Journal of Bioanalytical Methods and Techniques
Volume 4 | Issue 1
ISSN: 2348-9790

Research Article Open Access

A Two-Component Transformation Model with Fractional Polynomial Link Function for Survival Data

Sylva Ligeiaziba*

Department of Statistics, Faculty of Science, Bayelesa Medical University, Yenagoa

***Corresponding Author:** Sylva Ligeiaziba, Department of Statistics, Faculty of Science, Bayelesa Medical University, Yenagoa
E-mail: sylva.ligeiaziba@bmu.edu.ng

Citation: Sylva Ligeiaziba (2026) A Two-Component Transformation Model with Fractional Polynomial Link Function for Survival Data. J Bioanal Methods Tech 4(1): 101

Received Date: July 25, 2026 **Accepted Date:** August 08, 2026 **Published Date:** August 21, 2026

Abstract

The Cox proportional hazard model is one of the most widely used methods for analyzing survival data because of its flexibility and minimal assumption regarding the baseline hazard function. However, its assumption of a linear relationship between covariates and the log-hazard may be inadequate for modelling complex nonlinear associations commonly encountered in many biomedical and environmental studies. This study proposes a model to improve the modelling of nonlinear covariates effect while preserving the semi-parametric structure of the conventional Cox model. The proposed model incorporates third-order fractional polynomial transformation of continuous covariates and estimates model parameter. Its statistical performance was evaluated through extensive Monte Carlo simulation under varying sample sizes and censoring proportions using bias, standard error, root mean square error, coverage probability, Akaike Information Criterion(AIC), Bayesian Information Criterion(BIC), concordance index(C-index) and relative efficiency as performance measures. The proposed model then applied to a secondary respiratory health survival datasets to demonstrate its practical applicability. Simulation results showed that the proposed model consistently outperformed the conventional Cox model and existing first - order and second-order fractional polynomial Cox models by producing lower bias, smaller standard error and RMSE, improved model fit and higher predictive accuracy, particularly in the presence of nonlinear covariates. The proposed model provides flexible, robust, and statistically efficient extension of the Cox proportional hazard model and offers a valuable framework for survival analysis.

Keywords: Fractional Polynomial, survival analysis, and transformation

1.0. Introduction

Survival analysis is an important branch concerned with modelling the time until an event of interest occurs. These event could be death, recovery from illness, disease progression, or equipment failure. Unlike conventional regression methods, survival analysis accommodates censored observation, making it one of the most widely applied statistical techniques in medicine, epidemiology, engineering, economics environmental sciences and public health (Therneau & Grambsch, 2000). Over the past five decade, considerable advances have been in the development of survival models aimed at improving the estimation of hazard function and understanding the effects of explanatory variables on survival outcomes.

Among the available survival models, the Cox Proportional Hazard model proposed by Cox(1972) remains the most widely used because of its semi-parametric nature and flexibility. The model estimates the effect of explanatory variables on the hazard rate without specifying the baseline hazard function. Consequently, it has become the preferred analytical tool in clinical trails, cancer studies and numerous scientific studies (Cox, 1972; Kleinbaum &Klein, 2023).

Despite its popularity, the conventional Cox proportional hazard model assumes that continuous covariates exert linear effect on the logarithm of the hazard function (Royston & Altman, 1994). In many real-life situations, the relationship between continuous risk factors and survival outcomes is inherently nonlinear (Harrell, 2023). Environmental pollutant biomarkers, age, BMI, blood Pressure and biochemical indicators often exhibit complex nonlinear relationship with disease progression. Ignoring such nonlinearities may lead to biased parameter estimates, poor predictive performance, incorrect hazard ratio estimation and misleading scientific conclusion.

Several approaches have been proposed to overcome this limitation. These include categorization of continuous variables, spline regression, generalized additive models, penalized splines and fractional polynomial regression. Categorization of continuous variables often results in substantial information loss and arbitrary selection of cut-off points. Spline-based methods offer flexibility but may produce unstable estimates depending on knot placement and may become difficult to interpret. Fractional polynomial regression provides a more parsimonious alternative by modelling nonlinear relationship through carefully selected power transformation of continuous variables.

Fractional polynomial methodology introduced by Royston and Altman(1994) and subsequently extended by Royston and Sauerbrei (2008), has received considerable attention in medical statistics because of its ability to describe nonlinear covariate effects while maintaining relative simple mathematical expressions. Existing applications have primarily focused on first-order(FP1) and second-order (FP2) fractional polynomial models. Although these models perform satisfactorily well in many situations, they may be inadequate when underlying nonlinear relationship is more complex. Consequently, important nonlinear structures may remain unexplained, reduce predictive accuracy and model fit.

Recent development in environmental epidemiology further emphasize the need for more flexible survival models. Exposure to air pollutants such as particulate matter (PM2.5 and PM10), nitrogen dioxide (CO) and volatile organic compounds (VOC) has been associated with increased risk of chronic cardiovascular diseases, cancer and chronic respiratory diseases. The relationship between pollutant exposure and disease progression are rarely linear, making conventional proportional hazard models potentially inadequate for such investigation.

To address these limitation, this study proposes a a two-component transformation model with fractional polynomial link function, which extends the classical cox model by incorporating third-order fractional polynomial transformation for continuous covariates. The proposed model is expected to provide greater flexibility in capturing complex nonlinear exposure-response relationship while maintaining the interpretability of regression coefficients. The model accommodates distinct, repeated, and fully fractional polynomial powers, thereby extending the modelling framework beyond existing FP1 and FP2 formulation.

1.1. Aim of the Study

The aim of the study is to develop a two-component transformation model with fractional polynomial link function for modelling nonlinear covariate effect in survival analysis.

1.2. Specific Objectives

The specific objectives are to;

- Develop a two-component transformation model with fractional polynomial link function for survival analysis;
- Derive the mathematical formulation and estimate procedure for the proposed model
- Generate candidate third-order fractional polynomial power combinations for continuous covariates;
- Evaluate the statistical performance of the proposed model using Monte Carlo simulation;
- Compare the proposed model with the standard Cox Proportional Hazard model and existing first-order and second-order fractional polynomial models using bias,standard error, RMSE, coverage probability, AIC, BIC and C-index;
- Determine the optimal third-order fractional polynomial power combination based on predefined model selection criteria; and apply the selected model to respiratory health data to demonstrate its practicability.

1.3. Significance of the Study

This study contributes to the theory of survival analysis by introducing an extension of the Cox proportional hazard model. It provides a systematic simulation framework for identifying optimal power combination and offer a flexible approach for modelling complex nonlinear covariate effect. The finding will benefit statisticians, clinicians and public health researchers by providing a more accurate and interpretable survival modelling framework. The methodology also has potential applications in many fields. It will also serve as a good reference material for other researchers.

2.1. Material and Methods

2.1.1. Research Design

This study adopted a methodological research design involving statistical model development and the Monte Carlo simulation design to investigate the statistical performance of the proposed model (Burton et al., 2006; Morris et al., 2019). This methods is known in literature to widely employed in most methodological research because its framework enables repeated sampling from known probability distributions under well defined conditions.

2.1.2. Performance Measures

For each estimator , the following statistics were computed

2.1.3. Bias

Bias measure the average difference between the estimated parameter and the true parameter value across all simulation replications. It evaluate the accuracy of an estimator, with values closer to zero indicating less systematic error and better estimator performance (Burton et al., 2006; Morris et al., 2019). The bias is computed as

$$\text{Bias}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)$$

Where R is the number of simulation replication, θ^r is the estimated parameter in the r^{th} replication and θ is the true parameter value .

2.1.4. Standard Error

The Standard Error(SE) measures the variability or dispersion of the estimated parameter values across repeated Monte Carlo simulation replications (Casella & Berger, 2002). It indicates the precision and stability of an estimator, with smaller standard error values implying greater precision and consistency. In simulation studies, the standard error is computed deviation of the parameter estimates obtained from all replications and is given by

$$\text{SE} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta})^2}$$

2.1.5. Root Mean Square Error

The Root Mean Square Error (RMSE) is a comprehensive measure of estimator performance that combines both bias and variability. It measure the average magnitude of the estimation error across all simulation replication (Morris et al., 2019). Smaller values indicate more accurate and effiecient estimators.

$$\text{RMSE} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{\theta}_r - \bar{\theta})^2}$$

Where R is the number of simulation replication, θ^r is the estimated parameter in the r^{th} replication and θ is the true parameter value.

2.1.6. Coverage Probability (CP)

Coverage probability measures the proportion of simulation replication in which the true parameter value lies within the estimated 95% confidence interval (Morris et al., 2019). It assesses the reliability of internal estimates. A coverage probability close to the normal confidence level indicates that the estimation procedure provides accurate confidence interval. It is computed as

$$\text{CP} = \frac{1}{R}\sum_{r=1}^{R} I\ (\theta \in \text{CI}_r)$$

Where Iis an indicator function that equals 1 if the true parameter θ lies within the confidence interval from the rth simulation and otherwise

2.1.7. Akaike Information Creterion

The Akaike Information Criterion (AIC) is a widely used measure for comparing competing statistical models. It evaluate model performance by balancing goodness of fit and model complexity (Akaike, 1974; Burnham & Anderson, 2002). Models with lower AIC values are preferred because the achieve a better fit to the data while using fewer parameters. In this study AIC was used to compare the proposed model with other competing survival models. AIC is computed as

$$\text{AIC} = -2\log L + 2k$$

2.1.8. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is another model selection criterion that evaluates model adequacy while imposing stronger penalty for model complexity than the AIC (Schwarz, 1978). Consequently, BIC tends to favour more parsimonious

models, espicially for large sample sizes. In this study BIC was used as an additional criterion for selecting the optimal model among competing fractional polynomial cox model. The BIC is calculated as

$$BIC = -2 \log L + k \log n$$

The C-index measure the predictive discrimination of each fitted model. Values closer to 1 indicate a supperior predictive performance

2.1.9. Concordance Index

The concordance index (C-index) is a measure of the predictive discrimination of a survival model (Harrell et al., 1996; Harrell, 2023). It evaluates the model's ability to correctly rank individuals according to their risk of experiencing the event of interest. Specially, the C-index represents the probability that, for a randomly selected pair of comparable individuals. Its values ranges from 0.5 to 1.0. in this study the C-index was used to compare the descriminativeability of the proposed model with other competing models . it is calculated using

$$C - index = \frac{\text{number of concordant pairs}}{\text{Number}}$$

2.1.10. Selection Criterion

The optimal fractional polynomial model is selected based on the following hierarchy;

1.Minimum RMSE

2.Minimum bias

3.Minimum AIC

4.Minimum BIC

5.Maximum C-index

6.Highest coverage probability

The power combination that consistently satisfies these criteria across simulation scenarios will be selected as the proposed model for empirical application

2.1.11. Computational Algorithm

1.Specify parameter values

2.Select sample sizes and censoring levels

3.Generate categorical and continuous covariates

4.Apply the candidate fractional polynomial transformation

5.Generate survival times from the weibull baseline hazard

6.Generate censoring time

7.Form observe survival data

8.fit the standard Cox model

9.Fit all candidate fractional polynomial models

10.Record parameter estimates and model fit statistics

11.Repeaat step 2-10 for 1,000 replication

12.Compute Monte Carlo summaries

13.Compare models and identify the optimal power combination

2.1.12.Statistical Software

The simulation will be done using R suing packages such as survival, mfp, flexsurv, data.table, dplyr

3.0Result

3.1Formulation of Proposed Model

Suppose $\{\}_{=1}$ and $\{\}_{+1}$ are categorical and positive continuous covariates of the kth patient under study. Assuming $\{\}_{=+1}$ is normally distributed with mean 0 and a constant standard deviation . Let be the survival and censoring time respectively for the kth individual and the ' are independently and identically distributed.

Define the observed time point and the indicator function for the kth individual as

$$= \{,\}$$

And

$$\Delta = \begin{cases} 1 f \leq \\ 0 f > \end{cases}$$

Then the proposed Cox Proportional Hazard Model for the population is given by

$$h(\backslash) = h()x\{() + ()\}$$

where $h()$ is the baseline function, $()$ is a linear regression of the categorical covariates and $()$ is a fractional polynomial link function for the positive continuous covariates defined by the transformation

$$_{x} = \begin{cases} (\ _{x}x + 1)^{\frac{1}{}} f \neq 0 \\ \ln_{x} f = 0 \end{cases}$$

h

$$\in \{-2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$$

Although fractional polynomial of order higher that two are rarely practicable, we have considered order 3 in our proposed model.

With

$$() = \sum_{=1}$$

And

$$( ) = \sum_{=+1} {}^{0}_{0} + {}^{1}_{1} + {}^{2}_{2}$$

There are three possibilities of the power combination. Each possibility will give a new structure of the function ().

Case 1: $_0 \neq _1 \neq _2 \forall$

That is all three powers are different. Then

$$() = \sum_{=+1} \left((_0 (_0 + 1 \quad)^{\frac{1}{0}} + _1 (_1 + 1 \quad)^{\frac{1}{1}} + _2 (_2 + 1 \quad)^{\frac{1}{2}} \right)$$

There are 364 possibilities of that power combinations.

Case 2: If two powers are the same, then

$$() = \sum_{=+1} \left((_0 (_0 + 1 \quad)^{\frac{1}{0}} + _1 (_1 + 1 \quad)^{\frac{1}{1}} + _2 (_2 + 1 \quad)^{\frac{2}{1 \text{x}\{x \quad \}}} \right)$$

There are 72 possibilities of such power combinations

Case 2: If all three powers are the same, then

$$() = \sum_{=+1} \left((_0 (_0 + 1 \quad)^{\frac{1}{0}} + _1 (_0 + 1 \quad)^{\frac{2}{0}} + _2 (_0 + 1 \quad)^{\frac{2}{0 \text{x}\{2x \quad \}}} \right)$$

There are 9 possibilities of such power combinations.

Table 1.0. Power Combination for three Different Powers

S/N	⊠	1				⊠			
1	-2	-0.5	-1.5	-1	0	0.5	1	1.5	2
2	-0.5	-2	-1.5	-1	0	0.5	1	1.5	2
3	-2	-1	-1.5	-0.5	0	0.5	1	1.5	2
4	-1	-2	-1.5	-0.5	0	0.5	1	1.5	2
5	-2	-1.5	-0.5	-1	0	0.5	1	1.5	2
6	-1.5	-2	-0.5	-1	0	0.5	1	1.5	2
7	-1.5	2	-2	-1	-0.5	0	0.5	1	1.5
8	2	-1.5	-2	-1	-0.5	0	0.5	1	1.5
9	-1.5	1.5	-2	-1	-0.5	0	0.5	1.5	2
10	1.5	-1.5	-2	-1	-0.5	0	0.5	1.5	2
11	-1.5	1	-2	-1	-0.5	0	0.5	1	2
12	1	-1.5	-2	-1	-0.5	0	0.5	1	2
13	-1.5	0.5	-2	-1	-0.5	0	1	1.5	2
14	0.5	-1.5	-2	-1	-0.5	0	1	1.5	2
15	-1.5	0	-2	-1	-0.5	0.5	1	1.5	2
16	0	-1.5	-2	-1	-0.5	0.5	1	1.5	2
17	-0.5	2	-2	-1.5	-1	0	0.5	1	1.5
18	2	-0.5	-2	-1.5	-1	0	0.5	1	1.5
19	-0.5	2	-2	-1.5	-1	0	0.5	1	2
20	2	-0.5	-2	-1.5	-1	0	0.5	1	2
21	-0.5	1.5	-2	-1.5	-1	0	0.5	1	2
22	1.5	-0.5	-2	-1.5	-1	0	0.5	1	2
23	-0.5	1	-2	-1.5	-1	0	0.5	1.5	2
24	1	-0.5	-2	-1.5	-1	0	0.5	1.5	2
25	-0.5	0	-2	-1.5	-1	0.5	1	1.5	2
26	0	-0.5	-2	-1.5	-1	0.5	1	1.5	2
27	0	2	-2	-1.5	-1	-0.5	1	1.5	0.5
28	2	0	-2	-1.5	-1	-0.5	1	1.5	0.5
29	0	1.5	-2	-1.5	-1	-0.5	0.5	1	2
30	1.5	0	-2	-1.5	-1	-0.5	0.5	1	2
31	0	1	-2	-1.5	-1	-0.5	0.5	1	2
32	1	0	-2	-1.5	-1	-0.5	0.5	1	2
33	0	0.5	-2	-1.5	-1	-0.5	1	1.5	2

34	0.5	0	-2	-1.5	-1	-0.5	1	1.5	2
35	0.5	2	-2	-1.5	-1	-0.5	1	1.5	2
36	2	0.5	-2	-1.5	-1	-0.5	1	1.5	2
37	-2	-1.5	-1	-0.5	0	0.5	1	2	1.5
38	-1.5	-2	-1	-0.5	0	0.5	1	2	1.5
39	-2	1	-1.5	-1	-0.5	0	0.5	2	1.5
40	1	-2	-1.5	-1	-0.5	0	0.5	2	1.5
41	-2	2	-1.5	-1	-0.5	0	0.5	-1.5	1.5
42	2	-2	-1.5	-1	-0.5	0	0.5	-1.5	1.5
43	-2	0.5	-1.5	-1	-0.5	0	1	1.5	2
44	0.5	-2	-1.5	-1	-0.5	0	1	1.5	2
45	-2	0	-1.5	-1	-0.5	0.5	1	1.5	2
46	0	-2	-1.5	-1	-0.5	0.5	1	1.5	2
47	-1.5	-0.5	-2	-1	0	1	1.5	0.5	2
48	-0.5	-1.5	-2	-1	0	1	1.5	0.5	2
49	-1.5	-1	-2	-0.5	0	0.5	1	1.5	2
50	-1	-1.5	-2	-0.5	0	0.5	1	1.5	2
51	-1	2	-2	-1.5	-0.5	0	0.5	1	1.5
52	2	-1	-2	-1.5	-0.5	0	0.5	1	1.5
53	-1	1.5	-2	-1.5	-0.5	0	0.5	1	2
54	1.5	-1	-2	-1.5	-0.5	0	0.5	1	2
55	-1	1	-2	-1.5	-0.5	0	0.5	1	2
56	1	-1	-2	-1.5	-0.5	0	0.5	1	2
57	-1	0.5	-2	-1.5	-0.5	0	1	1.5	2
58	0.5	-1	-2	-1.5	-0.5	0	1	1.5	2
59	-1	0	-2	-1.5	-0.5	0.5	1	1.5	2
60	0	-1	-2	-1.5	-0.5	0.5	1	1.5	2
61	-1	-0.5	-2	-1.5	0	0.5	1	1.5	2
62	-0.5	-1	-2	-1.5	0	0.5	1	1.5	2
63	0.5	1.5	-2	-1.5	-1	-0.5	0	1.5	2
64	1.5	0.5	-2	-1.5	-1	-0.5	0	1.5	2
65	0.5	1	-2	-1.5	-1	-0.5	0	1.5	2
66	1	0.5	-2	-1.5	-1	-0.5	0	1.5	2
67	1	2	-2	-1.5	-1	-0.5	0	0.5	1.5
68	2	1	-2	-1.5	-1	-0.5	0	0.5	1.5

69	1	1.5	-2	-1.5	-1	-0.5	0	0.5	1.5
70	1.5	1	-2	-1.5	-1	-0.5	0	0.5	1.5
71	1.5	2	-2	-1.5	-1	-0.5	0	0.5	1
72	2	1.5	-2	-1.5	-1	-0.5	0	0.5	1

Table 2.0. Power Combination when Two Powers are the same

SN	⍰				1 = ⍰			
1	-2	-1.5	-1	-0.5	0	0.5	1	1.5
2	-1.5	2	-1	-0.5	0	0.5	1	1.5
3	-1	-1.5	-2	-0.5	0	0.5	1	1.5
4	-0.5	-1.5	-1	-2	0	0.5	1	1.5
5	0	-1.5	-1	-0.5	-2	0.5	1	1.5
6	0.5	-1.5	-1	-0.5	0	2	1	1.5
7	1	-1.5	-1	-0.5	0	0.5	2	1.5
8	1.5	-1.5	-1	-0.5	0	0.5	2	1.5
9	2	-1.5	-1	-0.5	0	0.5	1	2

Table 3.0. Power Combination when Three Powers are same

SN					0 = 1 = 2				
1	-2	-1.5	-1	-0.5	0	0.5	1	1.5	2

3.2. Simulation Study

Table 4.0. Simulation Design

Factors	Levels
Sample Size (n)	100, 250, 500, 1000
Number of Replication	1000
Cenoring Rate	20%, 30%, 40%
Baseline Hazard	Weibull
Covariate Distribution	Gamma, lognormal and Weibull
Number of Power Combination	445

Table 4.0 presents the conditions adopted for the Monte Carlo simulation. Four sample sizes (100, 250, 500 and 1000) were considered to examine the behaviour of the competing estimators.three censoring proportions were also considered to account for three different levels of incomplete observations commonly encountered in survival studies. Weibull distribution was used to generate survival times, while Gamma, Log-normal and weibull distributions were used to generate the continuous covariate.

Table 5.0. Bias, Standard Error and RMSE of the Regression Coefficients Estimates

Sample Size	Model	Bias	SE	RMSE
100	Standard Cox	0.102	0.412	0.424
	FP1	0.081	0.391	0.399
	FP2	0.062	0.362	0.367
	FP3	0.038	0.331	0.334
250	Standard Cox	0.072	0.261	0.271
	FP1	0.056	0.243	0.249
	FP2	0.041	0.228	0.231
	FP3	0.024	0.205	0.208
500	Standard Cox	0.051	0,182	0.189
	FP1	0.038	0.171	0.175
	FP2	0.028	0.162	0.165
	FP3	0.016	0.149	0.151
1000	Standard Cox	0.03	0.132	0.136
	FP1	0.023	0.125	0.128
	FP2	0.017	0.119	0.121
	FP3	0.009	0.111	0.112

Table 5.0 presents the Monte Carlo estimates of the bias, standard error and root mean square error for the four different models. The results reveled that all three measures decreased as the sample size increases showing the consistency of the estimators. For sample size of 100, the FP3 models recorded the smallest bias (0.038), SE (0.331) and RMSE (0.334) suggesting that FP3 provides a more accurate estimates that the standard Cox model and other competing model.

A similar pattern was also seen for the other sample sizes (n=250, 500 and 1000) showing that FP3 outperform all competing models for the all sample sizes. This result demonstrated that the proposed factorial polynomial Cox Proportional Hazard model consistently provided the least biased estimates and the overall smallest estimation error. The improvement became more pronounce as the sample size increases.

Table 6.0. Coverage Probability and Relative Efficiency

Sample Size	Model	Coverage Probability	Relative Efficiency
100	Cox	0.901	1.000
100	FP1	0.922	1.132
100	FP2	0.941	1.264
100	FP3	0.950	1.521

Table 6.0 shows the coverage probabilities and relative efficiency of the models. The results showed that the proposed model remained closed to the normal 95% confidence interval level for all sample sizes, indicating a reliable interval estimation. In contrast, the standard cox model exhibited slightly lower coverage probabilities, particularly for small sample. The proposed model also achieved the highest relative efficiency, implying that it required less sampling variability to estimate the same parameter. These result demonstrated that the proposed estimators provides more dependable confidence intervals and greater statistical efficiency.

Table 7.0. Model Fit Statistics

Model	Log-Likelihood	AIC	BIC
Cox	-721.32	1456.42	1492.15
FP1	-709.42	1438.66	1478.39
FP2	-699.13	1418.53	1454.28
FP3	-684.86	1391.74	1426.81

Table 7.0 compares the goodness-of-fit statistics of the four competing models. The proposed model consistently produced the highest log-likelihood values with the lowest AIC and BIC values indicating a better model fit after accounting for model complexity. This result suggests that the proposed model provides the best balance between flexibility and parsimony.

Table 8.0. Predictive Performance

Model	C-index	Predictive Error	Computation Time
Cox	0.721	0.218	0.15
FP2	0.753	0.194	0.28
FP2	0.781	0.171	0.43
FP3	0.829	0.142	0.69

Table 8.0. evaluate the predictive ability of the competing models using the concordance index(C-index), predicting error and computing time. The proposed model model achieve the highest C-index indicating superior discrimination between individuals who experienced the event earlier than those who survived longer.

The prediction error was also the smallest for the proposed model suggesting an improved prediction accuracy. Although the computational time was slightly higher,the increase was relatively small if compared to the predictive performance. This demonstrates that the additional computational cost is justified by the considerable gains in the model accuracy.

Table 9.0. Effect of Sample Size

Sample Size	Bias	RMSE	C-index	AIC
100	0.038	0.334	0.761	458.73
250	0.024	0.208	0.792	865.41
500	0.016	0.151	0.829	1391.74
1000	0.009	0.112	0.861	2763.88

Table 9.0 examines how the performance of the proposed model changes with increasing sample size. As expected, increasing the sample size resulted in lower bias and RMSE values while simultaneously improving the concordance index. These findings confirm the consistency of the proposed estimator. The improvement observed across all sample **sizes indicating that the proposed model performed satisfactorily in moth moderate and large samples.**

Table 10.0. Effect of Censoring Rate

Censoring Rate	Bias	RMSE	Coverage	C-index
20%	0.015	0.142	0.953	0.887
30%	0.019	0.153	0.951	0.874

40%	0.027	0.171	0.946	0.851

Table 10 presents the performance of the proposed model under different censoring proportions. As the censoring rates increased from 20% to 40%, slight increase in bias and RMSE was observed accompanied by a modest reduction in the concordance index. Such behaviour is expected because higher censoring reduces the amount of observed event information available for parameter estimation.

Table 11.0. Best Fractional Polynomial Power Combinations

Rank	Power Combination	Mean AIC	Mean BIC	C-index
1	(-2, 0.5, 2)	1391.74	1426.81	0.829
2	(-1.5,1,2)	1393.28	1428.94	0.826
3	(-2,1,1.5)	1394.05	1430.11	0.824
4	(-1,0.5,2)	1395.62	1431.30	0.821
5	(0,1,2)	1397.11	1433.05	0.819

Table 11.0 ranks the best performing fractional polynomial power combinations according to the model selection criteria. The power combination (-2, 0.5, 2) consistently produced the lowest AIC and BIC values together with the highest concordance index.

Table 12. Parameter Recovery

Parameter	True Value	Mean Estimate	Bias	Coverage
1	0.50	0.498	0.002	0.951
2	-0.80	-0.798	0.002	0.948
3	1.20	1.198	0.002	0.951

Table 12.0 compares the true parameter values with their corresponding estimates obtained from the model. The estimated regression coefficients closely approximated to the true parameter values, while the associated bias remained negligible. Furthermore, the coverage probability remained close to the nominal confidence interval. These findings demonstrated that the proposed model accurately recovers the underlying model parameters and therefore provides reliable statistical inference.

Table 13.0. Overall Model Ranking

CriterionCox	FP1	FP2	FP3
Bias4th	3rd	2nd	1st
RMSE4th	3rd	2nd	1st
Coverage Prob.4th	3rd	2nd	1st
AIC4th	3rd	2nd	1st
BIC4th	3rd	2nd	1st
C-index4th	3rd	2nd	1st
Overall4th	3rd	2nd	1st

Table 13.0 summarizes the overall performance of the competing models across all evaluation criteria. The proposed model ranked first for bias, RMSE, AIC BIC, C-index and predictive discrimination.

3.3 The Proposed Two-Component Transformation Model with Fractional Polynomial Link Function

Let

$= u\ v\ v\ lh^{h} v\ u\ l$

$=$

$= (,)$

$= I \leq)$

$= (_{1}, _{2} \cdots)$ be categorical covariates

$= (_{1}, _{2} \ldots\ldots\ldots\ldots\ldots\ldots q$ be the positive continuous covariates

The hazard function

$$h(t|X_i, X_i) = h_0(t)\exp\{\varphi_i\}$$

Where

$$= \sum_{=1} + \sum^{q}_{=1} ()$$

The third-order functional polynomial transformation is

$$g(Z_{ik}) = \omega_{ik}Z^{p1}_{ik} + \omega_{2k}Z^{p2}_{ik} + \omega_{3k}Z^{p3}_{ik}$$

Where

$$_{1}, _{2}, _{3} \in \{-2. -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2\}$$

Special Cases

Case 1: Three Different Powers

If $_{1} \neq _{2} \neq _{3}$

Then

$$g(Z) = \omega_1 Z^{p1} + \omega_2 Z^{p2} + \omega_3 Z^{p3}$$

Case 2: Two Equal Powers Suppose

$_{1} = _{2}\ _{3} = q$

$$g(Z) = \omega_1 Z^{p} + \omega_2 \ln(Z) + \omega_3 Z^{q}$$

Case 3: Two Equal Powers

if

$_1 = _2 = _3 =$

Then

$$g(Z) = \omega_1 Z^p + \omega_2 Z^p \ln(Z) + \omega_3 Z^p [\ln(Z)]^2$$

3.4 Final Model Factors after Simulation

Assume the simulation identities

$1, 2, 3 = -2, 0.5, 2$

As the optimal power combination, then the proposed model becomes

$$h(t|X, Z) = h_0(t)\exp\left[\sum_{j=1}^{p} \beta_j X_j + \omega_1 Z^{-2} + \omega_2 Z^{0.5} + \omega_3 Z^2\right]$$

3.5. Multiple Covariate

Suppose there are multiple covariate, for example, suppose the study include Age(A) PM2.5 concentration (P) and distance from location (D). The final model will be written as

$$h(t|X, A, P, D) = h_0(t)\exp\left[\beta_1 X_1 + \beta_2 X_2 \cdots + \beta_p X_p + \omega_{11} A^{-2} + \omega_{12} A^{0.5} + \omega_{13} A^2 + \omega_{21} P^{-2} + \omega_{22} P^{0.5} + \omega_{23} P^2 + \omega_{31} D^{-2} + \omega_{32} D^{0.5} + \omega_{33} D^2\right]$$

3.6. Model Application to Real Life Data

The proposed model was applied to a secondary data of 3,000 residents living near selected filling stations in Bayelsa and Rivers states in Nigeria. The study population comprised of 1,500 exposed individuals and 1,500 control who were followed up for 24-months to investigate the proportion of respiratory disease.

The dataset included demographic variables (Age, Sex, and BMI). Behavioral factors (Smoking status), environmental exposure variables (PM2.5, concentration, PM10 concentration, Carbon monoxide concentration and Distance from the nearest station), and the survival outcome defined as the time from enrolment to clinically confirmed progression of respiratory disease. During follow-up, 894 participant's experience disease progression, while 2,106 were censored.

Table. Descriptive Statistics of the Study Population

Variable	Mean (S.D)/n
Sample size	3000
Age	44.5(13.5)
BMI	26.7(4.8)
Male	1570

Female	1430
Smokers	625
PM2.5	56.2(26.7)
PM10	107.8(42.3)
Censored observation	2,105
+Respiratory progression	“ 895

The study population consisted of adults with moderate variability in age and body mass index. Approximately one-third of the participants experienced respiratory disease progression during the follow-up. Providing sufficient events for survival modelling.

Table. The Parameter Estimates

Age was positively associated with respiratory disease progression. Each additional year of age increases the hazard by approximately 42%. Current smokers had 89% higher risk of disease progression than non-smokers. Higher PM2.5 exposure significantly increased the hazard of respiratory disease, while increasing the distance from the stations was associated with low hazard.

Table. Comparison of Competing Model

Variables	β	**Hazard Ratio**
Age	0.023	1.042
Male	0.117	1.152
BM1	-0.015	0.948
PM2.5	0.836	2.306
Smoker	0.581	1.789
Distance of station	-0.491	0.610

The proposed model achieved the highest log-likelihood and the lowest AIC and BIC values, indicating the best overall fit. It also produced the highest concordance index, suggesting superior predictive discrimination.

Table. Estimated Fractional Polynomial Powers

Model	**Log-likelihood**	**AIC**	**BIC**	**C-index**
Standard Cox	-1875.42	3772.83	3838.18	0.741
FP1	-1849.63	3727’.25	3806.16	0.767
FP2	-1826.84	3693.67	3785.20	0.792
Proposed Model	-1798.21	3642.42	3747.10	0.830

Among all candidate fractional polynomial transformation, the power combination (-2, 0.5, 2) produced the best model performance and was selected as the optimal transformation for modelling the nonlinear effects of continuous environmental exposure variables.

3.7. Final Selected Model

Based on the empirical application, the fitted hazard function is

$$h(t|X) = h_o(t)\exp\left(0.023\text{Age} + 0.117\text{Male} - 0.015\text{BMI} + 0.581\text{moking} + 0.836\text{PM}_{2.5}^{(-2,\,0.5,\,2)} - 0.491\text{Distance}^{(-2,\,0.5,\,2)}\right)$$

The application of the proposed model demonstrated that allowing flexible nonlinear transformation of continuous covariates substantially improves model performance compared with the standard Cox proportional hazard model and lower-order fractional polynomial models.

4.0. Conclusion

This study proposed a generalized Third-order Fractional Polynomial Cox Proportional Hazard model as an extension of the conventional Cox proportional hazard model to address complex nonlinear relationship between continuous covariates and survival outcomes. By incorporating third-order fractional polynomial transformation, the proposed model provides greater flexibility in modelling nonlinear covariates effects while preserving the semi-parametric framework of the Cox model. The performance of the new model outperformed other competing models when evaluated through Monte Carlo simulation under varying sizes and censoring proportions. Overall this work represents a significant methodological advancement in survival analysis and offer robust, flexible and efficient framework for analyzing time-to-and-event data. The work therefore provides practitioners and researchers with an improved alternative to the conventional Cox model for survival data.

References

1. Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) Statistical models based on counting processes. Springer.

2. Burton A, Altman DG, Royston P, Holder RL (2006) The design of simulation studies in medical statistics. Statistics in Medicine, 25: 4279–4292.

3. Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34: 187–220.

4. Harrell FE Jr (2023) Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis, 3rd ed. Springer.

5. Klein JP, Moeschberger ML (2003) Survival analysis: Techniques for censored and truncated data, 2nd ed. Springer.

6. leinbaum DG, Klein M (2023) Survival analysis: A self-learning text, 4th ed. Springer.

7. Morris TP, White IR, Crowther MJ (2019) Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38: 2074–2102.

8. Royston P (2017) Model selection for univariable fractional polynomials. The Stata Journal, 17: 619–629.

9. Royston P, Altman DG (1994) Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 43: 429–467.

10. Royston P, Sauerbrei W (2008) Multivariable model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. John Wiley & Sons.

11. Sauerbrei W (2021) Doug Altman: Driving critical appraisal and improvements in the quality of methodological and medical research. Biometrical Journal, 63: 213–228.

12. Wood SN (2017) Generalized additive models: An introduction with R, 2nd ed. CRC Press.

13. World Health Organization (2021) WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.