continuous	Raw	11.06%	12.43%	12.39%	669.28	0
Fasting Plasma Glucose	Continuous	Raw	9.61%	9.98%	3.85%	185.88	2.65E-305
Gamma-Glutamyl Transferase	Continuous	Raw	19.28%	20.50%	6.33%	698.48	0
Heart Failure	Binary	Raw	8.68%	8.79%	1.27%	121.94	1.62E-24
Hemoglobin A1c	Continuous	Raw	26.35%	26.83%	1.82%	321.65	0
Neutrophil Count	Continuous	Raw	21.98%	22.22%	1.09%	167.41	1.08E-255
Peripheral Artery Disease	Binary	Raw	4.43%	4.65%	4.97%	228.77	3.25E-46
Smoking Initiation	Binary	Raw	4.77%	4.84%	1.47%	201.11	1.93E-40
Total Bilirubin	Continuous	Raw	39.10%	47.61%	21.76%	6732.64	0
Total Bilirubin	Continuous	IRNT	41.90%	44.17%	5.42%	1739.57	0
Triglycerides	Continuous	Raw	24.94%	25.61%	2.69%	466.56	0
Vitamin D	Continuous	Raw	14.84%	15.26%	2.83%	249.14	0
Vitamin D	Continuous	IRNT	16.43%	16.75%	1.95%	193.72	1.46E-305

the jump or slope change test was significant at an FDR-adjusted threshold ($q < 0.1$). Details of analysis are shown in **Method** and **Supplemental Materials**.

We systematically tested for jumps and slope changes across 61 traits (Table S5-S6). Based on the aforementioned filtering criteria, 2 traits exhibited significant jumps at age 65 (Fig. 4A), while 13 traits showed significant slope changes (Fig. 4B). Among those with slope changes, the majority demonstrated a decrease in slope following retirement, suggesting a potential deceleration in the rate of health deterioration or lifestyle change. Notably, cystatin C (Cys C) was the only trait showing a significant increase in slope, indicating a possible acceleration in kidney function decline. We also observed post-retirement declines in lipid-related traits, including apolipoprotein A (APOA), apolipoprotein B (APOB), HDL-C, total cholesterol (TC), and triglycerides (TG). Fig. 4C presents representative traits that passed both TAPS structural adequacy and causal significance criteria, with panels colored green for linear discontinuities and yellow for piecewise linear effects. Furthermore, we compared the 60–70 age profiles to full-age trajectories (Fig. S9) to better contextualize these changes. For lipid traits, the observed slope shifts near retirement appear to be embedded within broader age-related declines, suggesting that retirement may act as a turning point that accelerates ongoing metabolic changes. In contrast, the increase in Cys C slope at age 65 stands out against an otherwise gradual trajectory, indicating a localized inflection point in kidney function decline. Overall, these findings suggest that while many traits remain unaffected, a distinct subset shows retirement-related changes consistent with interpretable causal patterns.

3 Discussion

In this paper, we propose TAPS, a new framework for estimating and evaluating whether a target function in a GAM can be fully described by a parametric structure. Existing hypothesis tests are typically limited to testing whether a smooth function is zero, rather than whether it follows a scientifically meaningful parametric form. TAPS addresses this gap by formulating the null hypothesis as the function belonging to a given parametric structure, and the alternative as a general smooth function. This formulation allows us to directly answer the question: does the function follow the specified structure? Methodologically, TAPS ensures model identifiability by constructing an orthogonal smooth deviation component, and provides both Wald and score tests for hypothesis testing. The Wald test can be applied to all families of outcomes which `mgcv` supports, while the score test generally offers greater statistical power. In addition, we develop the R package `mgcv.taps`, which is tightly integrated with `mgcv`, one of the most widely used GAM software, allowing users to apply TAPS with minimal additional learning, supports a wide range of outcome types, and scales efficiently to biobank-scale data.

Our two real data analyses also provide novel insights. First, the linearity test of PRSs across 48 traits revealed that the null hypothesis of linearity was rejected for nearly all traits, while the improvement in predictive performance from nonlinear modeling was modest for most traits (median of gain in PVE = 0.34%). These results confirm that for highly polygenic traits such as height, linear additive models remain

sufficient: modeling height nonlinearly yielded a negligible gain in PVE (0.03%), consistent with prior findings that linear models suffice for many common traits [14]. However, some traits, including TBL and CRP, showed clear departures from linearity, with PVE gains exceeding 10%, suggesting that linear models were inadequate for these outcomes. Moreover, we found that the presence and impact of nonlinear effects were influenced by data transformation. After applying IRNT, the score test continued to detect significant nonlinear patterns for most traits (median p-values 5E-12); however, the gain in PVE from nonlinear modeling substantially decreased, with a median improvement of only 0.03%. This underscores that the predictive contribution of nonlinear PRSs is often limited, especially after transforming traits to reduce skewness and outliers. Our findings align with Albiñana et al. [36], who reported that improvements from non-linear models (e.g., XGBoost [37]) primarily stem from capturing non-linear interactions between covariates, such as sex and age.

Second, our results revealed that nearly 25% of traits (15 over 61) showed significant changes in post-retirement slopes, indicating that retirement is associated with altered long-term health trajectories for a substantial subset of physiological indicators. These include accelerated increases in Cys C, along with steeper declines in height, hemoglobin, and key metabolic markers such as APOA and HDL-C. Such patterns suggest possible deterioration in renal, hematologic, and cardiometabolic health after retirement. Our findings broadly support the view that retirement may, in certain contexts, contribute to physiological decline. This aligns with prior literature: for example, Chen et al. [38] used an instrumental variable (IV) approach to show improvements in mental health post-retirement, but also increased risks of lifestyle-related diseases among men. Nonetheless, our analysis may suffer from unmeasured confounding. For instance, our sample is limited to White British participants, who generally worked in less physically demanding environments. Effects of retirement may be found to be positive to health in populations with harsher work conditions or weaker safety nets. Future research should incorporate more diverse cohorts and institutional settings, and apply stronger causal inference tools, such as combining the IV approach with RDD or RKD [39], to better isolate retirement's health effects and assess their generalizability across labor markets and social contexts.

Several limitations of TAPS warrant further investigation. First, TAPS requires users to specify the parametric structure to be tested, including hyperparameters such as the location of a jump or slope change. In applications like RDD and RKD, these locations correspond to cutoffs of treatment assignments and are not subject to tuning [16]. However, identifying appropriate structures remains challenging, and data-driven methods for selecting or validating structural forms, such as the number and position of breakpoints, are still underdeveloped [40, 41]. Second, the Wald test relies on numerical approximations and is consistently less powerful than the score test, especially when the smooth term has low degrees of freedom. Whereas, the score test currently applies only to outcomes from exponential family distributions. For other types, such as survival and ordinal outcomes, only the Wald test is available. We therefore recommend using the score test as the first choice when applicable, and applying the Wald test as a practical alternative when the score test is unavailable. In addition, our PRS analysis did not incorporate GxG or GxE interactions, which

Kelemen et al. [30] found to contribute 7% and 4% gains in PVE, respectively. In other words, our findings suggest that PRS estimated by SBayesRC (and potentially other existing PRS methods) may capture non-linear effects, but the improvement in prediction is limited. Incorporating GxG and GxE interactions into our model warrants future investigation.

4 Method

4.1 Construction of Basis Functions

Recall the new mixed-effect representation (3):

$$f(x) = \boldsymbol{A}(x)^\top \boldsymbol{\alpha} + b(x),$$

where $\boldsymbol{A}(x) = (A_1(x), \ldots, A_q(x))^\top$ is a vector of the basis functions of the parametric structure, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^\top$ is a fixed vector, $b(x)$ is a Gaussian process with covariance function $\mathrm{cov}(b(x_{i1}), b(x_{j1})) = \phi K(x_i, x_j)$, $K(x_i, x_j)$ is a known bivariate function, and $\phi \geq 0$ is the variance. In particular, $\boldsymbol{A}(x)^\top \boldsymbol{\alpha}$ is orthogonal to $b(x)$, which guarantees the identifiability of (3). Note that all covariates involved in smooth terms are assumed to be continuous or approximately continuous. Variables with only a few distinct values (e.g., categorical or highly discretized variables) are not suitable, as smooth function estimation becomes unstable or ill-defined in such settings.

The construction of $b(x)$ starts from representing it using the RKHS [42], under the alternative hypothesis that $b(x)$ is a general smooth function. According to the functional analysis theory, every RKHS $\mathcal{H}_K$ corresponds to a unique semi-positive definite kernel function $K(\cdot, \cdot)$ defined in the Cartesian product space $[0, 1] \times [0, 1]$. For any function $b \in \mathcal{H}_K$,

$$\langle b, K(\cdot, x) \rangle_{\mathcal{H}_K} = b(x), \tag{6}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$ is the inner product in the space $\mathcal{H}_K$. This property ensures that for n observation $\{x_1, \ldots, x_n\} \subset [0, 1]$, there must exist an $(n \times n)$ matrix $\mathbf{K}$ with the (i, j)th element being $K(x_i, x_j)$ and an $(n \times 1)$ vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ such that

$$\boldsymbol{b} = \mathbf{K}\boldsymbol{\theta}, \tag{7}$$

where $\boldsymbol{b} = (b(x_1), \ldots, b(x_n))^\top$.

We construct the basis functions of $\mathcal{H}_B = \mathcal{H}_K / \mathcal{H}_A$ by finding an $((n + q) \times n)$ orthogonal matrix $\mathbf{Q}$ satisfying the following linear constraint

$$\mathbf{A}^\top \mathbf{F}\mathbf{Q} = \mathbf{0}_{q \times n}, \tag{8}$$

where $\mathbf{F} = (\mathbf{A}, \mathbf{K})$ is the combination of $\mathbf{A}$ and $\mathbf{K}$. Here, the orthogonal matrix $\mathbf{Q}$ describes a special linear transformation that resorts the columns of $\mathbf{F}$, i.e., the basic functions of $\mathcal{H}$, into a new one perpendicular to $\mathbf{A}$. As a result, the columns of $\mathbf{FQ}$ can be regarded as the basis functions of $\mathcal{H}_B$ and thus it can represent $\boldsymbol{b} = \mathbf{FQ}\boldsymbol{\pi}$ where

$\boldsymbol{\pi}$ is an unconstrained vector in $\mathbb{R}^n$. The QR decomposition can yield a candidate of the orthogonal matrix $\mathbf{Q}$. Consider the following QR decomposition:

$$\mathbf{F}^\top \mathbf{A} = (\mathbf{Q}_1, \mathbf{Q}_2) \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{0}_{n\times q} \end{pmatrix}, \tag{9}$$

where $\mathbf{Q}_1$ is an $((n+q)\times q)$ orthogonal matrix, $\mathbf{Q}_2$ is an $((n+q)\times n)$ orthogonal matrix, and $\mathbf{R}_1$ is an $(q\times q)$ upper triangular matrix. Here $\mathbf{Q}_1$ is an orthogonal basis matrix of $\text{span}(\mathbf{F}^\top\mathbf{A})$ and $\mathbf{Q}_2$ is an orthogonal basis matrix of $\overline{\text{span}}(\mathbf{F}^\top\mathbf{A})$, thus we can choose $\mathbf{Q} = \mathbf{Q}_2$.

In practice, most of the commonly used kernel functions can generate a general smooth function, such as the Gaussian kernel, spherical kernel, the Matérn family correlation function, and the kernels of the Sobolev space [27]. In addition, the expression (7) also holds for multivariate function $\boldsymbol{b} = (b(\boldsymbol{x}_1), \ldots, b(\boldsymbol{x}_n))^\top$ where $\boldsymbol{x}_i$ is a multivariate vector. In this case, the kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is usually chosen as the isotropic kernel function, i.e., $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = K(||\boldsymbol{x}_i - \boldsymbol{x}_j||)$ with a certain norm $||\cdot||$.

4.2 Prior Distribution

Mathematically, there is a well-known equivalence between a smooth function in a RKHS induced by a kernel function $K(\cdot,\cdot)$, and a Gaussian process defined on an L_2 space with covariance function K. Leveraging this equivalence, for a nonparametric regression model $y_i = b(x_i) + \epsilon_i$, one typically estimates the function $b \in \mathcal{H}_K$ by solving the following penalized least squares problem:

$$\hat{\boldsymbol{b}} = \arg\min_{\boldsymbol{b}} \left\{ \|\boldsymbol{y} - \boldsymbol{b}\|_2^2 + \lambda_n \boldsymbol{b}^\top \mathbf{K}^+ \boldsymbol{b} \right\}, \tag{10}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{b} = (b(x_1), \ldots, b(x_n))^\top$, $\hat{\boldsymbol{b}} = (\hat{b}(x_1), \ldots, \hat{b}(x_n))^\top$, $\mathbf{K}$ is the $(n\times n)$ kernel matrix with entries $K(x_i, x_j)$, and $\mathbf{K}^+$ denotes its Moore–Penrose generalized inverse. This RKHS-based formulation is statistically equivalent to placing a Gaussian process prior on $b(x)$, with zero mean and covariance function $\text{cov}(b(x_i), b(x_j)) = \sigma^2 \lambda_n^{-1} K(x_i, x_j)$, where $\sigma^2 = \text{var}(\epsilon_i)$. However, in our mixed-effects representation (3), the smooth component $b(x)$ is not a free element of $\mathcal{H}_K$, but is constrained to be orthogonal to the structured term $a(x) = \boldsymbol{A}(x)^\top \boldsymbol{\alpha}$. As a result, deriving the implied Gaussian process prior for $b(x)$ is nontrivial, since the prior must account for the orthogonality constraint imposed by the model.

The prior distribution of $b(x)$ can be constructed as follows. Specifically, since $\boldsymbol{b}$ also falls within $\mathcal{H} = \mathcal{H}_A \cup \mathcal{H}_K$, it can be alternatively represented as

$$\boldsymbol{b} = \mathbf{F}\tilde{\boldsymbol{\vartheta}} = \mathbf{A}\tilde{\boldsymbol{\alpha}} + \mathbf{K}\tilde{\boldsymbol{\theta}}, \tag{11}$$

where the vector $\tilde{\boldsymbol{\vartheta}} = (\tilde{\boldsymbol{\alpha}}^\top, \tilde{\boldsymbol{\theta}}^\top)^\top$. The constraint to ensure $\boldsymbol{b}$ fully fall within $\mathcal{H}_B$ is

$$\mathbf{A}^\top \mathbf{F} \tilde{\boldsymbol{\vartheta}} = \mathbf{0}_{q\times 1}. \tag{12}$$

This constraint is similar to the boundary condition of the polynomial smoothing spline [42, Example 12.29], but it has a clearer meaning: the random effect $\boldsymbol{b}$ represented by $\mathbf{F}\tilde{\boldsymbol{\vartheta}}$ is perpendicular to the fixed parametric effect $\mathbf{A}\boldsymbol{\alpha}$.

Based on this random effect representation of smooth functions in RKHS, the augmented vector $\tilde{\boldsymbol{\vartheta}}$ should follow $\mathcal{N}(\mathbf{0}_{(n+q)\times 1}, \phi\tilde{\mathbf{K}}^{+})$ where $\tilde{\mathbf{K}} = \mathrm{diag}(\mathbf{0}_{q\times q}, \mathbf{K})$ is an $((n+q)\times(n+q))$ block-diagonal matrix. The block corresponding to $\tilde{\boldsymbol{\alpha}}$ is a zero matrix because $\mathbf{A}\tilde{\boldsymbol{\alpha}}$ is a fixed effect and hence this is no randomness of $\tilde{\boldsymbol{\alpha}}$. Furthermore, since $\tilde{\boldsymbol{\vartheta}} = \mathbf{Q}\boldsymbol{\pi}$, the prior distribution of $\boldsymbol{\pi}$ is then

$$\boldsymbol{\pi} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \phi(\mathbf{Q}^{\top}\tilde{\mathbf{K}}\mathbf{Q})^{+}). \tag{13}$$

Subsequently, since $\boldsymbol{b} = \mathbf{F}\mathbf{Q}\boldsymbol{\pi}$, the prior distribution of $\boldsymbol{b}$ is

$$\boldsymbol{b} \sim \mathcal{N}(\mathbf{0}_{n\times 1}, \phi\mathbf{F}\mathbf{Q}(\mathbf{Q}^{\top}\tilde{\mathbf{K}}\mathbf{Q})^{+}\mathbf{Q}^{\top}\mathbf{F}^{\top}). \tag{14}$$

Thus, testing whether $\boldsymbol{f} = \mathbf{A}\boldsymbol{\alpha}$ reduces to testing whether $\phi = 0$.

4.3 Dimension Reduction

The construction of $\boldsymbol{b}$ requires the QR decomposition of the $((n+q)\times q)$ matrix $\mathbf{F}^{\top}\mathbf{A}$, which can be computationally expensive when n is large. In addition, the dimension of the vector $\boldsymbol{\pi}$ is n, which may lead to the curse of dimensionality in practice. To address these numerical challenges, we adopt two dimension-reduction techniques to simplify the representation of $\boldsymbol{b}$. First, we use a low-rank RKHS approximation [43] to express the smooth component. Second, we apply principal component analysis (PCA) for further dimension reduction, following the strategy used by Wood [4].

4.4 Score Test

Recall the GAM (1):

$$y_i \sim \mathcal{Y}(\mu_i, \phi_0), \quad g(\mu_i) = \eta_i = \boldsymbol{Z}_i^{\top}\boldsymbol{\gamma} + f_1(x_{i1}) + \cdots + f_J(x_{iJ}),$$

where y_i is a response variable belonging to a certain distribution $\mathcal{Y}(\mu_i, \phi_0)$ with mean μ_i and dispersion parameter ϕ_0, g is a known link function, η_i is a linear predictor, $\boldsymbol{Z}_i$ is the ith row of a parametric model matrix $\mathbf{Z}$, $\boldsymbol{\gamma}$ is a regression coefficient vector, and f_j is a smooth function of covariate x_{ij}.

We extend the score test proposed by Zhang and Lin [18] to evaluate whether the smooth deviation term $b(x)$ is equal to zero. The core idea of this test is based on the equivalence between GAM and GLMM. Specifically, the score test is constructed from the following GLMM model:

$$\boldsymbol{Y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{15}$$

where $\mathbf{A}$ is the design matrix for fixed effects $\boldsymbol{\alpha}$, and $\mathbf{B}$ is the design matrix for random effects $\boldsymbol{\beta}$. The noise term follows a Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \phi_0\mathbf{W}^{-1})$, where

ϕ_0 is the dispersion parameter, $\mathbf{W} = \mathrm{diag}(W_1, \ldots, W_n)$ and $W_i = 1/(V(\mu_i)g'(\mu_i)^2)$, with $V(\mu)$ denoting the variance function and g the canonical link function. For the GAM (1), $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_J, \mathbf{Z})$, $\mathbf{B} = (\mathbf{F}_1\mathbf{Q}_1, \ldots, \mathbf{F}_J\mathbf{Q}_J)$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_J^\top, \boldsymbol{\gamma}^\top)^\top$, and $\boldsymbol{\beta} = (\boldsymbol{\pi}_1^\top, \ldots, \boldsymbol{\pi}_J^\top)^\top$. Since both $\mathbf{B}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are random, it is reasonable to consider them as a new noise $\boldsymbol{\varepsilon} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, whose prior distribution is $\mathcal{N}(\mathbf{0}_{n\times 1}, \mathbf{V}(\boldsymbol{\phi}))$ with $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_J)^\top$ and

$$\mathbf{V}(\boldsymbol{\phi}) = \phi_0\mathbf{W}^{-1} + \sum_{j=1}^{J} \phi_j \mathbf{F}_1\mathbf{Q}_1\mathbf{S}_j\mathbf{Q}_1^\top\mathbf{F}_1^\top. \tag{16}$$

The restricted maximum likelihood (REML) function of this GLMM is:

$$\begin{aligned}\mathcal{L}_{\mathrm{REML}}(\boldsymbol{\phi}|\widehat{\boldsymbol{\alpha}}) = &-\frac{1}{2}(\boldsymbol{Y} - \mathbf{A}\boldsymbol{\alpha})^\top\mathbf{V}(\boldsymbol{\phi})^{-1}(\boldsymbol{Y} - \mathbf{A}\boldsymbol{\alpha}) \\ &- \frac{1}{2}\log\det(\mathbf{V}(\boldsymbol{\phi})) - \frac{1}{2}\log\det\{\mathbf{A}^\top\mathbf{V}(\boldsymbol{\phi})^{-1}\mathbf{A}\},\end{aligned} \tag{17}$$

The score test is based on the score equation $\mathbf{U}(\cdot)$ and Hessian matrix $\mathbf{H}(\cdot,\cdot)$ of the REML function (17) below:

$$\begin{aligned}\mathbf{U}(\phi_j) &= \left.\frac{\partial\mathcal{L}_{\mathrm{REML}}(\boldsymbol{\phi}|\widehat{\boldsymbol{\alpha}})}{\partial\phi_j}\right|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j} \\ &= \underbrace{\frac{1}{2}(\widehat{\boldsymbol{Y}} - \mathbf{A}\widehat{\boldsymbol{\alpha}})^\top\mathbf{V}(\boldsymbol{\phi})^{-1}\mathbf{G}_j\mathbf{V}(\boldsymbol{\phi})^{-1}(\widehat{\boldsymbol{Y}} - \mathbf{A}\widehat{\boldsymbol{\alpha}})\}}_{\mathbf{u}(\phi_j)} - \underbrace{\frac{1}{2}\mathrm{tr}(\mathbf{P}(\boldsymbol{\phi})\mathbf{G}_j)}_{\mathbf{e}(\phi_j)}\Bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j},\end{aligned} \tag{18}$$

$$\mathbf{H}(\phi_j, \phi_k) = \mathrm{E}\left\{\frac{-\partial^2\mathcal{L}_{\mathrm{REML}}(\boldsymbol{\phi}|\widehat{\boldsymbol{\alpha}})}{\partial\phi_j\partial\phi_k}\right\}\Bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j} = \frac{1}{2}\mathrm{tr}(\mathbf{P}(\boldsymbol{\phi})\mathbf{G}_j\mathbf{P}(\boldsymbol{\phi})\mathbf{G}_k)\Bigg|_{\boldsymbol{\phi}=\widehat{\boldsymbol{\phi}}_0^j}, \tag{19}$$

where $\widehat{\boldsymbol{\phi}}_0^j = (\hat{\phi}_1, \ldots, \hat{\phi}_{j-1}, 0, \hat{\phi}_{j+1}, \ldots, \hat{\phi}_J)^\top$, $\mathbf{G}_j = \partial\mathbf{V}(\boldsymbol{\phi})/\partial\phi_j = \mathbf{F}_1\mathbf{Q}_1\mathbf{S}_j\mathbf{Q}_1^\top\mathbf{F}_1^\top$, and

$$\mathbf{P}(\hat{\boldsymbol{\phi}}) = \mathbf{V}(\hat{\boldsymbol{\phi}})^{-1} - \mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}\mathbf{A}(\mathbf{A}^\top\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}. \tag{20}$$

Zhang and Lin [18] used the moment matching method to show that $\mathbf{u}(\phi_1)$ in (19) asymptotically follows a scaled χ^2 distribution:

$$\mathbf{u}(\phi_1) \sim \kappa\chi_\nu^2, \tag{21}$$

where $\kappa = \mathbf{H}(\phi_1, \phi_1)/(2\mathbf{e}(\phi_1))$ and $\nu = 2\mathbf{e}(\phi_1)^2/\mathbf{H}(\phi_1, \phi_1)$. They thus used $\mathbf{u}(\phi_1)$ as the testing statistics and yielded the p-value by $p(\phi_1) = 1 - F_\chi(\mathbf{u}(\phi_1)/\kappa, \nu)$, where $F_\chi(\cdot, \nu)$ is the cumulative distribution function of the χ_ν^2 distribution.

When the sample size n is large, direct construction of the $n \times n$ precision matrix $\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}$ and the $n\times n$ projection matrix $\mathbf{P}(\hat{\boldsymbol{\phi}})$ is computationally infeasible. To address this, we develop an operator-based strategy to compute the score statistic $\mathbf{u}(\phi_1)$, its expectation $\mathbf{e}(\phi_1)$, and its variance component $\mathbf{H}(\phi_1, \phi_1)$, without explicitly forming

these matrices. In particular, we construct matrix-vector product operators that evaluate the action of $\mathbf{V}(\hat{\boldsymbol{\phi}})^{-1}$ and $\mathbf{P}(\hat{\boldsymbol{\phi}})$ on arbitrary vectors, and use them to implement both residual projection and trace computation efficiently. For the trace terms, we exploit the low-rank structure of the form $\mathbf{G}_j = \mathbf{F}_1\mathbf{Q}_1\mathbf{S}_j\mathbf{Q}_1^\top\mathbf{F}_1^\top$. These computational tricks make the score test scalable and memory-efficient for large datasets.

4.5 Wald Test

We adopt the Wald test proposed by Wood [26] to evaluate whether the smooth deviation term $b(x)$ is equal to zero. Specifically, let $\hat{\boldsymbol{\pi}}$ denote the estimate of $\boldsymbol{\pi}$ under the estimated variance component $\hat{\phi}$. The null hypothesis is H_0: $b(x) = 0$. The Wald test statistic is given by $\hat{\boldsymbol{b}}^\top\mathbf{S}_b^{-1}\hat{\boldsymbol{b}}$, where $\hat{\boldsymbol{b}} = \mathbf{F}\mathbf{Q}\hat{\boldsymbol{\pi}}$ and $\mathbf{S}_b = \hat{\phi}\mathbf{F}\mathbf{Q}(\mathbf{Q}^\top\tilde{\mathbf{K}}\mathbf{Q})^+\mathbf{Q}^\top\mathbf{F}^\top$. Although this statistic is theoretically assumed to follow a χ^2 distribution, Wood [26] showed that it is in fact a mixture of χ^2 distributions due to the penalization imposed on $b(x)$, which renders its effective degrees of freedom non-integer. To account for this, a numerical algorithm was developed to accurately approximate the null distribution of the test statistic.

4.6 Regression Discontinuity Design

RDD is a quasi-experimental evaluation method that identifies the causal effects of interventions by assigning a cutoff above or below which an intervention is assigned [16]. The cutoff is indeed a kind of structural breakpoint that reflects the "just-barely-treated" against the "just-barely-didn't" to get the effect of treatment. Suppose $f(x)$ is a general discontinuous function of covariate x that has a breakpoint ν_0. The causal effect of $f(x)$ is a local concept, which is defined as the jump of $f(x)$ at ν_0:

$$\theta := \lim_{h\downarrow 0} f(\nu_0 + h) - \lim_{h\uparrow 0} f(\nu_0 - h), \tag{22}$$

where h is a small bandwidth of the local neighborhood. This expression defines the causal effect as the size of the discontinuous jump at the cutoff of the running variable, corresponding to the difference in outcomes between units just above and just below the threshold where treatment is assigned [16].

The linearity discontinuity is the most used structure in RDD, where the function is represented as:

$$f(x) = \alpha_0 + \alpha_1(x - \nu_0) + \alpha_2\mathrm{I}(x > \nu_0) + \alpha_3\{(x - \nu_0)\mathrm{I}(x > \nu_0)\}. \tag{23}$$

Under this parametric form, the jump is explicitly modeled by the coefficient α_2, which captures the immediate shift in the outcome at the cutoff. The remaining terms control for linear trends on either side of the threshold, allowing for different slopes before and after treatment assignment.

4.7 Regression Kink Design

RKD is a quasi-experimental method that identifies causal effects by exploiting known kinks of a treatment assignment rule [16]. Unlike RDD, which focuses on a discontinuous jump at a cutoff, RKD leverages a structural change in the slope of a response variable for a running variable. Suppose the function $f(x)$ is continuous at a known kink point ν_1, but its derivative changes. Then the causal effect is defined as the difference in the slopes immediately to the right and left of ν_1:

$$\theta := \lim_{h \downarrow 0} f'(\nu_1 + h) - \lim_{h \uparrow 0} f'(\nu_1 - h). \tag{24}$$

This expression defines the causal effect as the change in slope (i.e., the derivative) of the outcome variable at the kink point, capturing how the treatment intensity or incentive structure shifts at a known threshold in the running variable [16].

The most commonly used specification is a piecewise linear model with a kink at ν_1, expressed as:

$$f(x) = \alpha_0 + \alpha_1 x + \alpha_2 (x - \nu_1)_+. \tag{25}$$

In this parametric form, the kink is modeled by the term $(x - \nu_1)_+$, and the causal effect $\theta = \alpha_2$ represents the difference in slopes before and after the kink, while α_1 accounts for the baseline trend.

4.8 Linearity Test of PRS

For continuous traits, we consider the following model:

$$\text{Trait}_i = f_1(\text{Trait PRS}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma}) + \epsilon_i, \tag{26}$$

and for binary traits, we adopt the following generalized additive model:

$$\Pr(\text{Trait}_i = 1) = \frac{\exp(f_1(\text{Trait PRS}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma})}{1 + \exp(f_1(\text{Trait PRS}_i) + f_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma})}, \tag{27}$$

where $\boldsymbol{Z}_i = (1, \text{Sex}_i, \text{PC}_{i1}, \ldots, \text{PC}_{i10})^\top$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{12})^\top$. For the term $f_1(\text{Trait PRS})$, we use the mixed-effect representation (3) with a linear parametric component to test whether the PRS effect can be sufficiently modeled as linear. For computational efficiency, the age effect $f_2(\text{Age})$ is modeled using cubic regression splines. Smoothing parameters were estimated via REML as implemented in `mgcv`. Note that for hypertension (HTN), we consider specifying the bivariate function of SBP PRS and DBP PRS, where the parametric structure is the linear interaction between them, as SBP, DBP, and medication directly define HTN.

To compute the proportion of variance explained (PVE), we first fit a fully parametric linear model for $f_1(\text{Trait PRS})$ by replacing the smooth term with a linear effect.

For continuous outcomes, PVE is defined as the standard coefficient of determination:

$$R^2_{\text{continuous}} = \frac{\text{var}(\hat{\epsilon}_i)}{\text{var}(\text{Trait}_i)}, \tag{28}$$

where $\hat{\epsilon}$ denotes the residuals from the linear model. For binary outcomes, we report the pseudo R^2 on the logit liability scale [44], which approximates the variance explained on an underlying continuous liability. Specifically, the transformation is:

$$R^2_{\text{binary}} = \frac{\text{var}(\hat{f}_1(\text{Trait PRS}_i) + \hat{f}_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}})}{\text{var}(\hat{f}_1(\text{Trait PRS}_i) + \hat{f}_2(\text{Age}_i) + \boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}}) + 3.29}. \tag{29}$$

This formulation accounts for the inherent noise in binary traits and allows more interpretable comparisons across traits with different prevalences.

To reduce the influence of outliers and skewed distributions, we applied the following IRNT to all continuous outcomes:

$$\text{IRNT}(y_i) = \Phi^{-1}(n^{-1} r_i - 0.5), \tag{30}$$

where r_i is the rank of y_i among all y values in ascending order, and $\Phi^{-1}(\cdot)$ denotes the quantile function of the standard normal distribution. This transformation ensures that the transformed variable approximately follows a standard normal distribution and improves the robustness and comparability of model fitting. Moreover, to mitigate the influence of outliers and skewed distributions in the raw trait analyses, we excluded individuals falling in the top and bottom 1% of the trait distribution.

4.9 Regression Discontinuity and Kink Designs of Retirement

We applied RDD and RKD, using age = 65, the statutory retirement age in the UK, as the cutoff. For each continuous outcome, we modeled the trait as a function of age using the following GAM:

$$\text{Trait}_i = f_1(\text{Age}_i) + f_2(\text{Trait PRS}_i) + \boldsymbol{Z}_i^\top \boldsymbol{\gamma} + \epsilon_i \tag{31}$$

where $\boldsymbol{Z}_i = (1, \text{Sex}_i, \text{Income}_i, \text{PC}_{i1}, \ldots, \text{PC}_{i10})^\top$, and $\boldsymbol{\gamma}$ denotes the corresponding coefficients. We further included income in the model, given the documented association between socioeconomic status and retirement outcomes. Smoothing parameters were estimated via REML as implemented in `mgcv`. For moderate physical activity, vigorous physical activity, walking time, fortified food intake, spirits intake, and water intake, we did not specify their PRSs in the model because SBayesRC reported that their heritabilities estimates were insignificant.

For $f_1(\text{Age})$, we applied the mixed-effect representation (3), where the parametric component corresponds to either a linear discontinuity (for RDD) or a piecewise linear structure with a change in slope (for RKD) at the threshold of age 65. Specifically, for

RDD, the parametric component is:

$$f_1(\text{Age}) = \alpha_0 + \alpha_1 \text{I}(\text{Age} > 65) + \alpha_2(\text{Age} - 65) + \alpha_3(\text{Age} - 65)\text{I}(\text{Age} > 65). \quad (32)$$

For RKD, the parametric component is:

$$f_1(\text{Age}) = \alpha_0 + \alpha_1 \text{Age} + \alpha_2(\text{Age} - 65)_+, \quad (33)$$

where $(x)_+ = \max(x, 0)$. We used the score test and Wald test implemented in TAPS to determine whether these parametric structures adequately capture the age effect and to test for a significant jump (in RDD) or slope change (in RKD) at age 65.

We adopted a two-stage inference strategy to avoid misinterpreting nonlinear patterns as causal effects of retirement. First, we used the TAPS score test to determine whether the age effect could be adequately modeled by a parametric structure (linearity with a jump for RDD, or piecewise linearity for RKD). Only if this test was non-significant (score test $p > 0.05$), indicating no evidence of nonlinearity, did we proceed to the second stage: testing whether the parametric component exhibited a significant jump or slope change at age 65.

Acknowledgments

We would like to thank the editors and the anonymous referees for their very careful reviews and constructive suggestions. The first author thanks Dr. Jianxin Pan for multiple insightful discussions. We also thank Dr. Simon Wood for providing the flexible `mgcv` interface, which facilitates user-defined smooth terms and enables seamless methodological extensions. This work was supported by grants HG011052 and HG011052-03S1 (to X.Z.) from the National Human Genome Research Institute (NHGRI), USA.

Data Availability

All GWAS summary statistics used in this study are publicly available, with their sources listed in Table S1. The individual-level data from the UK Biobank used for replication analyses are available under Application ID: 81097.

Code Availability

The R package `mgcv.taps`, which implements the TAPS method, is available at https://github.com/harryyiheyang/mgcv.taps. All analysis code used in this paper is available at https://github.com/harryyiheyang/mgcv.taps.code.

References

[1] Wood, S. N. *Generalized Additive Models: An Introduction with R* (CRC Press, 2017).

[2] Nelder, J. A. & Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)* **135**, 370–384 (1972).

[3] Hastie, T. J. & Tibshirani, R. J. *Generalized additive models* (CRC Press, 1990).

[4] Wood, S. N. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**, 95–114 (2003).

[5] Eilers, P. H. & Marx, B. D. Flexible smoothing with b-splines and penalties. *Statistical Science* **11**, 89–121 (1996).

[6] Wood, S. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **62**, 1025–1036 (2006).

[7] Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association* **99**, 673–686 (2004).

[8] Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 3–36 (2011).

[9] McKeigue, P. M. *et al.* Relation of incident type 1 diabetes to recent covid-19 infection: cohort study using e-health record linkage in scotland. *Diabetes Care* **46**, 921–928 (2023).

[10] Wood, S. N. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466**, 1102–1104 (2010).

[11] Shang, L., Wu, P. & Zhou, X. Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics. *Nature Communications* **16**, 1059 (2025).

[12] Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature protocols* **15**, 2759–2772 (2020).

[13] Sulc, J., Sjaarda, J. & Kutalik, Z. Polynomial mendelian randomization reveals non-linear causal effects for obesity-related traits. *Human Genetics and Genomics Advances* **3** (2022).

[14] Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS genetics* **4**, e1000008 (2008).

[15] Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**, 1193–1198 (2012).

[16] Huntington-Klein, N. *The effect: An introduction to research design and causality* (Chapman and Hall/CRC, 2021).

[17] Cox, D., Koh, E., Wahba, G. & Yandell, B. S. Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. *The Annals of Statistics* **16**, 113–119 (1988).

[18] Zhang, D. & Lin, X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics* **4**, 57–74 (2003).

[19] Crainiceanu, C., Ruppert, D., Claeskens, G. & Wand, M. P. Exact likelihood ratio tests for penalised splines. *Biometrika* **92**, 91–103 (2005).

[20] Nummi, T., Pan, J., Siren, T. & Liu, K. Testing for cubic smoothing splines under dependent data. *Biometrics* **67**, 871–875 (2011).

[21] Wood, S. N., Pya, N. & Säfken, B. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* **111**, 1548–1563 (2016).

[22] Wood, S. N., Li, Z., Shaddick, G. & Augustin, N. H. Generalized additive models for gigadata: modeling the uk black smoke network daily data. *Journal of the American Statistical Association* **112**, 1199–1210 (2017).

[23] Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

[24] Kanagawa, M., Hennig, P., Sejdinovic, D. & Sriperumbudur, B. K. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582* (2018).

[25] Breslow, N. E. & Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9–25 (1993).

[26] Wood, S. N. On p-values for smooth components of an extended generalized additive model. *Biometrika* **100**, 221–228 (2013).

[27] Ruppert, D., Wand, M. P. & Carroll, R. J. *Semiparametric Regression* (Cambridge University Press, 2003).

[28] Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. & Goude, Y. Fast calibrated additive quantile regression. *Journal of the American Statistical Association* **116**, 1402–1412 (2021).

[29] He, R. *et al.* Delivr: a deep learning approach to iv regression for testing nonlinear causal effects in transcriptome-wide association studies. *Biostatistics* **25**, 468–485 (2024).

[30] Kelemen, M. *et al.* Performance of deep-learning based approaches to improve polygenic scores. *medRxiv* 2024–10 (2024).

[31] Privé, F. *et al.* Portability of 245 polygenic scores when derived from the uk biobank and applied to 9 ancestry groups from the same cohort. *The American Journal of Human Genetics* **109**, 12–23 (2022).

[32] Lab, N. UK Biobank GWAS Round 2 Results (2018). URL http://www.nealelab.is/uk-biobank/. Accessed: 2025-04-06.

[33] Zheng, Z. *et al.* Leveraging functional genomic annotations and genome coverage to improve polygenic prediction of complex traits within and between ancestries. *Nature Genetics* **56**, 767–777 (2024).

[34] Odone, A. *et al.* Does retirement trigger depressive symptoms? a systematic review and meta-analysis. *Epidemiology and Psychiatric Sciences* **30**, e77 (2021).

[35] Furuya, S. & Fletcher, J. M. Retirement makes you old? causal effect of retirement on biological age. *Demography* **61**, 901–931 (2024).

[36] Albiñana, C. *et al.* Multi-pgs enhances polygenic prediction by combining 937 polygenic scores. *Nature communications* **14**, 4702 (2023).

[37] Chen, T. & Guestrin, C. *Xgboost: A scalable tree boosting system*, 785–794 (2016).

[38] Chen, F., Wakabayashi, M. & Yuda, M. The impact of retirement on health: Empirical evidence from the change in public pensionable age in japan. *The Journal of the Economics of Ageing* **28**, 100513 (2024).

[39] Oldenburg, C. E., Moscoe, E. & Bärnighausen, T. Regression discontinuity for causal effect estimation in epidemiology. *Current Epidemiology Reports* **3**, 233–241 (2016).

[40] Muggeo, V. M. Estimating regression models with unknown break-points. *Statistics in Medicine* **22**, 3055–3071 (2003).

[41] Zhang, F. & Li, Q. Segmented correspondence curve regression for quantifying covariate effects on the reproducibility of high-throughput experiments. *Biometrics* **79**, 2272–2285 (2023).

[42] Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press, 2019).

[43] Kammann, E. & Wand, M. P. Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**, 1–18 (2003).

[44] Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genetic Epidemiology* **36**, 214–224

(2012).

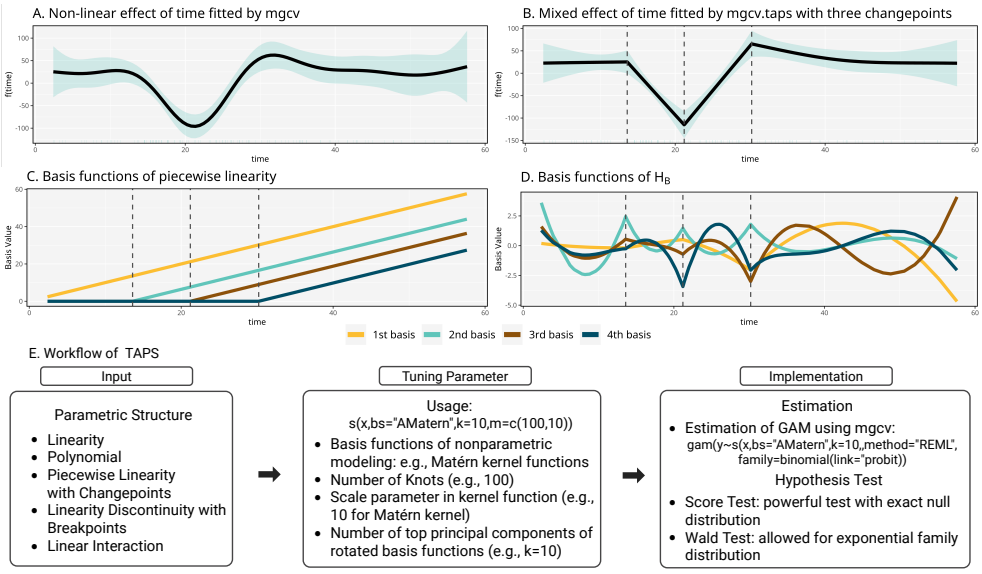


Fig. 1: Overview of the TAPS framework. This figure uses the `mcycle` data, which records head acceleration over time in a simulated motorcycle crash experiment, to illustrate the workflow of TAPS. This data is provided by the R package `MASS` and is commonly used for benchmarking smooth curve fitting methods. **A**. Non-linear effect of time fitted by `mgcv` with default smooth terms. **B**. Fitted smooth using `mgcv.taps` with three user-specified changepoints. **C**. Basis functions of the space of piecewise linearity space $\mathcal{H}_A$. **D**. Basis functions from the kernel-smoothed space $\mathcal{H}_B$, automatically learned via principal component decomposition. **E**. Summary of the TAPS components. Input allows users to specify parametric structure in the adaptive smooth term, such as linearity, polynomial trends, piecewise linearity, or discontinuities. Tuning Parameter panel outlines the usage of the adaptive smooth term with key options, including the number of knots and kernel parameters. Implementation involves model fitting using `mgcv` and hypothesis testing via either the score test or the Wald test. For implementation details and source code, please visit the GitHub.

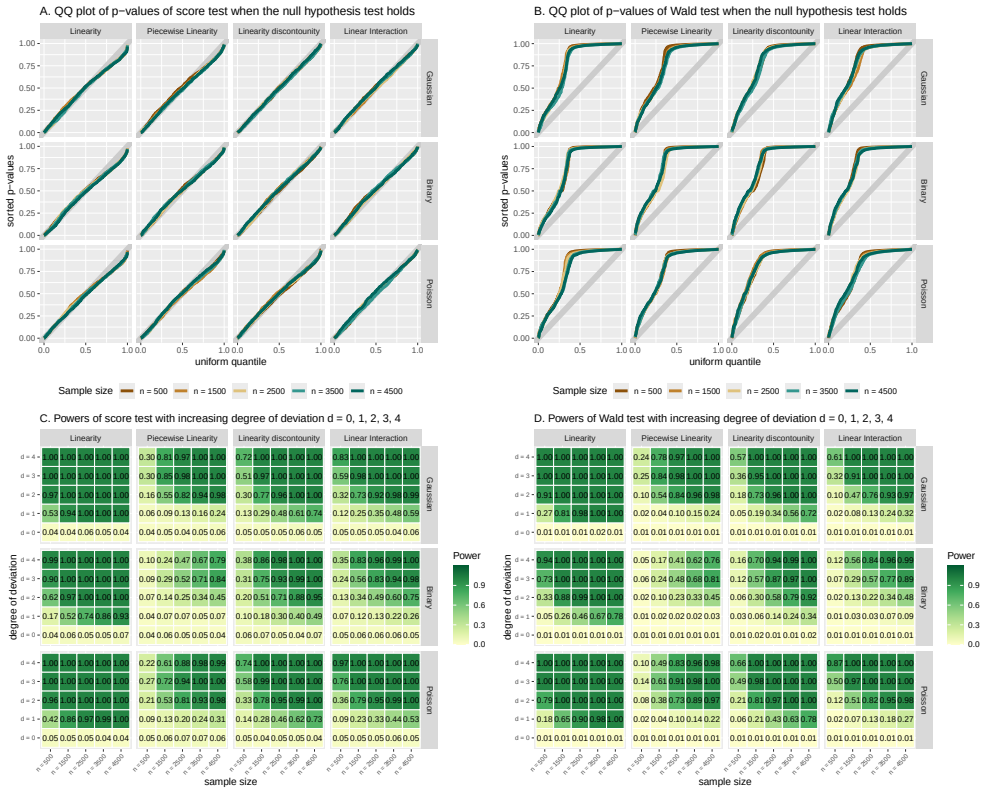


Fig. 2: Performance of the Wald and score tests under the null and alternative hypotheses. **A**. QQ plots of p-values under the null hypothesis for the score test, and **B**. for the Wald test, across four structure types (columns) and three outcome distributions (rows), with sample sizes ranging from $n = 500$ to $n = 4500$. Each line compares the empirical distribution of p-values against the expected uniform distribution (gray dashed line). **C**. Heatmaps of power under the alternative hypothesis for the score test, and **D**. for the Wald test, with increasing deviation from the null controlled by parameter d (rows) and sample size (columns). Both color and overlaid numeric values indicate rejection frequencies (power).

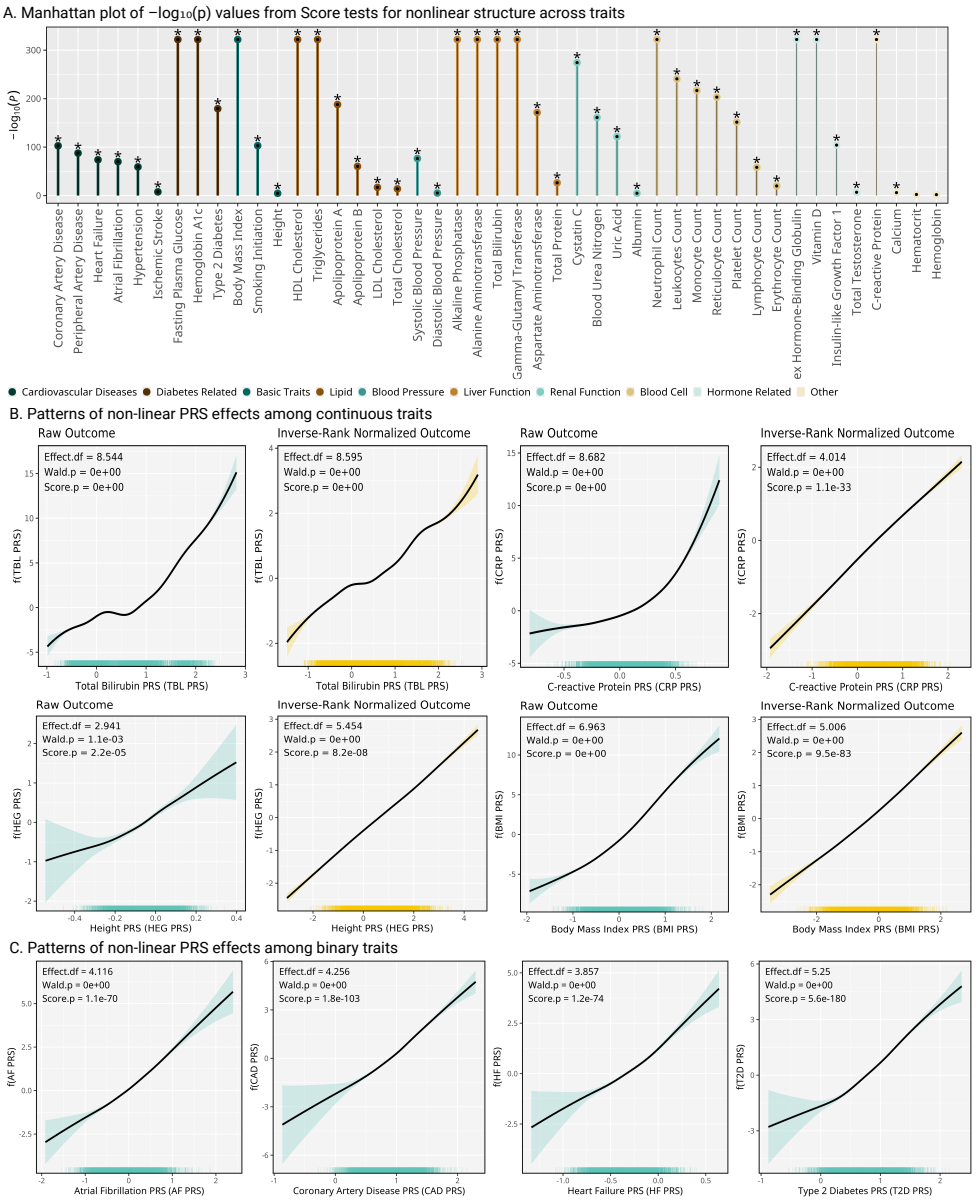
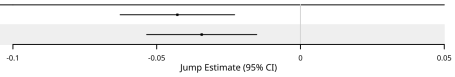


Fig. 3: Linearity tests and nonlinear effect patterns of PRSs across metabolic and cardiovascular traits. **A**. Manhattan plot of $-\log_{10}(p)$-values from score tests evaluating the linearity of PRS effects across a range of traits. Each point represents one trait, with higher values indicating stronger evidence of deviation from linearity. Asterisk * indicates score test p-value $< 0.05/48$. **B**. Estimated nonlinear effect functions for continuous traits. Green curves correspond to raw traits, and yellow curves correspond to IRNT traits. **C**. Estimated nonlinear effect functions for selected binary traits. In both **B** and **C**, shaded bands indicate 95% confidence intervals, and tick marks on the x-axis show the distribution of PRS values. Each panel also reports the edf, the p-value from the score test, and the p-value from the Wald test for nonlinearity.

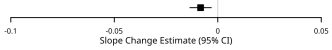
A. Summary of effect jump test at age = 65 across multiple traits

Outcome	Score Test P	Jump P	Jump Q
Driving Time	9.65e-02	2.47e-05	5.26e-03
Walking Time	9.29e-01	4.32e-04	6.16e-02



B. Summary of effect slope-change test at age = 65 across multiple traits

Outcome	Score Test P	Slope P	Slope Q
Alanine Aminotransferase	8.27e-01	1.40e-03	1.12e-02
Apolipoprotein A	2.83e-01	1.79e-06	3.80e-05
Apolipoprotein B	1.37e-01	2.17e-03	1.89e-02
Cooked Vegetable Intake	7.19e-02	4.05e-05	1.01e-03
Cystatin C	1.04e-01	5.87e-03	6.72e-02
Erythrocyte Count	2.32e-01	7.03e-04	1.65e-02
Gamma-Glutamyl Transferase	3.35e-01	1.28e-04	3.30e-03
HDL Cholesterol	3.34e-01	1.33e-03	2.45e-02
Hematocrit	4.08e-01	2.52e-08	2.03e-06
Hemoglobin	3.33e-01	1.86e-07	1.09e-05
Television Time	1.59e-01	2.25e-03	3.64e-02
Total Cholesterol	3.18e-01	1.27e-03	2.85e-02
Triglycerides	7.74e-01	1.63e-03	2.95e-02



C. Patterns of age effect for traits with significant jump or slope change at age = 65

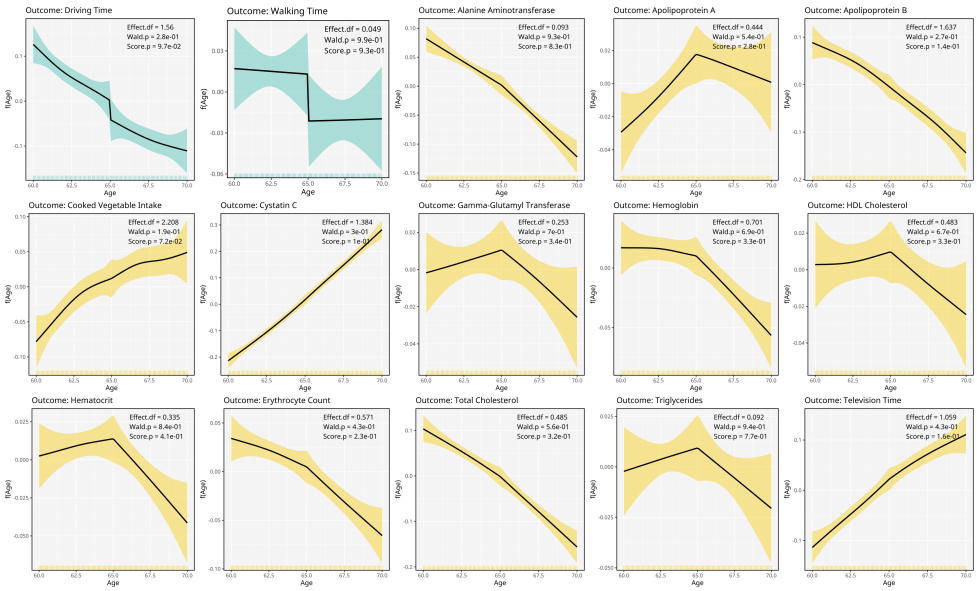


Fig. 4: Regression discontinuity design analysis of the impact of retirement (age = 65) on health and lifestyle outcomes in the UK Biobank. **A**. Summary of effect jump tests at age 65 across multiple traits. **B**. Summary of effect slope change tests at age 65. **C**. Estimated age effect patterns for selected traits with significant jump or slope change. In **A** and **B**, significant results are determined based on two criteria: (1) a non-significant score test for nonlinear age effect ($p > 0.05$), and (2) a significant jump or slope change test at age 65 (FDR $q < 0.1$). To the right of each smooth function plot is a forest plot summarizing the estimated effect size at age 65 and its 95% confidence interval. In **C**, each panel displays the fitted age effect function, with the vertical dashed line at age 65 denoting the retirement threshold. The upper-left corner of each panel reports the effective degrees of freedom (edf), the p-value from the score test, and the p-value from the Wald test for nonlinearity. Green panels indicate traits with linear breakpoint structures, while yellow panels correspond to piecewise linear effects.