

**Fractional polynomial and restricted cubic spline
models as alternatives to categorising continuous data:
applications in medicine**

Onkabetse Vincent Mabikwa

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds

School of Medicine

March 2019

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Published research

Chapter 3 of this PhD thesis contains work from the following publication:

1. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. (2017).

Assessing the reporting of categorised quantitative variables in observational epidemiological studies. *BMC Health Services Research*, 17, 201.

(<http://doi.org/10.1186/s12913-017-2137-z>).

Attributable contents to Onkabetse Vincent Mabikwa: Data extraction, analysis, interpretation of the results and writing of the manuscript.

The contribution of other authors: GDC, BPD, and FSJ contributed to the concept and provided input on the drafting of the manuscript.

Published protocol

Chapter 6 contains work from the following publication:

1. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. (2017).

Modelling the alcohol-blood pressure associations in type 2 diabetes patients: UK Biobank. UK Biobank, United Kingdom.

(<http://www.ukbiobank.ac.uk/2017/07/mr-onkabetse-mabikwa-modelling-the-alcohol-blood-pressure-associations-in-type-2-diabetes-patients-uk-biobank/>).

Attributable contents to Onkabetse Vincent Mabikwa: Writing the research protocol and selecting the data suitable for the chapter.

The contribution of other authors: GDC, BPD, and FSJ contributed to the concept and drafting of the research protocol by proof reading and guiding the selection of data fields.

Conference Presentations

1. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *A simulation study investigating the performance of traditional and alternative approaches for fitting nonlinear exposure-outcome relationships in epidemiology*. Oral Presentation at the LICAMM Early Career Group Science Day. 27th May 2017, University of Leeds, United Kingdom.

Attributable contents to Onkabetse Vincent Mabikwa: Data analysis, interpretation of the results and writing of the abstract and presentation.

The contribution of other authors: GDC, BPD, and FSJ contributed to the concept and proof read the abstract and presentation for the conference.

2. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *A survey based study evaluating the incidence and categorisation of quantitative variables in medical research*. Oral Presentation at 39th Research Students' Conference in Probability and Statistics. 14th – 17th June 2016, Dublin, Ireland.

Attributable contents to Onkabetse Vincent Mabikwa: Data extraction, analysis, interpretation of the results and writing of the abstract and presentation.

The contribution of other authors: GDC, BPD, and FSJ contributed to the concept and proof read the abstract and presentation for the conference.

3. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *Fractional polynomial and restricted cubic spline models as alternatives to categorising continuous data: applications in medicine*. Poster Presentation at Faculty of Medicine and Health Postgraduate Symposium. 29th June 2015, University of Leeds, United Kingdom.

Attributable contents to Onkabetse Vincent Mabikwa: Data analysis, interpretation of the results and writing of the abstract and presentation.

The contribution of other authors: GDC, BPD, and FSJ contributed to the concept and proof read the abstract and presentation for the conference.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Onkabetse Vincent Mabikwa to be identified as the Author of this work has been asserted by Onkabetse Vincent Mabikwa in accordance with the Copyright, Designs and Patent Act 1988.

©2018 The University of Leeds Onkabetse Vincent Mabikwa

Acknowledgements

- The Government of Botswana and Botswana International University of Science & Technology (BIUST) for providing me with the PhD opportunity to further my studies.
- Leeds Institute of Cardiovascular and Metabolic Medicine (LICAMM) for providing training support and an environment conducive for me to complete my PhD
- Special thanks to my supervisors, Dr. Sarah Fleming, Dr. Darren Greenwood and Dr. Paul Baxter for their unlimited feedback and support throughout my journey in this PhD.
- The UK Biobank team for approving my research proposal to use their data in Chapter 6 of this PhD thesis.
- My wife, Blessy Mabikwa for unlimited support and raising our daughter during my stay in the UK.
- My father Alfonse Mabikwa for believing in me and always supporting my educational needs and dreams.

Abstract

Continuous predictor variables are often categorised when reporting their influence on the outcome of interest. This does not make use of within category information. Alternative methods of handling continuous predictor variables such as fractional polynomials (FPs) and restricted cubic splines (RCS) exist.

This thesis first investigates the current extent of categorisation in comparison to these alternative methods. The performances of categorisation, linearisation, FPs and RCS approaches are then investigated using novel simulations, assuming a range of plausible scenarios including tick-shaped associations. The simulation starts with continuous outcomes, and then move onto predictive models where the outcome itself is dichotomised into a binary outcome. Finally, a novel application of the four methods is performed using the UK Biobank data – incorporating additional issues of confounding and interaction.

This thesis shows that the practice of categorisation is still widely used in epidemiology, whilst alternative methods such as FPs and RCS are not. In addition, this research shows that categorising continuous variable into few categories produce functions with large RMSEs, obscure true relations and have less predictive ability than the linear, FP and RCS models. Finally, this thesis shows that nonlinearity and interaction terms are more easily detected when applying FPs and RCS methods. The thesis concludes by encouraging medical researchers to consider the application of FPs and RCS models in their studies.

Table of Contents

Acknowledgements.....	iv
Abstract.....	v
List of Tables	xi
List of Figures.....	xiii
List of Abbreviations	xvii
Chapter 1	1
Introduction.....	1
1.1 Background.....	1
1.1.1 The importance of correctly specifying nonlinear relationships in epidemiology	2
1.1.2 Continuous predictor variables are predominately analysed as categorical measures in the current publications.....	2
1.1.3 Why is categorisation popular in medical studies	2
1.1.4 Challenges of categorising continuous variables during statistical modelling.....	5
1.2 A brief summary of the research motivation and rationale.....	6
1.3 Summary of objectives	7
1.4 Significance and contribution to knowledge.....	10
1.5 Outline of the thesis	12
Chapter 2	14
Statistical approaches for modelling exposure-outcome relationships	14
2.1 Overview of statistical approaches for analysing exposure-outcome relationships in epidemiology	14
2.2 Statistical methods used in the thesis.....	17
2.2.1 Modelling scenario	18
2.2.2 Categorisation/Dichotomisation	19
2.2.3 Linearisation	21
2.2.4 Fractional polynomials (FPs)	21
2.2.5 Restricted cubic splines (RCS).....	26
2.3 Statistical modelling framework	29
2.3.1 The concept of generalized linear models (GLMs).....	30
2.4 Conclusion	31
Chapter 3	32
Cross sectional survey of research methods in current use	32
3.1 Background.....	32

3.2	Methods.....	34
3.2.1	Study selection.....	34
3.2.2	Data extraction.....	38
3.2.3	Statistical analysis.....	38
3.3	Results.....	38
3.3.1	General characteristics.....	38
3.3.2	The incidence of categorisation amongst the exposures or main risk factors	40
3.3.3	Summary of key findings	43
3.4	Discussion	45
3.5	Conclusions.....	48
Chapter 4	50
	Comparison of different approaches for modelling associations between an exposure and a continuous outcome – a simulation study	50
4.1	Introduction.....	50
4.1.1	Aim and objectives:	52
4.2	Simulation framework	53
4.2.1	Introduction	53
4.2.2	Simulation set-up.....	54
4.3	Methods evaluating the performance of statistical models.....	59
4.3.1	Goodness of fit.....	61
4.3.2	The type I error and power rates.....	63
4.3.3	Confidence intervals, turning points, and coverage probabilities ...	65
4.4	Results.....	66
4.4.1	Goodness of fit.....	66
4.4.2	Estimated type I errors and statistical power.....	72
4.4.3	Median predicted associations shapes, their confidence intervals, and turning points	74
4.5	Discussion	88
4.5.1	General approach.....	88
4.5.2	Summary of main results	89
4.5.3	Challenges and limitations.....	94
4.5.4	Strengths and opportunities	95
4.5.5	Novelty	96
4.5.6	Future work.....	97
4.6	Conclusions.....	97

Chapter 5	100
Extensions to prognostic models with binary outcomes – a simulation study. 100	
5.1 Introduction.....	100
5.1.1 Aims and objectives	102
5.2 Monte Carlo simulation framework.....	103
5.2.1 Simulation set-up.....	104
5.2.2 Extension of the alcohol-blood pressure example.....	106
5.3 Approaches for handling continuous predictors	108
5.3.1 Model evaluation	109
5.4 Results.....	116
5.4.1 Comparison of various modelling techniques based on different association shapes of the predictor	117
5.4.2 Discrimination	129
5.4.3 Calibration plots	131
5.4.4 Clinical usefulness of statistical models under investigation	137
5.5 Discussion.....	145
5.5.1 General approach.....	145
5.5.2 Summary of main results	146
5.5.3 Limitations.....	150
5.5.4 Future work	151
5.6 Conclusions.....	152
Chapter 6	153
Examining the alcohol-hypertension association in type 2 diabetes patients using the UK Biobank	153
6.1 Background.....	153
6.1.1 Introducing DAGs	155
6.1.2 Specific study objectives	161
6.2 Subjects and Methods	162
6.2.1 UK Biobank participant’s characteristics	162
6.2.2 Alcohol intake estimates.....	165
6.2.3 Statistical analysis	168
6.2.4 Stratification analysis	171
6.2.5 Sensitivity analysis	172
6.3 Results.....	173
6.3.1 General characteristics.....	173
6.3.2 Model fits.....	176

6.3.3	Alcohol-hypertension association curves	182
6.3.4	Stratification analysis	186
6.3.5	Sensitivity analysis	195
6.4	Discussion	196
6.4.1	Summary of key findings	196
6.4.2	Challenges and limitations.....	201
6.4.3	Strengths and opportunities	205
6.4.4	Novelty	206
6.4.5	Future studies.....	207
6.5	Conclusions.....	207
Chapter 7	209
Discussion	209
7.1	Introduction.....	209
7.2	Synthesis and interpretation of key findings.....	212
7.2.1	The current practice of reporting and analysing continuous variables in observational studies	212
7.2.2	Implications of categorisation and comparison of alternative methods 213	
7.2.3	The differences between continuous and binary outcome models	218
7.2.4	Application study - Examining the alcohol-hypertension association in type 2 diabetes patients using the UK Biobank.....	221
7.3	Research implications and contribution to knowledge	223
7.4	Challenges and limitations	227
7.4.1	PhD scope/coverage	227
7.4.2	Methods applications	228
7.4.3	Data.....	229
7.5	Strengths and opportunities.....	231
7.6	Recommendations for future research	232
7.7	Publications arising from this PhD thesis	233
7.8	Conclusions.....	235
Appendix A	236
A.1	Data collection form	236
Appendix B	244
B.1	Example of the simulation codes used to generate data in chapter 4.....	244
B.2	Example of the simulation codes used to generate data in chapter 5.....	245

Appendix C	246
C.1 Additional tables in Chapter 4	246
Appendix D	273
D.1 Additional tables in Chapter 5	273
Appendix E	276
E.1 Justification of assumed associations in the DAG	276
Appendix F	284
F.1 Additional tables and figures in Chapter 6	284
Appendix G	290
G.1 Research protocol submitted at the UK Biobank	290
List of References	297

List of Tables

Table 2.1: A list of transformations defined by specific values of p_j	22
Table 2.2: Knots positions expressed in quantiles/percentiles of the exposure (x_1)	28
Table 3.1: Key findings showing the characteristics of categorisation amongst the exposure variables in epidemiological studies.....	44
Table 4.1: Nonlinear associations investigated with continuous outcome models.....	56
Table 4.2: The proportion of times the test of linearity was rejected in 1000 simulations under linear association datasets fitted using FPs and RCS models (assuming different number of observations and random error (σ)).	73
Table 5.1: Proposed linear and nonlinear logit functions used in the simulation to compare various approaches of handling the continuous predictor (x) when developing prognostic models.	107
Table 5.2: Comparison of optimal predictor and probability estimates obtained across 1000 simulations after fitting thresholds and quadratic association datasets using different modelling approaches.	126
Table 5.3: The coverage probabilities of 'true' optimal outcome events in 1000 simulations (replications) obtained in thresholds and quadratic datasets after fitting categorisation (CAT3 and CAT5), RCS, and FPs.	127
Table 5.4: The median estimates for the area under the ROC curve (AUC) and their 95% confidence intervals obtained after fitting FP, RCS, categorisation, and linearisation models in a simulation study replicated 1000 times. The reported estimates were obtained after applying these methods in log linear, thresholds and quadratic datasets.	130
Table 5.5: Comparison of net benefits and reduction of false positive results per 100 patients according to different statistical prediction models assuming various threshold probabilities.....	140
Table 6.1: The conversion units for estimating the contents of alcohol drinks	168
Table 6.2: The general characteristics of $n=23,842$ diabetes patients included in the study.....	175

Table 6.3: Summary statistics obtained after fitting various unadjusted logistic regression models	177
Table 6.4: The unadjusted and adjusted Odds ratios (ORs) of hypertension and their 95% confidence intervals obtained using the method of categorisation (CAT).	179
Table 6.5: The unadjusted and adjusted odds ratios (ORs) of hypertension & their 95% confidence intervals obtained from the best fitting linearisation (LIN), fractional polynomials - first order degree (FP1) and the restricted cubic spline with 3 knots (RCS3) models. The odds of hypertension was modelled as a function of alcohol consumption, g/day.....	181

List of Figures

- Figure 3.1: A detailed flow chart summarising the selection and identification process of eligible articles 37
- Figure 4.1: The proposed 'true' association functions used in the simulations to compare the properties and performances of fractional polynomials, restricted cubic spline, categorisation and linear regression techniques 58
- Figure 4.2: The estimated median RMSEs obtained when fitting (a) linear association shape (b) linear piecewise association shape (c) nonlinear piecewise association shape and (d) quadratic or U association shape using the linearisation, categorisation (CAT3 and CAT5), RCS, and FP models in a simulation study with 1000 replicates. Various noises (σ) were considered in the simulation with 200 observations (sample size). The 95% CI of each median RMSE are provided at the top of each bar..... 68
- Figure 4.3: The estimated median RMSEs obtained when fitting (a) linear association shape (b) linear piecewise association shape (c) nonlinear piecewise association shape and (d) quadratic or U association shape using the linearisation, categorisation (CAT3 and CAT5), RCS, and FP models in a simulation study with 1000 replicates. Various sample sizes were considered in the simulations where $\sigma=5.0$ 70
- Figure 4.4: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations (replicates) after fitting the linear and linear piecewise association datasets using the methods of categorisation (CAT3 and CAT5), linearization (LIN), fractional polynomials (FP), and restricted cubic splines (RCS). The results were taken from a sample with 200 observations and moderate noise, $\sigma=5.0$ 78
- Figure 4.5: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations (replicates) after fitting the nonlinear piecewise and quadratic or U association datasets using the methods of categorisation (CAT3 and CAT5), linearisation, fractional polynomials (FP), and restricted cubic splines (RCS). The results were taken from a sample with 200 observations and moderate noise, $\sigma = 5.0$ 80

- Figure 4.6: The outcome (at $c^*=20$) predicted by fitting the CAT3, CAT5, FP and RCS regression models in linear and nonlinear threshold datasets..... 84
- Figure 4.7: The outcome (at $c^*=20$) predicted by fitting the CAT3, CAT5, FP, and RCS regression models in Quadratic or U-shaped datasets. 87
- Figure 5.1: A decision tree diagram. The probabilities of disease and no disease are given by p and $1-p$ respectively. The values of true positive, false positive, false negative and true negative are given by A , B , C , and D respectively. 115
- Figure 5.2: Comparison of FP (green), RCS (blue), CAT3 (brown), CAT5 (red) and linearisation (orange) methods in a simulation where continuous predictor variable assume various shapes for prediction of event outcome. Median probability functions obtained with these methods after 1000 simulations are presented to compare them against true shapes (black) in linear, thresholds and quadratic datasets..... 119
- Figure 5.3: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations (replicates) after fitting linear and linear threshold association datasets using linearisation, categorisation, FP and RCS modelling approaches 122
- Figure 5.4: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations after fitting nonlinear threshold and quadratic or U association datasets using linearisation, categorisation, FP, and RCS modelling approaches 124
- Figure 5.5: Calibration plots of the event probabilities obtained in log odds models. The plots were obtained in a simulation with 1000 replicates comparing linearisation, categorisation, FPs and RCS approaches in linear and linear threshold datasets respectively. For each approach, the median observed probabilities of an event were plotted against the predicted probabilities. 134
- Figure 5.6: Calibration plots of the event probabilities obtained in log odds models. The plots were obtained in a simulation with 1000 replicates comparing linearisation, categorisation, FPs and RCS approaches in nonlinear thresholds and quadratic or U shaped datasets respectively. For each approach, the median observed probabilities of an event were plotted against predicted probabilities. 136

- Figure 5.7: The median predicted curves attained from 1000 simulations showing the Net Benefits of applying various statistical models (FP, RCS, CAT3, CAT5 and linearisation approaches) in linear predictor-outcome relationship datasets 138
- Figure 5.8: Comparison of various statistical prediction models showing the net reduction of false positives per 100 patients in the linear predictor-outcome datasets..... 141
- Figure 5.9: The median predicted curves attained from a simulation study (with 1000 replicates) showing the Net Benefits and Reduced false positive per 100 patients using various statistical models (FP, RCS, CAT3, CAT5 and linearisation approaches) in (a) linear threshold, (b) nonlinear threshold and (c) quadratic or U predictor-outcome relationship datasets. 144
- Figure 6.1: DAGitty schematic view of confounding adjustment for an alcohol-hypertension relationship 159
- Figure 6.2: The UK Biobank application and review process 163
- Figure 6.3: Histogram of alcohol consumption (in g/day) with normal curve..... 176
- Figure 6.4: The unadjusted and adjusted odds of hypertension (on log scales) estimated using categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption (in g/day). 184
- Figure 6.5: The adjusted odds of hypertension (on log scales) together with their 95% CIs estimated using the categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption, g/day. 185
- Figure 6.6 (a): The predicted probabilities of hypertension between patients on medication against non-medication users. The probability differences were computed based on the four categories of alcohol consumption. Figure 6.6 (b) shows the probabilities of hypertension between different categories of alcohol consumption against patient's age. The predicted probabilities were attained through the adjusted multivariable categorical model..... 187
- Figure 6.7 (a): The predicted probabilities of hypertension between patients on anti-hypertension against non-users across the different amount of alcohol consumption (g/day). Figure 6.6 (b) shows the difference in probabilities

between anti-hypertension medication users and non-users at different quantities of alcohol consumption (g/day). The predicted probabilities were attained through the adjusted multivariable linear model. 190

Figure 6.8: The predicted probabilities of hypertension across different levels of alcohol consumption (g/day) and age from the adjusted multivariable linear model 191

Figure 6.9: shows the difference in probabilities of hypertension between patients on antihypertensive medication and non-users at different levels of alcohol consumption (g/day) from the FP1 model with a power transformation of 0.5 and RCS with 3 knots. 193

Figure 6.10: The predicted probabilities of hypertension across different levels of alcohol consumption (g/day) and age from the adjusted FP and RCS models.. 195

Figure 7.1: Conceptual framework and linkages of the thesis results chapters 211

List of Abbreviations

AIC	Akaike's Information Criterion
AUC	Area under the receiver operating characteristic (ROC) curve
CI	Confidence interval
DAG	Directed acyclic graph
FP	Fractional polynomial
FP1	First order degree fractional polynomials
FP2	Second order degree fractional polynomials
GAM	Generalized additive model
GLM	Generalized linear model
LOESS	Locally weighted regression smoother
MCE	Monte Carlo error
MFP	Multivariable fractional polynomials
MRVS	Multivariable regression splines
OR	Odds ratio
RCS	Restricted cubic spline
RCS3	Restricted cubic spline model (with 3 knots)
RMSE	Root mean square error
SAZ	'Spike' at zero
SBP	Systolic blood pressure
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology

Chapter 1

Introduction

This chapter describes the purpose of the PhD, lists the main contributions and provides the thesis structure.

The chapter is divided into sub-headings listed below:

- 1) A brief background,
- 2) A brief motivation of the research thesis
- 3) A brief summary of objectives
- 4) Significance and contributions of the thesis
- 5) Thesis outline.

1.1 Background

The central objective of epidemiology is to assess the causes of disease amongst individuals identified with certain characteristics, such as diet, blood groups and smoking habits (Wakeford and McElvenny, 2007). Besides that, epidemiologists are also faced with the challenge to correctly define predictor-outcome relationships in their studies (Philippe and Mansi, 1998). When predictor variables are measured in continuous scales their relationship with the outcomes may be complex due to nonlinearity. Thus, nonlinear predictor-outcome relationships have to be considered during statistical modelling. Unfortunately, nonlinearity is often ignored when reporting the relationships between continuous predictors and the outcome variables in medical studies (Brenner and Blettner, 1997, Royston et al., 2006, Williams, 2011).

1.1.1 The importance of correctly specifying nonlinear relationships in epidemiology

In epidemiology, it is very important to accurately characterise the nonlinear relationships being investigated. This is because such relationship studies inform policies that influence individuals' health outcomes. For instance, where the nonlinear predictor-outcome relationships are well established, health practitioners may identify individuals who could benefit from some targeted interventions.

1.1.2 Continuous predictor variables are predominately analysed as categorical measures in the current publications

An assessment performed by Turner et al. (2010), found a common practice of categorisation amongst medical studies working with continuous predictor variables. Turner et al. (2010) reviewed observational studies published in December 2007 and January 2008 from five medical journals and found 86% of articles with categorised continuous predictors or risk factors. A similar trend was also observed in a different survey conducted in the early 2000 - reporting a proportion of 84% amongst epidemiological studies (Pocock et al., 2004). Reflecting on these findings, the incidence of categorisation could still be large. This thesis will start by demonstrating the current extent for categorisation of continuous variables in medical studies.

1.1.3 Why is categorisation popular in medical studies

The potential reasons and suggestions for categorisation include the following:

Clinical audience to enable decision making: In clinical studies, categorisation of continuous variables may be performed for the sake of the audience to enable clinical decisions (Baneshi and Talei, 2011, Sauzet and Peacock, 2014). For example, the body mass index (BMI) could be grouped into four categories such that $18.5 \leq \text{BMI} < 25.0$, $25.0 \leq \text{BMI} < 30.0$, $30.0 \leq \text{BMI} < 40.0$, and $\text{BMI} \geq 40.0 \text{ kg/m}^2$ represent individuals with healthy weight, overweight, obese and severely obese, respectively (NICE, 2014).

This kind of grouping offers simple risk classifications that are easily understood by clinicians for decision making or treatment recommendations. Although useful in clinical settings, such simplicity is unnecessary in medical research since it is gained at a cost (Royston et al., 2006). The practice of categorising continuous variables does not make use of within-category information (Bakhshi et al., 2008, Bakhshi et al., 2012).

The existence of distinct categories within variables of interest: Evidence of distinct groups within continuous variables has been suggested as another reason for categorisation in medical studies (MacCallum et al., 2002). For example, researchers investigating the influence of coffee consumption (g/day) on health outcomes may categorise the intake using two distinct groups (coffee drinkers versus nondrinkers) in their analysis; where nondrinkers are participants with 'zero' responses and coffee drinkers are 'nonzero' values. Although these two groups exist, it is important to recognise that analysts employing such categories will lose all the information showing some variations amongst the coffee drinkers.

Lack of awareness on the potential consequences of categorisation: Since the categorisation method is popular in medical studies, many researchers may be using it unconsciously unaware of the associated problems (Royston et al., 2006).

Lack of awareness on the appropriate methods of analysis: Due to inexperience and lack of knowledge on appropriate methods for handling continuous variables during statistical analysis, researchers may categorise continuous variables to perform association tests using t and χ^2 statistics (MacCallum et al., 2002, Royston et al., 2006). This situation is mostly common in studies investigating the associations between multiple independent variables and interaction terms. Hence, misleading (or biased) measures of effect sizes and spurious associations are likely to occur in such scenarios (Breitling and Brenner, 2010, Williams, 2011). Regression models using methods such

as fractional polynomials (Royston and Altman, 1994) and restricted cubic spline models have been suggested for studies working with continuous variables to incorporate nonlinear relationships, interaction terms and multiple independent variables. In the absence of nonlinearity, the simple linear regression models may also be useful. However, it is more convincing to assume that many medical researchers are not sufficiently familiar with the application of these regression models. Evidence provided by Turner et al. (2010) shows a high incidence of categorisation amongst medical studies.

Beliefs that categorisation improves reliability: Other researchers believe raw measures of continuous data provide unreliable and imprecise information. Hence, they opt for categorisation with the belief that it would produce reliable measures (Cohen, 1983, Royston et al., 2006). This kind of reasoning is incorrect; categorisation does not refine the original measurements but worsens reliability substantially – weakening the correlations or associations between variables (MacCallum et al., 2002).

Due to statistically significant associations after categorisation: According to Royston et al. (2006), researchers may also justify their reasons for categorisation when significant relationships exist between categorised continuous variables. Such arguments are based on the assumption that categorisation is a conservative method - with reduced statistical power due to loss of information. Therefore, the underlying relationships should be strong (MacCallum et al., 2002). However, this reasoning is not entirely true, statistical significant tests are mainly influenced by sample sizes, levels of α and the distribution of sampling errors.

Based on the arguments above, many of the perceived reasons for categorisation are misguided hence this practice might be unnecessary. There exist several challenges

associated with the practice of categorising continuous data during statistical modelling. The next section highlights some of the challenges.

1.1.4 Challenges of categorising continuous variables during statistical modelling

The potential challenges of categorising continuous variables during statistical modelling include (1) deciding the choice of cut points, (2) minimising information loss, and (3) the issue of bias due to incomplete adjustment for confounding when using categorised variables.

When categorising continuous variables, researchers must decide how many categories to choose and where to place the category cut points. However, deciding these choices is not easy - there exist some uncertainties due to lack of knowledge of the most appropriate approach (O'Brien, 2004, Royston et al., 2006). Briefly, some of the methods available for establishing categories include using the median as the cut point - an extreme form of categorisation that yields two categories in the data (known as dichotomisation in the literature). For more than two categories, percentiles/quantiles could be used as cut points to decide categorical boundaries for ordered groups. Alternatively, analysts could adopt equally spaced categories such as 0-5, 5-10, 10-15 and 15+ or use 'zero/never' categories as the reference group amongst the variables with large portions of zeros. The former is common amongst skewed variables, where the proportion of zeros is relatively large and comparable to non-zero values. Such variables are common in medicine and often referred to as 'spike' at zero (SAZ) variables (Jenkner et al., 2016, Lorenz et al., 2017). So far the methods defined are data dependent; this makes it difficult to compare results between studies. From a statistical point of view, the cut point should be chosen a priori to avoid data-driven inferences. However, lack of knowledge on the appropriate cut points in many medical settings

makes this difficult. Dichotomisation or categorisation may lead to a misleading measure of effect size if the cut points are not properly chosen (Cohen and Chen, 2009).

The practice of categorisation simplifies relationships in the data restricting analysts to work with step functions which may be inappropriate for the final models (Royston et al., 1999). Thus, the final models may overlook potential curves in the data resulting in biased functions. The amount of bias in the estimated function could be large depending on the sample sizes and information loss (MacCallum et al., 2002).

In regression models controlled for confounding, the practice of categorisation does not yield optimal estimates - due to inefficient use of within-category information. The categorisation method has been demonstrated to be inadequate when controlling for continuous confounders, with crudely categorised covariates resulting in misleading estimates (Brenner and Blettner, 1997, Brenner, 1998).

1.2 A brief summary of the research motivation and rationale

The existing evidence provided by Turner et al (2010) shows high incidence of categorisation of continuous data in medicine. The practice of categorisation may be unnecessary; many of the perceived reasons are misguided, information is lost and there exist serious challenges such as choosing the cut points. The alternative methods of analysing continuous data such as linear regression, fractional polynomials (FPs) and restricted cubic spline (RCS) are available but they are rarely used. The reasons why such methods are not widely used could be lack of examples in their applications, differing views on the most appropriate alternative approach, perceived difficulties in application and interpretation of estimates.

This research will be demonstrating the current extent of categorisation in medicine, and the performances and applications of the linear, FPs and RCS regression

models against the practice of categorising continuous data - focusing in the area of epidemiology. The intention is to encourage and promote the use of the alternative methods in medical studies especially amongst clinicians with little statistical background for analysing continuous data. To reach out for this audience and other researchers interested in the applications and performances of these methods; the findings in the thesis will be demonstrated by conducting a new survey of current statistical methods used for analysing continuous data, novel plausible simulations (assuming continuous and binary outcomes) and real application studies. The objectives of the thesis are summarised in the next section.

1.3 Summary of objectives

Objective 1 - Cross sectional survey of research methods in current use

To conduct a cross sectional survey on the current practice of reporting and analysing continuous variables in observational studies

Following the STROBE guidelines (Von Elm et al., 2007) aimed at improving the reporting in observational studies; a new survey of current statistical methods used to analyse and report continuous variables in medical research will be needed. The previous findings are based on the surveys that reviewed studies that were published before (Pocock et al., 2004) and immediately after (Turner et al., 2010) the STROBE guidelines. Thus, many researchers were not aware of the STROBE guidelines at the time of their publications. To bridge this gap, the first research objective will be to conduct a new survey on the current practice of reporting and analysing continuous variables in observational studies. Details of the survey and its findings are presented in Chapter 3.

Objective 2 – Simulations based on normal error models

To investigate using continuous outcome models - simulation based study; the performances and properties of categorisation, linearisation, FP, and RCS approaches assuming plausible exposure-outcome relationships in epidemiology.

A simulation based study will be set out as the second objective to investigate the performances and properties of categorisation, linear, FP, and RCS regression models using pre-defined scenarios. Motivated by application of these methods to real data, the simulations will be covering several exposure-outcome associations in epidemiology using the reported alcohol-blood pressure relationships as example scenarios. The alcohol-blood pressure example scenarios are meant to assess the models directly in datasets where the distribution of the exposure and shapes of the underlying exposure-outcome relationships are known. The performance measures that will be used for model evaluation include the RMSE, type I errors, statistical power, coverage probability of turning points and the ability of each model to recover the ‘true’ exposure-outcome functions in the simulations. Further details of the simulation and the associated findings are provided in Chapter 4 of this thesis.

Objective 3 – Simulations based on binary outcome models

To investigate using binary outcome models - simulation based study; the performance and choice of categorisation, linearisation, FP and RCS approaches for handling continuous predictors in prognostic models.

The third objective will be addressed using a simulation study with binary outcome models - commonly found in medical studies after dichotomisation of underlying continuous outcomes, e.g. blood pressure and hypertension. The intention will be to examine the influence of different approaches used for handling continuous

predictors when developing prognostic (or predictive) models. Therefore, this chapter differs from the previous simulations; not only on the type of outcome, and being based on prognostic models but also that it forms an example of categorisation for both the outcome and exposure variables. Moreover, prognostic models are rarely investigated in epidemiology, so this will also be an opportunity to consider other suitable performance measures for evaluation. For instance, some of the performance measures that will be used to evaluate the performance and properties of various prognostic models in the simulations include the c-index scores, calibration plots, and decision analysis curves. The findings of this simulation study are provided in Chapter 5.

Objective 4 – Application study using the UK Biobank data

To investigate the association between alcohol consumption and hypertension in patients with type 2 diabetes using categorisation, linearisation, FP and RCS modelling approaches.

The application of categorisation, linearisation, FP and RCS modelling approaches in this area has been identified as appropriate for investigations for several reasons including the conflicting evidence of guidelines regarding the associations between alcohol consumption and hypertension, previous evidence of use and common practice of categorisation in the area, importance to public health and the interest on identifying whether alcohol consumption thresholds exist in the patients with type 2 diabetes. The specific objectives in this chapter are as follows:

- i. To investigate the association between alcohol consumption and the odds of hypertension in patients with type 2 diabetes adjusting for selected confounding variables identified by the use of a directed acyclic graph (DAG).

- ii. To investigate effect modification of age and antihypertensive medication use in the adjusted multivariable alcohol-hypertension model.

Explanatory modelling (or causal inference) is very popular in epidemiology thus the reason to consider it for application instead of prognostic modelling (or predictive analysis). The details of study are presented in Chapter 6. The chapter demonstrates the application of the alternative methods covering issues of nonlinearity, confounding, interactions and interpretations - for medical researchers who might find it difficult to deal with such issues in their studies when applying these methods.

1.4 Significance and contribution to knowledge

The thesis will provide the following as novel contributions:

- i. A new piece of research providing the reporting of continuous variables in medical studies according to the STROBE guidelines. This will be the first update covering research articles that were published several years after the STROBE guidelines. The results of this update have relevance to authors and readers working with observational studies. Key issues necessary for improvement when reporting and analysing continuous variables will be highlighted in this research for the purpose of promoting and preserving scientific knowledge for synthesis and clinical decision making.
- ii. A novel simulation study investigating the properties and performances of categorisation, linearisation, FP, and RCS models on plausible relationships found in epidemiology. The simulations will be based on continuous outcome models to provide an inferential guide in similar situations – focusing on the ability of these methods to characterise the ‘true’ relationships in the data. The simulation will be covering nonlinear associations with turning points for

estimation. The ability of these methods to accurately predict turning points is uncertain. Therefore, this will be the first study conducting an assessment of these methods - under different simulation conditions taking into account issues of turning points. Ultimately, this research should also provide guidance to medical researchers on the applications of these methods.

- iii. A novel simulation study examining the influence of categorisation, linearisation, FP and RCS approaches used for handling continuous predictors when developing prognostic (or predictive) models with binary outcomes. Binary outcomes are common in epidemiology and prognostic models are rarely investigated. However, with the emerging field of machine learning, predictive analysis may be on the rise. Predictive analytics go hand-in-hand with machine learning where big data – large volumes of raw structured, semi structured and unstructured data are used to estimate or predict future outcomes. Due to spurious correlations and possible biases during data collections, the predicted outcomes from these data sources need to be validated for accuracy to guard against overly optimistic and exaggerated claims. This is where this simulation study comes in; different methods for developing prognostic models are demonstrated and evaluated using novel measures for the purpose of providing guidance to medical researchers working in the same area.
- iv. An investigation assessing the association between alcohol consumption and hypertension in patients with type 2 diabetes using the UK Biobank. Different association functions fitted with the methods of categorisation, linearisation, FP and RCS will be compared to demonstrate the application of these methods in real dataset characterised by issues of nonlinearity, confounding and

interactions. This will be the first study to demonstrate and compare the application of these methods in the UK Biobank. Clinically, this is also a novel study that should inform strategies for managing and controlling alcohol induced hypertension amongst type 2 diabetes patients taking into consideration (a) the efficacy of antihypertension medication use and (b) the severity of alcohol drinking in different age groups.

1.5 Outline of the thesis

The chapters in this research thesis are structured as follows:

Chapter 1 is an introductory chapter that gives the research background, research motivation and rationale, summary of objectives, significance and contribution to knowledge and finally the thesis outline.

Chapter 2 summarises the methods used in this thesis. This chapter presents the choice and descriptions of modelling approaches including categorisation, linearisation, FP, and RCS as applied to objectives 2, 3 and 4.

Chapter 3 present a survey on the current practice of reporting and analysing continuous variables in observational epidemiological studies (Objective 1).

Chapter 4 and Chapter 5 are simulation studies investigating the performance of methods discussed in Chapter 2. These chapters address objective 2 and 3 respectively. In each chapter, details covering the gap in the literature, specific objectives, and procedures on how the simulations were set up are provided. Other contents include descriptions of performance measure, results, finding, discussions, and conclusions.

Chapter 6 investigated the association between alcohol consumption and hypertension in patients with type 2 diabetes using the UK Biobank dataset. In the analysis different modelling techniques were applied for comparison.

Chapter 7 is a conclusion chapter. The chapter reflects on key findings, research implications, challenges and limitations, strengths and opportunities and the recommendations for future studies.

Chapter 2

Statistical approaches for modelling exposure-outcome relationships

This chapter provides an overview of statistical approaches for analysing the relationships between continuous exposure and outcome variables in epidemiology. This chapter also explains the decisions informing the choice of methods used in this thesis and their general applications.

The chapter is divided into sub-headings listed below:

- 1) An overview of statistical methods
- 2) The details of statistical methods used in the thesis
- 3) The statistical modelling framework of methods used in the thesis
- 4) A conclusion

2.1 Overview of statistical approaches for analysing exposure-outcome relationships in epidemiology

In epidemiology, the potential influence of continuous exposures on health outcomes can be ascertained using various statistical approaches that depend on the assumptions of categorisation, linear and nonlinear relationships (Schmidt et al., 2013a). The simplest method based on the assumption of categorisation involves categorising the exposure into two or more group categories, creating dummy variables and then reporting the outcome of each category (using one group as the reference) or fitting a linear trend over the ordered categorical exposure (Maclure and Greenland, 1992, Figueiras and Cadarso-Suárez, 2001). Alternatively, the analysts may keep the exposure as a continuous variable and assume a linear relationship between the exposure and

outcome during statistical modelling (Royston et al., 1999, Figueiras and Cadarso-Suárez, 2001). When the relationship is complex, nonlinearity may be assumed and explored using different methods that include:

- i. A polynomial parametrization of the exposure. In this method, the exposure is transformed and expressed as a quadratic or cubic term to report nonlinear relationships (Schmidt et al., 2013a). The advantage of this method is that it offers simpler function forms that define the overall relationships between the exposure and the outcome. On the contrary, simple transformations offer limited flexible functions especially at the tails of the exposure distribution whilst high order degree polynomials are susceptible to artefacts and over-fitted functions (May and Bigelow, 2005, Schmidt et al., 2013a).
- ii. Fractional polynomial approach. The fractional polynomials (FPs) offer some improvement on the polynomial parametrization approach. The FP approach was first developed by Royston and Altman (1994) to allow a combination of polynomials and logarithmic functions. The main advantage of the FP approach is that it offers more flexibility and wider set of functional forms for the relationship between the exposure and the outcome variables (Royston and Altman, 1994, Royston and Sauerbrei, 2008).
- iii. Splines based approaches. Like FPs, splines extend from the polynomial parametrization approach. Splines are piecewise functions whose ‘pieces’ are polynomials defined over the adjacent intervals. The junction between two intervals is called the ‘knot’ and the number of knots (specified by the user) ranges between 3 and 7 (Desquilbet and Mariotti, 2010). Linear, quadratic and cubic splines are typical examples of these functions but cubic splines are common (Steenland and Deddens, 2004, Schmidt et al., 2013b). Quadratic or cubic splines may be used to improve the smoothness of linear

spline functions when sudden change of slopes occurs between adjacent intervals. But, both the quadratic and cubic splines may behave poorly at the tails (Schmidt et al., 2013b). To overcome this problem, restricted cubic splines (cubic splines constrained to be linear at the tails) have been developed to improve the fitted curves (Desquilbet and Mariotti, 2010). In addition, there exist other different types of spline functions that include the B-splines, smoothing splines, penalized splines and thin-plate splines (Hastie and Tibshirani, 1990, Schimek and Turlach, 2000, Schmidt et al., 2013b). These functions offer non-interpretable parameter estimates hence some interpretations can only be made from the fitted curves (Steenland and Deddens, 2004).

- iv. Nonparametric approaches. Nonparametric methods relax the assumption of parametrization of the exposure emphasizing on the graphical approach to explore nonlinear relationships between the exposure and the outcome variables (May and Bigelow, 2005). There are no assumptions made on the nature (or shape) of the existing relationships or the distributions of the exposure and outcome variables; the estimated nonlinear functions are data driven (Keele, 2008). LOESS (for locally weighted regression) plot is one basic example of the available nonparametric approaches (Cleveland and Devlin, 1988). LOESS is a nonparametric procedure for generating a moving weighted average of the outcome variable within specified local regions of the exposure (x -axis). The weights reduce as one move from the center of each specific region and become zero beyond the range of the region (Cleveland and Devlin, 1988, Steenland and Deddens, 2004). Examples of common fits include the linear (first order LOESS) or quadratic (second order LOESS) models. However, the LOESS plots are only excellent for

exploring nonlinearity in single exposure-outcome relationship studies. Their use in multivariable settings is limited (May and Bigelow, 2005). Splines are also classified under the nonparametric approaches (Keele, 2008, Schmidt et al., 2013b). For estimation, both the LOESS and Spline functions can be incorporated into the standard linear models using the Generalized Additive models (GAMs) procedure (Hastie and Tibshirani, 1990, Beck and Jackman, 1998). However, the GAMs are computationally expensive. Furthermore, fitted GAMs can be harder to communicate than a vector of parameter estimates and their standard errors (Beck and Jackman, 1998).

2.2 Statistical methods used in the thesis

With the variety of statistical approaches listed above, this PhD thesis focused on the method of categorisation, linearisation (i.e. assuming linearity between continuous exposure and outcome variables), fractional polynomials (FP) and restricted cubic splines (RCS) methods. The methods of linearisation and categorisation were chosen for assessment in more details because traditionally, the relationships between the continuous predictors and outcome variables are assumed to be linear or the predictors are grouped and entered into the model as dummy variables. A review in observational studies showed 86% of categorisation amongst published articles that investigated continuous predictors or risk factors (Turner et al., 2010). The FPs and RCS approaches are considered powerful and flexible in fitting both complex and linear associations (Royston and Altman, 1994, Desquilbet and Mariotti, 2010). The other alternative approaches (polynomials with powers < 2) have limitations in fitting complex nonlinear curves especially at their tails whilst those with powers (> 3) are susceptible to artefacts and over-fitted models (Schmidt et al., 2013b). The other reasons for choosing to investigate and compare RCS and FP are as follows:

- i. They are readily available for implementation in most statistical programs
- ii. Comparisons between RCS and FP are lacking and little is known about their results and properties (Sauerbrei et al., 2006).
- iii. The features of the FP & RCS functions are intriguing when imagining the potential differences between the two models. When applied in nonlinear datasets, the RCS function might be able to get around the data at the extreme regions due to the linearity constraint at their tails. On the other hand, the FP functions offer flexibility - constrained by the selected set of powers available for use in such models.

2.2.1 Modelling scenario

To further describe the application of categorisation, linearisation, FPs and RCS methods, consider the following modelling scenario presented by Cumsille et al (2000):

Let assume that the interest is to study the relationship between one continuous outcome (y) and one continuous exposure (x_1), controlling for confounding variables (both continuous and categorical) $x_2 \dots x_n$. For simplicity, suppose the random vector $\mathbf{V} = (y, x_1, \dots, x_n) = (\mathbf{Y}, \mathbf{X}')$ has a multivariate normal distribution such

that $\mathbf{V} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where mean vector $\boldsymbol{\mu} = \begin{pmatrix} \mu_y \\ \mu_1 \\ \cdot \\ \cdot \\ \cdot \\ \mu_n \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}$, and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \sigma_{y1} & \cdot & \cdot & \cdot & \sigma_{yn} \\ \sigma_{y1} & \sigma_1^2 & \cdot & \cdot & \cdot & \sigma_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{yn} & \sigma_{1n} & \cdot & \cdot & \cdot & \sigma_n^2 \end{pmatrix} = \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{yx} & \Sigma_{x1xn} \end{pmatrix}. \quad \text{Eq. 2. 1}$$

Given the modelling scenarios above, the methods of categorisation, linearisation, FPs, and RCS are described in sections 2.2.2 to 2.2.5 below.

2.2.2 Categorisation/Dichotomisation

Based on the scenario in 2.2.1, suppose the exposure variable, x_1 is dichotomised (or categorised into two groups) using a cut point, c such that the dichotomised variable Z is created based on the following rule:

$$Z = \begin{cases} 1 & \text{if } x_1 \leq c \\ 0 & \text{if } x_1 > c \end{cases}, \quad \text{Eq. 2. 2}$$

where c is the cut point.

Using the new variable, Z , the unadjusted exposure-outcome model could be fitted as follows:

$$f(Z) = E(y|Z) = \beta_0 + \beta_1 Z, \quad \text{Eq. 2. 3}$$

where $f(Z) = E(y|Z)$ is the expected function of the outcome (y), β_0 is the estimated outcome when $Z = 0$, and β_1 represent the change in the outcome when Z takes the value 1.

Alternatively, x_1 can be grouped or divided into k categories resulting in the process known as polychotomisation (categorisation of a continuous variable into many groups). For example, suppose Z takes value 1 if the exposure is in the k^{th} category and 0 otherwise for $k = 1, 2, \dots, K - 1$. Then, the unadjusted exposure-outcome relationship assuming k categories could be written as follows:

$$f(Z) = E(y|Z_k) = \beta_0 + \beta_1 Z_1 + \dots + \beta_k Z_k = \beta_0 + \sum_{k=1}^{K-1} \beta_k Z_k, \quad \text{Eq. 2. 4}$$

where β_k represent the associated outcome value of an exposure at k category.

For multivariable models incorporating confounders, the adjusted exposure-outcome functions could be written as follows when dichotomised:

$$f(Z, X) = E(y|Z, X) = \beta_0 + \beta_1 Z + \beta_2 x_2 + \dots + \beta_n x_n = \beta_0 + \beta_1 Z + \sum_{i=2}^n \beta_i x_i, \quad \text{Eq. 2. 5}$$

where β_i 's are parameters associated with each confounding variable in the model.

Similarly, when the exposure was categorised into k categories (polychotomisation), the adjusted model explaining the exposure-outcome relationship could be expressed as follows:

$$f(Z_k, X) = E(y|Z_k, X) = \beta_0 + \sum_{k=1}^{K-1} \beta_k Z_k + \sum_{i=2}^n \beta_i x_i \quad \text{Eq. 2. 6}$$

Detailed information about dichotomisation (as presented above) could be found in the following references; Cumsille et al., (2000), Gustafson and Le (2002) and Natarajan (2009).

2.2.2.1 Strengths and challenges of categorisation

The main strength of categorisation is based on its simplicity - avoids strong assumptions about the exposure-outcome relationships. The challenge is that such simplicity is achieved at a cost of throwing away some information. Thus, categorisation may lead to some reduction in effect sizes and statistical power compared to models that keep continuous variables in the analysis (Cumsille et al., 2000, Williams, 2011). The other problem faced by researchers is deciding on the categories to use in the analysis; choosing the cut points and decisions on whether continuous variables should be transformed into dichotomous (binary) or ordinal scale measurements (Cumsille et al., 2000, Figueiras and Cadarso-Suárez, 2001).

2.2.3 Linearisation

From the modelling scenario provided in section 2.2.1, a typical regression function assuming linearity on the exposure-outcome relationship could be written as follows:

$$f(X) = E(y|x_1) = \beta_0 + \beta_1 x_1, \quad \text{Eq. 2. 7}$$

where $f(X) = E(y)$ is the estimated function of the outcome, β_0 is the predicted value of the outcome when the exposure is zero and β_1 is the estimated value of outcome per unit increase of the exposure (x_1).

In multivariable regression, the exposure-outcome relationship could be adjusted for other covariates such that

$$f(X) = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n = \beta_0 + \sum_{i=1}^n \beta_i x_i \quad \text{Eq. 2. 8}$$

Further theories of linear regression models could be found elsewhere (Marill, 2004, Schneider et al., 2010, Schmidt and Finan, 2017).

2.2.3.1 Strengths and challenges of linearisation

Linear models are simple to implement and interpret. However, the assumption of linearity imposes some restrictions on the data and forces linear parametric models that frequently do not fit nonlinear data closely (Figueiras and Cadarso-Suárez, 2001). Thus, appropriate relationships maybe missed if deviations from the assumption of linearity are strong.

2.2.4 Fractional polynomials (FPs)

Fractional polynomials (FPs) were first proposed by Royston and Altman (Royston and Altman, 1994) for modelling families of curves. An FP function of degree m defining the unadjusted exposure-outcome relationship for the modelling scenario in section 2.2.1 is given as follows:

$$FP_m(X) = \beta_0 + \sum_{j=1}^m \beta_j f_j(x_1), \quad \text{Eq. 2. 9}$$

where m is a positive integer, $\beta_0 \dots \beta_m$ are regression parameters, and $f_j(x_1)$ is the Box and Tidwell (1964) transformation defined by;

$$f_j(x_1) = \begin{cases} x_1^{p_j} & \text{if } p_j \neq 0 \\ \ln(x_1) & \text{if } p_j = 0 \end{cases}, \quad \text{Eq. 2. 10}$$

with the constraint $x_1 > 0$, so that all the transformations are possible.

The power terms, p_j in Eq. 2.10 are taken from a restricted set of integer and non-integer values suggested by Mosteller and Tukey (1977, chapter 4) for general curve fitting. This set is given by $p_j \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$, where x_1^0 denote $\ln(x_1)$. The set offer a wide range of curve shapes and transformations often used by applied researchers, such as the log ($p_1 = 0$), reciprocal ($p_1 = -1$) and square root ($p_1 = 0.5$). Table 2.1 provide a list of transformations defined by the specific values of p_j (assuming $m = 1$ in Eq. 2.9).

Table 2.1: A list of transformations defined by specific values of p_j .

FP powers (p_j)	Functions	FP Equations
-3	Inverse cubic	$FP_1(x_1) = \beta_0 + \beta_1(1/x_1^3)$
-2	Inverse square	$FP_1(x_1) = \beta_0 + \beta_1(1/x_1^2)$
-1	Inverse	$FP_1(x_1) = \beta_0 + \beta_1(1/x_1)$
-0.5	Inverse square root	$FP_1(x_1) = \beta_0 + \beta_1(1/\sqrt{x_1})$
0	log	$FP_1(x_1) = \beta_0 + \beta_1 \ln(x_1)$
0.5	Square root	$FP_1(x_1) = \beta_0 + \beta_1 \sqrt{x_1}$
1	Linear	$FP_1(x_1) = \beta_0 + \beta_1 x_1$
2	Square	$FP_1(x_1) = \beta_0 + \beta_1 x_1^2$
3	Cubic	$FP_1(x_1) = \beta_0 + \beta_1 x_1^3$

From Table 2.1, additional values of p_1 between -3 and 3 such as $p_1 = 1/4$ are possible but such transformations are rare. The curves produced by such transformation values are very similar to those in their neighbourhood (i.e. p_j). Hence, p_j provide a set of powers familiar to many applied researchers (Long and Ryoo, 2010).

From Eq. 2.10, repetition of powers is also allowed. However, this is only possible when $m \geq 2$. For instance, when $m = 2$, and $p_1 = p_2$ the FP model is written as follows:

$$FP_2(X) = \beta_0 + \beta_1 x_1^{p_1} + \beta_2 x_1^{p_1} \ln(x_1) \quad \text{Eq. 2. 11}$$

Otherwise, when $m = 2$, and the powers are not repeating (i.e. $p_1 \neq p_2$) the FP model is written as follows:

$$FP_2(X) = \beta_0 + \beta_1 x_1^{p_1} + \beta_2 x_1^{p_2} \quad \text{Eq. 2. 12}$$

Based on the descriptions above, fractional polynomials cover a wide range of shapes and transformations including conversional polynomials. For example, the first-order degree FP ($m = 1$) with $p_1 = 1$ gives a linear polynomial model whilst the second-order degree FP ($m = 2$) with $p_1 = 1$ and $p_2 = 2$ is a quadratic function. The regression parameters, β_j and polynomial powers p_j , $j = 1 \dots m$ for such functions are obtained using a closed-test procedure described by Sauerbrei & Royston (Sauerbrei and Royston, 1999) and Ambler & Royston (Ambler and Royston, 2001).

2.2.4.1 First and second order degree fractional polynomial functions

To be practically relevant, fractional polynomial functions with $m \leq 2$ are fitted. In reality, FP models with $m > 2$ would rarely be required for exposure-outcome relationship studies (Royston and Altman, 1994, Royston and Sauerbrei, 2008). For a permissible set of FP powers, $p_j \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ there exist 44 possible combinations of models (8 when $m = 1$ and 36 when $m = 2$) where the best fit is

chosen as the one with the lowest deviance ($-2 \times \log\text{-likelihood}$). Furthermore, the FP family with $m \leq 2$ has a richer set of functions. For example, Figure 4.5 (Royston and Sauerbrei, 2008, p.78) shows a variety of curves available when $m = 2$ and $p = (-2, 2)$ for different β values. The variety of curves attained for a single pair of powers, $p = (-2, 2)$ when $m = 2$ suggests the β coefficients are non-interpretable - one disadvantage of working with FPs. However, Royston and Altman (1994) advise analysts to interpret the FP curves instead of predicted parameters.

To adjust for confounding, the multivariable fractional polynomial (MFP) approach (combination of functional selection form for continuous covariates and backward elimination process of variable selection) has been suggested (Sauerbrei and Royston, 1999, Royston and Sauerbrei, 2005). However, application of such automated procedures have been criticised; the selection of variables through stepwise methods is known to produce biased results with inflated p-values and standard errors (Blanchet et al., 2008). To deal with the controversy of variable selection, realist ontology of causation of health outcome and covariates could be constructed to identify ‘minimal’ sufficient set of confounders for model adjustment. This approach incorporates clinical and epidemiological knowledge to achieve models with public health and scientific relevance.

2.2.4.2 The strengths and weaknesses of FPs

The advantages of FPs is that it avoids cut points and make full use of covariate information (Greenland, 1995b). Furthermore, FP functions of $m \leq 2$ offers a wide range of flexible shapes (e.g. monotonic and asymptotic curves) that could improve the reporting of nonlinearity in exposure-outcome studies (Royston and Altman, 1994).

The applications of FPs also have some limitations that need to be pointed out. Firstly, FP powers cannot be used with zero or negative values. This is because the FP

powers are based on natural logarithms (terms of the form $x^r [\ln(x)]^j$ are included in the family of FP curves) (Greenland, 1995b). If the exposure (x) assumes zero values, the FP function would be sensitive with unstable tails at zero levels of the exposure. In epidemiology, such variables are known as ‘spike’ at zero (SAZ) variables (Lorenz et al., 2017). Examples of such variables include tobacco consumptions and occupational exposures (e.g. asbestos). In FP models, the ‘spike’ at zero can be dealt with by shifting the origin of the exposure making it nonzero (by adding a constant δ) (Royston and Altman, 1994, Royston and Sauerbrei, 2008). However, such transformations have been criticised because the choice of the constant δ could influence the results of the FP models (Ambler and Royston, 2001). Ignoring the ‘spike’ at zero level of the exposure may also be a sensible choice of dealing with this problem since there are no biological interpretations associated with such behaviour (Royston and Sauerbrei, 2008). For non positive values, Royston and Altman (1994) recommend adding a positive number to force the exposure to be positive. The discomfort with this approach is that it is likely to introduce new parameters into the FP functions thus influencing the results (Greenland, 1995b). Secondly, the FP1 is based on limited powers to detect nonlinearity; the best function select powers from a candidate set of $p_j \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Although flexible than conventional polynomials, the FP1 models could still be misleading (in both the shapes and estimates) due to this limited choice of powers (Royston et al., 1999). Finally, the FP functions offer non interpretable parameters of the curves. However, the results of such models can be reported using graphs or tables, although such approaches have its problems. Visual presentations make it difficult for researchers to conduct a meta-analysis in similar studies. In addition, subjective biases are likely to be introduced by analysts when interpreting the results (Greenland, 1995b, Royston et al., 1999).

2.2.5 Restricted cubic splines (RCS)

Restricted cubic splines (RCS) are a special type of cubic splines (CS) constrained further to be linear above the last knot and below the first knot (Durrleman and Simon, 1989, Govindarajulu et al., 2009). The expression for RCS function with k knots, $t_1 < \dots < t_k$ can be written as follows:

$$f(X) = \beta_0 + \beta_1 x_1 + \sum_{j=1}^{k-2} \theta_j f_j(x_1), \quad \text{Eq. 2. 13}$$

where $f_1(x_1) \dots f_{k-2}(x_1)$ are cubic terms such that;

$$f_j(x_1) = (x_1 - t_j)_+^3 - \frac{(x_1 - t_{k-1})_+^3 [t_k - t_j]}{[t_k - t_{k-1}]} + \frac{(x_1 - t_k)_+^3 [t_{k-1} - t_j]}{[t_k - t_{k-1}]}, \quad j = 1 \dots k - 2 \quad \text{Eq. 2. 14}$$

The notation $(\cdot)_+$ is known as the ‘positive part’ function. It can be written as $(a)_+ = \max(0, a)$. It retains the maximum value between 0 and $a = (x_1 - t_{k-1})$ or $a = (x_1 - t_k)$ respectively.

From the RCS function above;

- i. The function, $f(X)$ is linear in parameters suggesting that standard procedures can be used for statistical inference. For example, when the interest is to estimate the unknown smooth function, $f(\cdot)$, a test that $\beta_1 = \theta_1 = \dots = \theta_{k-2} = 0$ represent a constant function and a test that $\theta_1 = \dots = \theta_{k-2} = 0$ is similar to a linearity test.
- ii. The estimated parameter estimates are hard to interpret thus the estimated $f(X)$ is the main output.
- iii. Although the $f(X)$ looks complicated to implement, readily available statistical packages (e.g. R and Stata software) can be used to estimate the function.

Additional information on the development of restricted cubic spline regression models may be found in the following references; Durrleman and Simon (1989), Heinzl and Kaider (1997) and Desquilbet and Mariotti (2010).

For multivariable models, a similar procedure as the MFP was made available for users working with regression splines (RS). The closed-test procedure was modified for spline functions resulting in a new algorithm called the multivariable regression spline (MRVS) (Royston and Sauerbrei, 2007, Royston and Sauerbrei, 2008). According to Royston & Sauerbrei (Royston and Sauerbrei, 2007), the MVRS and MFP procedures are similar in spirit and character, suggesting the application of MVRS will also produce a biased multivariable model. Therefore, the problem of minimising bias in the multivariable RCS model could also be dealt with by adjusting for a ‘minimally’ sufficient set of covariates as explained in section 2.2.4.

2.2.5.1 Choosing the number of knots and their position

In the absence of prior knowledge, analysts using cubic regression splines should carefully decide on the number and the placement of the knots across the exposure-outcome functions being fitted. Stone (Stone, 1986) found that more than 5 knots were rarely required to fit the RCS models. Studies performed after Stone’s investigation suggest 3-5 knots will suffice in the RCS functions (Durrleman and Simon, 1989, Heinzl and Kaider, 1997, Harrell, 2001). For knot placement, the available procedures are still unclear on whether the knots positions should be determined from the exposure-outcome curvature, shape or the size of the sample (Durrleman and Simon, 1989). A reasonable approach was recommended by Harrell (2001) to place the knots as follows:

- i. At the quantiles/percentiles distribution of the exposure (x_1),
- ii. At the extremes,

- iii. Equally spaced over the quantiles/percentiles

The pre-specific positions for knots placements proposed by Harrell (2001) are provided in Table 2.2 below.

Table 2.2: Knots positions expressed in quantiles/percentiles of the exposure (x_1)

No of knots, K	Knots positions expressed in quantiles of the exposure (x_1)				
3		0.1	0.5	0.9	
4		0.05	0.35	0.65	0.95
5	0.05	0.275	0.5	0.725	0.95

Source: (Harrell, 2001)

The advantages of Harrell's method of knots selection are as follows:

1. The knots selection are less subjective
2. A commonly used automatic knots selection scheme that allows reproducibility and comparison of results between studies (Heinzl and Kaider, 1997).

Therefore, this practice of knots selection using restricted cubic splines (RCS) models could be of interest to researchers investigating exposure-outcome relationships in medicine.

Other strategies include adaptive procedures based on standard algorithms for "optimal" knots selection (Morton, 1988, Friedman and Silverman, 1989, Luo and Wahba, 1997, Zhou and Shen, 2001). Knots selection based on such strategies could be subjective - there exists no standard algorithm which produces the best possible number and position of knots from the data alone (Morton, 1988, Zhou and Shen, 2001). Additionally, these methods exhibit computational burden in large samples because sets of candidate knots have to be examined to establish the 'optimal' number of knots and their positions (Zhou and Shen, 2001).

2.2.5.2 The strengths and weaknesses of RCS

Restricted cubic splines have advantages of parsimony and offer a wide range of flexible shapes (Desquilbet and Mariotti, 2010). Unlike in other spline functions, RCS models allows flexibility in the data without requesting too many parameters (Therneau and Grambsch, 2000, Harrell, 2001). The potential shortcoming of RCS is that they constrained to be linear at their tails; this could strongly affect the entire shape of the curve and enhance the sensitivity of the overall shape to outliers (Greenland, 1995b). Furthermore, the RCS offers parameter estimates that are hard to interpret thus their estimated curves are recommended for use as the main output (Heinzl and Kaider, 1997). Finally, it is not always clear what degree of smoothness should be imposed on the data when working with spline functions (Royston et al., 1999). Thus, the choice of knots should strike the balance between adequate smoothness and avoiding over-fitted functions or some artefacts.

2.3 Statistical modelling framework

The four methods of categorisation, linearisation, FPs, and RCS described above could be used in the framework of the generalized linear models (GLMs) (Nelder and Wedderburn, 1972, McCullagh and Nelder, 1989, Turner, 2008), to estimate exposure-outcome relationships in epidemiology. The GLMs extend from the concept of multiple linear regression and offer flexibility (e.g. in the form of nonlinear models and non-normal distributions), allowing researchers to work with various type of functions and responses or outcome variables (Royston and Sauerbrei, 2008). Section 2.3.1 briefly describes the concept of GLMs putting the application of categorisation, linearisation, FPs, and RCS in context.

2.3.1 The concept of generalized linear models (GLMs)

Crudely stated, the relationship between the covariate(s) $\mathbf{X} = (x_1, \dots, x_k)$ and the outcome (Y) modelled using the GLM models comprise three components:

- i. The outcome variable (Y) with mean μ , $\Rightarrow E(Y) = \mu$.
- ii. A model function $\eta = \eta(\mathbf{X}, \boldsymbol{\beta})$ based on \mathbf{X} and on a vector of $\boldsymbol{\beta}$ parameters.
- iii. A link function g such that $g(\mu) = \eta$

A model function in (ii) is an additive predictor $\eta = \beta_0 + \sum f_j(X_j)$, such that β_0 is a constant term and $f_j(j > 0)$ is a function of X_j and a set of parameters (Royston and Altman, 1994). This means a linear predictor could be generalized into an additive predictor with $f_j(X_j) = \beta_j X_j$ for each j . For example, a model incorporating a quadratic polynomial in X_j has a linear predictor of the form $\beta_0 + \beta_1 x + \beta_2 x^2$ which can be written in additive format as $\beta_0 + f_j(X_j)$, where $f_j(X_j) = \beta_1 x + \beta_2 x^2$. Here, the quadratic model is additive in x , nonlinear in x , and linear in (x, x^2) . Essentially, the functions, $f_j(X_j)$ can be estimated in many ways including categorisation, linearisation, FP and RCS regression models. Such methods produce models whose additive predictors are linear as described above.

The choice of a link function g enables analysts to incorporate many types of response or outcome (Y) in the data. The normal, binomial and Poisson models, have their link functions taken as $g(\mu) = \mu$, $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ and $g(\mu) = \ln(\mu)$ respectively (Turner, 2008, Müller, 2012). For survival time data, a standard formulation involves the Cox hazard function $\lambda(t; \mathbf{X}) = \lambda_0(t)\exp(\eta)$, where $\lambda_0(t)$ is the baseline hazard function (Royston and Altman, 1994, Hollander and Schumacher, 2006).

2.4 Conclusion

This chapter provided an overview of methods available for analysing or characterising the relationships between continuous exposure and outcome variables in epidemiology. The methods of categorisation, linearisation, fractional polynomials and restricted cubic splines used in this thesis were emphasized, explaining their choice, application (using modelling scenarios), strengths and weaknesses.

As the way forward, the next chapter investigated the current extent of categorisation against the alternative methods of analysing continuous exposure-outcome relationships. The investigation was conducted in medical journals publishing medical research.

Chapter 3

Cross sectional survey of research methods in current use

3.1 Background

Most studies in medicine exhibit serious weaknesses due to issues of reporting (Little et al., 2009, Sauerbrei et al., 2014). Inadequate and poor reporting practices restrict generalisability and implementation of results and subsequently, the clinical and scientific utility of such studies is lost (Von Elm et al., 2007, Little et al., 2009, Langan et al., 2010). To aid reporting in epidemiology, the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) (Von Elm et al., 2007) and STRATOS (Strengthening Analytical Thinking for Observational Studies) (Sauerbrei et al., 2014) guidelines were developed to guide researchers working on observational studies.

Realising the benefits of research might be achieved slowly without sufficient clarity on reporting; in 2004, researchers, methodologist and journal editors met in a 2-day workshop under the STROBE initiative and developed recommendations (checklist of 22 items) necessary for an accurate and complete observational study (Von Elm et al., 2007). The established recommendations aim at contributing to the improvement of reporting in three main study designs of analytical epidemiology: cohort, case-control designs and cross-sectional studies (Von Elm et al., 2007). One aspect to consider when presenting the results of observational studies in epidemiology is how continuous risk factors are analysed and reported. The STROBE guidelines recommend authors describe how they handle quantitative variables when analysing the data; for categorised quantitative variables, the guidelines require researchers to explain and justify the

methods of categorisation. However, reviews in 2004 and 2010 suggest that few studies at that time were reporting the issues of categorisation in epidemiology appropriately (Pocock et al., 2004, Turner et al., 2010). These suggested that most continuous variables were categorised for analysis and presentation and that the basis for categorisation was rarely described. To investigate whether the analysis and presentations have improved in the past 7 years, we aimed to assess the practice of categorisation in the field of epidemiology according to the STROBE guidelines. Therefore, this was the first study assessing the reporting of categorised continuous variables several years after the STROBE guidelines.

For the purpose of this study, categorisation was defined as the practice of converting continuous variables such as age, body mass index (BMI) and blood pressure (BP) into two or more groups by splitting them at some points and designating individuals above or below the points as separate groups (MacCallum et al., 2002). For example, age could be divided into several age groups such as 1-5, 6-10, and 10+ or below/above 25th, 50th or 75th percentiles or based on quantiles (e.g. tertiles, quartiles, quintiles or deciles). In addition, binary variables were defined as measures assuming any two distinct values. For example, gender (coded 0 or 1 for male or females respectively) and medication use (coded 0 or 1 for No and Yes respectively).

The research seeks to highlight key issues necessary for improvement when reporting and analysing continuous variables in medical studies. It is suggested that categorisation have the potential to produce inaccurate estimates and clinical interpretations (Greenland, 1995b, Taylor and Yu, 2002, Chen et al., 2007, Bennette and Vickers, 2012). The latter consequences are linked with loss of information (Fedorov et al., 2009), reduced statistical power (Streiner, 2002, Peacock et al., 2012), efficiency (Zhao and Kolonel, 1992), reliability (MacCallum et al., 2002), higher type I (Austin

and Brunner, 2004) and type II (Streiner, 2002) errors likely to occur when analysing continuous variables as categorical measures. Thus, the findings in this study have relevance to authors and readers working with observational studies in epidemiology - for improved reporting and to promote or preserve scientific knowledge for synthesis and clinical decision making.

3.2 Methods

We based our assessment on five journals we would anticipate to be examples of current best practice in clinical epidemiology, using the highest impact factor (IF) ratings from the Web of Science citation report of July 2015 (Web of Science., 2015). Three journals were selected in the area of epidemiology and two general medical journals that publish epidemiological research. Journals selected were the International Journal of Epidemiology, Epidemiology, Journal of Clinical Epidemiology, the New England Journal of Medicine and Lancet. The rationale behind the selection of the five journals was based on impact factor to include journals with high levels of influence in the literature. The common use of categorisation in these leading journals would suggest the method is also widely applied in other journals with lower impact factors or more in specialist journals.

3.2.1 Study selection

For eligible articles, we considered observational studies published between 1st April and 30th June 2015. Articles published between this time intervals were selected to reflect current practice. Consideration was given to all publications with at least one independent continuous variable in the analysis. Specific eligibility criteria are as follows:

- i. Publications based on individual's data quantifying the risk or association between continuous exposures and outcomes.

- ii. The reported data should be from the original study. The study should not report pooled estimates in the form of systematic reviews and meta-analysis
- iii. The study should be based on observational designs such as cohort, case-control and cross-sectional (a requirement in the STROBE guidelines).

3.2.1.1 Exclusion criteria

We excluded all systematic reviews or meta-analyses, clinical trials or experimental studies and genetic epidemiology studies. Epidemiological studies other than cohort, cross-sectional and case-control studies such as ecological studies were also excluded because they are not covered by the STROBE recommendations. Additionally, non-related articles (e.g. comments, correspondence, editorials, non-full text abstracts) and non-related original (full text) publications (e.g. simulations, methodological papers) were also excluded. Details are provided in Figure 3.1.

3.2.1.2 Search strategy

The search for eligible articles was done amongst all publications obtained in the five journals. We reviewed all publications to identify those investigating associations between risk factors and disease outcomes or any measures in individuals. The search was done electronically, and the identified articles were later reviewed in more detail. Figure 3.1 presents a summary of the identification and selection process for eligible articles.

As shown in Figure 3.1, we identified 1005 articles from the five Journals: Lancet (540), NEJM (272), IJE (102), Epidemiology (28) and Journal of Clinical Epidemiology (63). From the 1005 publications identified, 944 articles were excluded after screening through their abstracts and titles. Reasons for excluding an article's title or abstract were based on studies identified and classified as follows; systematic reviews, meta-analyses or pooled analyses (45), non-related articles (648), non-related

original articles (60) and cohort or profile update studies (30), clinical trials and other experimental studies (121) and genetic studies (40).

The screening resulted in 61 articles which were retrieved and reviewed as full-text for inclusion in the analysis; 23 observational studies met the eligibility criteria, and 38 were excluded (see Figure 3.1). Amongst the 38 studies which were excluded, 22 were not related to the objective of the review, 4 were clinical trials and other experimental studies, 2 were meta-analyses and genetic studies and the other 10 studies investigated exposures or risk factors which were not continuous.

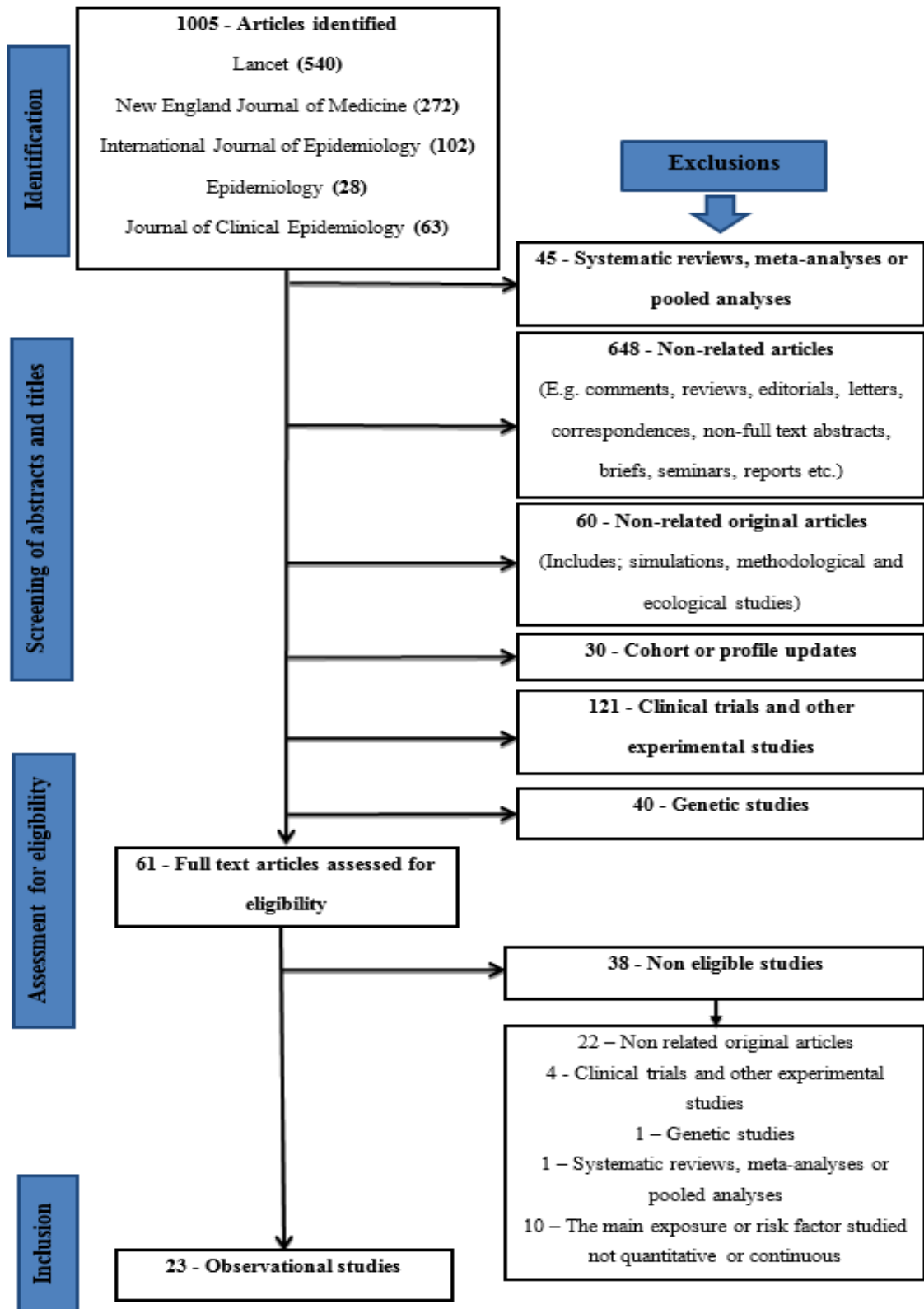


Figure 3.1: A detailed flow chart summarising the selection and identification process of eligible articles

3.2.2 Data extraction

We used a modified data collection form (see appendix - A.1) prepared by Turner et al. (2010) in their previous survey. The study variables and characteristics collected through this form are as follows: title of the study, lead author surname, date of publication, journal name, type of study design, sample size or number of participants, outcomes and exposures or risk factor characteristics (e.g. specialty, types, and whether they are categorised), details of grouping or categorisation, details of other adjusted variables included in the study, presentation and types of statistical results used in reporting, type of effect estimates (e.g. odds ratios, relative risks, confidence intervals, p-values).

3.2.3 Statistical analysis

The data collected was captured in a Microsoft Access database and exported to Stata 13 for analysis (StataCorp LP, 2013). The patterns of reporting for observational studies were quantified and reported using proportions. Where possible, examples from the data are provided for illustration. Only predominant findings or issues and practices of categorisation are reported.

3.3 Results

3.3.1 General characteristics

In this section, we provide a summary of results describing general characteristics of 23 observational studies included in the study. Overall, the three epidemiological journals produced 57% (CI = 34%, 77%) of the total articles included in the study. The other articles - 43% (CI = 23%, 66%) were obtained from the New England Journal of Medicine and Lancet. The International Journal of Epidemiology (IJE) and Lancet contributed more articles in the study than the other journals. The IJE contributed 39% (CI = 20%, 61%) of the total articles whilst from the Lancet we

obtained 35% (CI = 16%, 57%) of the total articles. Amongst these articles, cohort or follow-up studies were common. We obtained 74% (CI = 52%, 90%) of cohort or follow-up studies. The other study designs included; cross-sectional and case-control with 17% (CI = 5%, 39%) and 9% (CI = 1%, 28%) respectively.

Non-communicable diseases such as diabetes, cancer, heart diseases and mental illness were commonly studied contributing 35% (CI = 16%, 57%) amongst principal diseases or outcomes being investigated and mortality followed with 30% (CI = 13%, 53%). HIV, physiological or biochemical markers such as anti-mullerian hormone (AMH) concentration levels, body mass index (BMI) and other conditions contributed 35% (CI = 16%, 57%). These outcome variables were commonly analysed as binary variables (44%, CI = 23%, 66%), continuous variables (30%, CI = 13%, 53%) and time-to-event variables (26%, CI = 10%, 48%). For binary and time-to-event studies, mortality was more predominant compared to other outcome variables.

Considering the exposures or main risk factor variables, socioeconomic exposures were commonly investigated; 30% (CI = 13%, 53%) of studies with such exposures were obtained. For example, Zhang and colleagues (Zhang et al., 2015) investigated the associations between neighborhood deprivation index (socioeconomic exposure) and BMI (outcome). The neighborhood deprivation index in this study was derived from the 2000 US Census housing and population data using variables such as income, poverty, housing, education, and employment and occupation status. The other exposures found included; diet and lifestyle exposures (17%, CI = 5%, 39%), environmental exposures (13%, CI = 3%, 34%) physiological or biochemical markers (9%, CI = 1%, 28%) pre-existing conditions (4%, CI = 0%, 22%) and other varied risk factors (26%, CI = 10%, 48%).

3.3.2 The incidence of categorisation amongst the exposures or main risk factors

Amongst the 23 studies, 61% (CI = 39%, 80%) transformed the continuous exposures or the main risk factor variables into categorical or grouped measures for analysis. The other 39% (CI = 20%, 61%) kept the exposures or the main risk factor variables continuous. For example, Li and colleagues (Li et al., 2015) investigated the association between BMI trajectories and adult BP across two generations keeping the exposure (BMI) continuous. Linear spline function with one knot was used to summarise longitudinal changes of the BMI curves in the two generations. In another example, Victora and colleagues (Victora et al., 2015) investigated the association between intelligence quotient (IQ) and breastfeeding duration (measured in months) and categorised the exposure (breastfeeding duration). The assumed categories for the exposure were varied, defined according to the total duration of breastfeeding and predominant breastfeeding duration (breastfeeding as the main form of nutrition with some other foods). The total duration of breastfeeding (in months) was categorised using five interval groups; <1, 1-2.9, 3-5.9, 6-11.9 and ≥ 12 which differed to the predominant breastfeeding categories defined as; <1, 1-1.9, 2-2.9, 3-3.9 and ≥ 4 . In most articles, whenever categorical analysis was deployed as in the latter example, the categories were assigned ordinal values or scores to depict distinct levels amongst the categorised groups. Further details on the practices of categorisation considering only articles where continuous exposures or the main risk factors were transformed into categorical or group measures (n=14) are discussed in the next sub-sections.

3.3.2.1 Decisions informing categorisation

Amongst all studies which employed categorisation (n=14), one (7%, CI = 0%, 34%) article explained their choice for reported categories. Categorical groupings adopted in the study were explained as hypothetically driven. Hypothesis-driven

categories were then used to construct a cut-off or dichotomised model which was tested against the non-categorical (continuous) model. Otherwise, the rest of the studies, 93% (CI = 66%, 100%) did not explain or state reasons informing their choices of categorisation.

3.3.2.2 Criteria used for categorisation

Criteria used in establishing categorical boundaries for the exposure variables were varied with 21% (CI = 5%, 51%) of the studies using quantiles (e.g. median, quartiles, quintiles, and deciles). Equally spaced intervals or arbitrary groupings (which does not appear to be data or clinically driven) were very popular criteria for deciding categorical boundaries. Both equally spaced interval and arbitrary grouping criteria were observed in 65% (CI = 35%, 87%) of studies where categorisation occurred (see Table 3.1). Altogether, a combination of articles consisting ordered categories (equally spaced intervals and quantiles) and arbitrary grouping produced 86% (CI = 57%, 98%) of studies.

Otherwise, the other 14% (CI = 2%, 43%) of articles selected their categories based on established guidelines. For example, Gardner and colleagues (Gardner et al., 2015) used the WHO standards to categorise BMI into four categories; underweight (BMI < 18.5), normal ($18.5 \leq \text{BMI} < 25$), overweight ($25 \leq \text{BMI} < 30$), and obese (BMI ≥ 30) and Kaukonen and colleagues (Kaukonen et al., 2015) defined systemic inflammatory response syndrome (SIRS) status (present/absent) based on consensus statement of the American College of Chest Physicians and Society of Critical Care of Medicine.

3.3.2.3 Number of categories

When transforming continuous exposure variables for categorical analyses, the number of categories used across the studies varied between two and ten categories (see Table 3.1). Studies employing four or five categories were common. For example,

Gauffin and colleagues (Gauffin et al., 2015) investigated the association between school performance (exposure) and alcohol-related disorders (outcome) in early adulthood population by dividing the population into five categories: high school marks ($> \text{mean} + 1 \text{ SD}$); high average (between mean and mean + 1 SD); lower average (between mean and mean -1 SD); low ($< \text{mean} - 1 \text{ SD}$) and missing. The practice of categorisation with four or five categories was found in 57% (CI = 29%, 82%) of the articles. Dichotomisation (or grouping into two categories) was observed in one (7%, CI = 0%, 34%) article whilst ten categories appeared in two (14%, CI = 2%, 43%) articles (see Table 3.1).

When comparing the practice of categorisation using quantiles against equally spaced interval grouping, four or five categories were more likely to occur with the latter practice. Amongst studies with four or five categories, equally spaced interval grouping occurred in 38% (CI = 9%, 76%) of the articles compared to 25% (CI = 3%, 65%) of quantiles.

3.3.2.4 Trend testing and analysis

Trend tests are often performed to assess the strength of any exposure-outcome relationships that may exist in an investigation (Kodell and Chen, 1991). The results show that 57% (CI = 29%, 82%) of the studies which employed categorisation, performed the trend tests. For example, Wang and colleagues (Wang et al., 2015) performed a trend test in risk estimates using the median values of the heart rate quintile categories. The five values were treated as a continuous measure and were used to evaluate the risk trend; p-values were presented as part of the trend testing. In another example, Victora et al. (2015) performed the linear trend test based on mean categories for months of breastfeeding.

Amongst all studies where trend testing was performed, various significance trend values ranging between 0.0001 and 0.001 were obtained and interpreted as significant. However, there was variation across studies on how these values were obtained. Guertin and colleagues (Guertin et al., 2015) obtained the overall trend value from the pairwise estimates comparing coffee drinkers (number of cups/day) against non-drinkers (reference group). Moreover, in some studies, floating estimates (where no reference group is assumed) were used to attain the trend values.

3.3.2.5 Covariate adjustment

Considerations were also made to establish the number of confounders or other variables often adjusted for in studies investigating exposure-outcome relationships. Amongst studies where the exposure or main risk factor was categorised, the number of confounders or adjusted variables ranged between 3 and 20 with an average of 10 variables. Cohort or follow-up studies tend to report large numbers of variables or confounders compared to cross-sectional and case-control studies.

3.3.3 Summary of key findings

Table 3.1 provides summary statistics of key findings emerging from the study results. The proportions and confidence intervals of main findings explaining the characteristics of categorisation are presented in the table.

Table 3.1: Key findings showing the characteristics of categorisation amongst the exposure variables in epidemiological studies

Characteristics of categorisation	% of articles & CI regions
<i>Prevalence of categorisation</i>	61% (CI = 39%, 80%)
<i>Decision informing categorisation</i>	
Hypothesis-driven categories	7% (CI = 0%, 34%)
Unknown (reasons not provided in the articles)	93% (CI = 66%, 100%)
<i>Criteria used for categorisation</i>	
Established external criteria (e.g. WHO standards)	14% (CI = 2%, 43%)
Arbitrary grouping	29% (CI = 8%, 58%)
Equally spaced interval grouping	36% (CI = 13%, 65)
Quantile grouping	21% (CI = 5%, 51%)
<i>Number of categories used amongst grouped exposures</i>	
2	7% (CI = 0%, 34%)
3	7% (CI = 0%, 34%)
4	29% (CI = 8%, 58%)
5	29% (CI = 8%, 58%)
6	14% (CI = 2%, 34%)
10	14% (CI = 2%, 34%)
<i>Proportion of trend testing</i>	57% (CI = 29%, 82%)

3.4 Discussion

The present study indicates a high occurrence of categorisation in epidemiological studies. Amongst the articles investigating the associations between the continuous exposures and disease outcomes, 61% of them transformed the exposure variables into categorical measures for analysis. The results are consistent with those obtained in previous reviews. Pocock et al. (2004) and Turner et al. (2010) respectively reported 84% and 86% of categorisation in epidemiological studies. However, compared to these studies, we recorded the lowest proportions of categorisation. This could be attributed to the numbers and journals selected for assessment. For instance, the American Journal of Epidemiology (AJE) which was not considered in this survey contributed more articles (about 53% of articles) in Turner's study. There is also a possibility of under-representation from other specialist areas since we only used high-ranking journals. High ranking journals may be strict and particular with the quality of work they wish to publish. Thus, this could limit the number of articles considered in our study. However, there are advantages to evaluating high impact journals. They offer us the opportunity to report on practices from leading researchers.

Amongst the transformed continuous exposures, nearly 60% of the articles reported ordered categories (using either equally spaced intervals or quantiles). This kind of categorisation when investigating the exposure-outcome relationship has some disadvantages (Bennette and Vickers, 2012). Quantiles produce estimates that are data dependent. On the other hand, equally spaced interval groupings produce categories that can be statistically inefficient and unjustifiable. With normally distributed data, it will be ideal to have more categories at the center and a few at the tails (Bennette and Vickers, 2012). One would expect this to be a justification for arbitrary grouping however none was provided for all articles where such criterion was used. Justifications informing categorisation or grouping were explained in 7% of the studies. This is

despite the call to describe, “*Why quantitative groupings are chosen in the studies*” (recommendation 11 of the STROBE guidelines). Hence, high proportions of articles not explaining their choice for categorisation could be an indication that authors are not aware of existing guidelines. Otherwise, authors are ignoring the guidelines or simply underestimating the consequences of categorising data when analysing continuous variables.

The assessment also shows that researchers use different categories when categorising exposures or risk factors. However, four and five categorical groupings were common amongst studies categorising continuous exposure variables. Approximately, 60% of the studies used four or five categories when transforming the exposures for analysis. The finding is consistent with what other researchers view as a common practice in epidemiology (Becher, 1992, Royston et al., 2006). According to Royston et al. (2006) and Becher (1992), four or five categories are often created in the field of epidemiology. Dichotomisation was not popular; the practice featured in one article only.

Of particular interest was also how the confounders and other variables were adjusted when investigating the exposure-outcome relationships. There are no clear procedures to decide on the choice and number of confounders and other variables when investigating exposures and outcome relationships (Sauerbrei et al., 2007). Quite often we rely on evidence from other studies, subject knowledge, statistical packages and correlations to choose the variables we wish to include as confounders in our analysis. In this study, we observed large numbers of unrelated confounders and variables being investigated. This could result in false positive claims. Careful consideration is needed to establish what true confounders are in our investigations. In one article in this assessment (Guertin et al., 2015), a multivariable model was adjusted for 20 variables.

Such models are hard to interpret and can be misleading. Variables might be dependent on each other making it difficult to explain their associations. The use of directed acyclic graphics (or DAGs) (Textor, 2013) offers a better solution to identify and establish relations. DAGs provide graphical models explaining causal relationships amongst variables of interest (Textor, 2013). Furthermore, studies with a large number of confounders and variables should also be accompanied by large samples. The samples should also incorporate the study designs. Otherwise, studies with small samples, categorising exposures and having too many variables are likely to be underpowered (Royston et al., 2006).

Taking into consideration trend testing and analysis, 57% of the articles performed the tests after categorising the exposure variables. Trend values such as ordinal scores, mean and median of categories were often used in fitting and evaluating the overall trends. In all the studies reviewed, the null hypothesis was not clearly provided. However, indications from the studies suggest the hypothesis of no exposure-disease association was always assumed. We found that small significance values for trend statistics were in some studies interpreted as the existence of a monotonic (continuously increasing or decreasing) relationship between the exposures and risk outcomes. For example, after obtaining a trend value of 0.0006, Liu and colleagues (Liu et al., 2015) concluded that the risk between nasopharyngeal carcinoma (NPC) and categorised sibling size was continuously increasing. Such interpretations could be misleading. Sometimes a significant trend statistic value does not imply a continuously increasing risk of exposure on the outcome. Trend tests are not tests for monotonic exposure-outcome relationships (Maclure and Greenland, 1992, Schmidt et al., 2013a). If the exposure-outcome relationship is unknown, the trend test may obscure rather than reveal the relationship (Maclure and Greenland, 1992). Trend or slope estimation

methods such as polynomial regression and non-parametric models should supplement trend testing when investigating relationships which are unknown.

3.5 Conclusions

In epidemiology, studies evaluating issues of categorisation according to the STROBE guideline are lacking. Based on recommendation 11 of the STROBE guidelines, this study highlights current practices for analysing continuous variables focusing on issues of categorisation. Findings obtained using five medical journals indicates high proportions of categorisation within epidemiological studies. The categorisation of continuous exposure or risk factors was found in 61 percent of articles assessed. Reasons and justifications informing the choices and practices of categorisation are rarely provided and remain unknown. The findings confirm the presence and claims of categorisation viewed by some researchers as a dominant feature for analysing continuous data in medicine.

Clearly, these findings raise concerns about the adequacies of analysis and quality of reporting. Categorisation enables researchers to assume simple relationships between the outcome and exposures and in the process the information is lost. How much information is lost will depend on cut points or categories used (Altman et al., 1994). In this study, we have seen four or five group categories being dominant. However, we cannot be certain of how much information is lost when four or five group categories are assumed under different exposure - outcome associations.

The majority of researchers also preferred to use equally spaced intervals or arbitrary grouping. In medicine, biologically meaningful cut points are necessary to inform decisions which relate to the pattern of the data. Establishing meaningful cut points where complex relationships or associations are present may not be easy. Alternative approaches such as fractional polynomials (Royston and Altman, 1994,

Royston et al., 1999) and splines (Desquilbet and Mariotti, 2010, Schmidt et al., 2013b) are available. However, the precision and performance of these approaches in the presence of complex associations are also not well known (Keogh et al., 2012). Further research evaluating these approaches, their performance and precision under different complex associations is required.

Other existing guidelines available for medical researchers can be found on online resources including the Enhancing the QUALity and Transparency Of health Research (EQUATOR) network website (www.equator-network.org) which have the aim of improving the reporting of epidemiological and clinical studies.

Chapter 4

Comparison of different approaches for modelling associations between an exposure and a continuous outcome – a simulation study

4.1 Introduction

Limited comparisons on the properties of FP and RCS and their performance against the method of categorisation and linear models (known as linearisation in this thesis) are available to date (Strasak et al., 2011, Binder et al., 2013). Lack of examples in the application of these methods, differing views on the most appropriate alternative approach, perceived difficulties in application and interpretation of estimates could be the reason why the FP and RCS approaches are not widely used in medical studies. The survey on current research practice revealed the common use of categorisation amongst medical studies investigating the relationships between continuous predictor variables and the outcomes. About 61% of the publications converted continuous predictor variables into categorical groups during statistical analysis. Further results on the current research practice of analysing continuous predictors in medicine could be found in Chapter 3.

One way to compare a variety of statistical methods such as categorisation, linearisation, FP, and RCS is through a simulation study. Simulation studies are known for their ability and strength in assessing the appropriateness and accuracy of statistical methods using pre-defined scenarios (Burton et al., 2006, Crowther and Lambert, 2013). This is because, in reality, the appropriateness and accuracy of statistical methods cannot solely be evaluated and achieved with real data alone (Burton et al., 2006). In the

literature, simulation studies comparing, fractional polynomials and spline based approaches have focused mostly on (1) variable or covariate selection and (2) global fits or functions relating to covariates and/or outcome variables. For example, Binder and colleagues (Binder et al., 2013) compared FP and RCS methods in multivariable setting to help provide guidance on the appropriate technique for model building in situations where moderate number of covariates with different shapes are of interest. In another simulation, the performance of FP and RCS models were assessed based on various association functions found in environmental and occupational epidemiology (Govindarajulu et al., 2009). To reflect studies in environment and occupational health, Govindarajulu et al. (2009) investigated six plausible exposure-response scenarios (with right skewed exposure distributions). One important aspect lacking in these simulations was the performances of FP and RCS methods against the turning or thresholds (Muggeo, 2003, Benedetti et al., 2009) points. Turning points or thresholds (words used interchangeably) are defined as the locations where nonlinear functions experience sudden changes in their directions or slopes.

In epidemiology, exposure-outcome relationships may be characterised by sudden changes when the exposure reaches an unknown threshold or turning point (Pastor and Guallar, 1998, May and Bigelow, 2005). When this occurs, some natural or biological phenomenon may be present; requesting some interpretations from the researchers for health policy implementation or planning. Therefore, reporting exposure-outcome relationships require reliable models that have the ability to predict not only the function but also its turning points or thresholds (if they present in the data). However, none of the usual methods of analysis provides an inferential guide for estimating the location of turning-points when modelling exposure-outcome relationships (Pastor and Guallar, 1998).

To bridge this gap, this chapter assessed and compared the performance of categorisation, linearisation, FP and RCS methods based on different exposure-outcome relationships (with predetermined thresholds or turning points) to provide guidance on the appropriate models. The assessment was performed using a simulation study under the ‘normal error’ regression framework - assuming different relationships between one continuous exposure variable and one continuous outcome. The idea was to evaluate the performance of these methods directly in simulated datasets where the distribution of the continuous exposure and shapes of the underlying exposure-outcome relationships are known. The ultimate aim was to encourage the use of FP and RCS models and inform researchers on their properties of estimating exposure-outcome functions and turning points in medical studies. Although, the FP and RCS models has potential features suitable for analysing exposure-outcome studies they are not widely used. Findings in Chapter 3 showed that the method of categorising continuous predictor variables was popular amongst studies investigating exposure-outcome relationships. Compared to the methods of linearisation, FP and RCS, the method of categorisation may be inadequate or limited for analysing exposure-outcome relationships and estimating thresholds in the data. To investigate this, the specific objectives of this chapter are provided in section 4.1.1. Detailed simulation procedures are provided in section 4.2. Sections 4.3, 4.4, 4.5 and 4.6 describe the performance measures in the simulations, results, discussion and conclusion of this chapter respectively.

4.1.1 Aim and objectives:

The main aim of this chapter was to investigate and compare the performance of categorisation, linearisation, FP, and RCS methods based on simulated exposure-outcome relationship datasets - focusing on the ability of these methods to (1) recover the ‘true’ relationships assumed in the simulations and (2) estimate the positions of ‘true’ turning points or thresholds in the data.

To achieve this aim, the predicted functions (attained using these methods) were compared with the ‘true’ exposure-outcome relationships in the simulations by quantifying or evaluating the following performance measures:

- i. The root mean square error (RMSE) for goodness of fit. The RMSE was computed by taking the difference between predicted functions and the true simulated exposure-outcome curves (without the error).
- ii. The rates of type 1 errors amongst the alternative regression models (fractional polynomials and restricted cubic splines). A function showing a linear relationship between an exposure and the outcome was simulated and fitted with FP and RCS models to estimate the proportion of times linearity was rejected in 1000 replicates (iterations).
- iii. The 95% confidence interval regions of median predicted functions used in estimating the ‘true’ exposure-outcome shapes. The 95% confidence intervals allowed inference on the coverage of median predicted models against ‘true’ shapes.
- iv. The precision of median predicted functions in estimating the actual turning points or thresholds assumed in the simulations. The estimated turning or threshold points from the mean predicted curves were provided with their corresponding 95% confidence intervals.

4.2 Simulation framework

4.2.1 Introduction

Motivated by application to real data, the simulations in this chapter were exemplified by alcohol-blood pressure relationships found in epidemiological studies. However, this work is generalizable to most observational studies investigating exposure-outcome relationships with continuous exposures or risk factor variables. Both

the continuous and binary outcome models were covered in the simulations. But, the present chapter only provides for continuous outcome models. The results showing the simulations with binary outcomes are presented in Chapter 5.

In the literature, several studies report positive associations between alcohol intake and blood pressure (Chang and Park, 1991, Marmot et al., 1994, Choudhury et al., 1995, Moreira et al., 1998). However, the shape of the alcohol-blood pressure association and threshold dose for hypertension remains unclear (Keil et al., 1998, Husain et al., 2014). This is because in some studies the relationship has been reported as linear (Chang and Park, 1991, Beilin et al., 1996), U-shaped (Jackson et al., 1985, Matsumoto et al., 2009) and sometimes with threshold effects or J-shaped (Klatsky et al., 1977, Gillman et al., 1995).

In most of these studies, the authors reported alcohol-blood pressures relationships after categorising the alcohol intake measurements (Jackson et al., 1985, Moreira et al., 1998). The possibility of misspecified models and risk estimates in such instances could occur. Assuming an adult population, alcohol consumption datasets (intake measured in continuous scale) were generated based on example scenarios found in the literature to compare the method of categorisation against the linear, FP and RCS models in a simulation study.

4.2.2 Simulation set-up

Let's consider a simulation structure with an outcome variable (y) and one continuous exposure variable (x). Suppose an exposure variable (x) was drawn from a uniform distribution with a range of values between 0 and $\max(x)$ and y as an outcome variable has normally distributed errors. To conduct simulations of various exposure-outcome relationships from an example study chosen from epidemiology, assume $y \sim N(E(y), \sigma^2)$ where $E(y) = f_i(x)$ represent the simulated true mean functions such

as those found in the example provided in section 4.2.2.2 and σ denote the standard errors of an outcome (y).

With any statistical program suitable to conduct simulations, a seeding number could be initiated to generate random observations and set up the simulation with multiple datasets assuming different number of observations ($N_i, i = 1, 2, \dots, k$) with various random errors ($\sigma_i, i = 1, 2, \dots, n$) replicated R times. For each dataset, the observations inside the loop should be varied when replicating the samples R times. Moreover, the simulation should be done such that R is sufficient to produce minimum Monte Carlo error (MCE). The MCE is linear in $1/\sqrt{R}$. Therefore, based on the asymptotic property, as R is increased, $1/\sqrt{R} \rightarrow 0$ as the MCE (Koehler et al., 2009).

4.2.2.1 Description of Stata (statistical simulation program)

The programming structure in Stata 13 (StataCorp LP, 2013) is widely supported for many estimators and functions hence the program was identified as a suitable environment to perform the simulations. The Monte Carlo simulations in Stata can be carried out using either the `simulate` or `postfile` command (Adkins and Gade, 2012). In this work, the `postfile` command was used. The `postfile` command works with loops (e.g. `forvalues`, `while` and `foreach` looping) that makes it more powerful and flexible to use. Using the `simulate` command has disadvantages; it requires a lot of intervention from the user. The joint analysis and results from different models with different parameters cannot be attained without manually changing the file names and merging different datasets (Adkins and Gade, 2012). Example codes of the simulation procedures carried out using the `postfile` command are provided in Appendix B.

4.2.2.2 Alcohol blood pressure example

Consider alcohol intake (measured in grams/day) as a uniformly distributed exposure variable (x) with a range of values between 0 and 60 grams/day (average intakes from the literature) and systolic blood pressure (measured in millimetres per mercury, mm Hg) as an outcome variable (y) with normally distributed errors. As an example, the proposed mean functions, $f_i(x)$ similar to those found in epidemiological studies explaining alcohol-blood pressure relationships are presented in Table 4.1.

Table 4.1: Nonlinear associations investigated with continuous outcome models

Type of associations	Functions	Proposed 'true' functions
Linear	$\beta_0 + \beta_1 x$	$f_1(x) = 120 + 0.38x$
Linear piecewise threshold	$\beta_0 + \beta_1(x - c)$	$f_2(x) = \begin{cases} 121 & \text{if } x \leq 20 \\ 121 + 0.78 * (x - 20) & \text{if } x > 20 \end{cases}$
Nonlinear piecewise threshold	$\beta_0 + \beta_1(x - c)^2$	$f_3(x) = \begin{cases} 121 & \text{if } x \leq 20 \\ 121 + 0.038 * (x - 20)^2 & \text{if } x > 20 \end{cases}$
U-shaped or Quadratic	$\beta_0 + \beta_1 x + \beta_2 x^2$	$f_4(x) = 134 - 1.44x + 0.036x^2$

From the example in Table 4.1, multiple datasets were created based on the proposed mean functions representing alcohol-blood pressure relationships such that each dataset has 200, 500, 1000, 5000, and 10000 observations respectively. Variation

within individual's blood pressure was accounted by generating observations with varying standard errors ($\sigma_i = 2.5, 5.0, 7.5$). NOTE: When $y = f_i(x) + \varepsilon$ then ε is the error term with mean zero and variance, ($\varepsilon \sim N(0, \sigma^2)$). This procedure produced blood pressure (BP) outcome values with minimum, moderate and large variations in the datasets respectively. These BP outcomes were then predicted using the categorisation, linearisation, FP, and RCS regression models taking alcohol intake as an exposure variable. In this example, the proposed standard errors assumed corresponds to findings reported by Saunders and colleagues (Saunders et al., 1981) in a cohort study of 132 alcoholic patients who were admitted in a hospital for monitoring. The study was performed to investigate the relationship between alcohol consumption and blood pressure on patients who regularly consumed more than 80g of alcohol/day. The investigation was done over a period of two years and the blood pressures of patients were measured at the day of admission while still drinking, during detoxification from alcohol and after a period of abstinence. After a day of admission, BP was taken at least twice a day. The BP readings were made when the patient was lying upright. Following detoxification and continued absenteeism of alcohol, a mean change of 17.4 mmHg from the normal systolic BP of 120 mmHg was reported. In the example, this mean change correspond to $\sigma = 7.5$ that yield mean fluctuation of approximately 18 mmHg from the normal blood pressure levels amongst the alcohol consumers. In datasets where $\sigma = 2.5$, the mean change of blood pressure in alcohol consumers was expected to be approximately 6 mm Hg from normal systolic BP of 120 mmHg. Similarly, for $\sigma = 5.0$ the expected mean change from the normal should be approximately 12 mm Hg.

Based on these example datasets, the categorical, linear, FP and RCS regression models were then fitted to predict the mean functions in Table 4.1. The latter was

performed in each dataset with varying observations ($n = 200, 500, 1000, 5000, 10000$) and $R = 1000$ replications. Replicating datasets 1000 times produced an approximate MCE of 3%. Given this, good regression models should be able to produce sufficient estimates in the simulated datasets since R was large and yields minimum MCE.

Graphically, the example mean functions in Table 4.1 are shown in Figure 4.1. These mean functions were referred as ‘true’ functions in this chapter and the interest was to establish how the categorical, linear, FP and RCS regression models performed against them in the simulation.

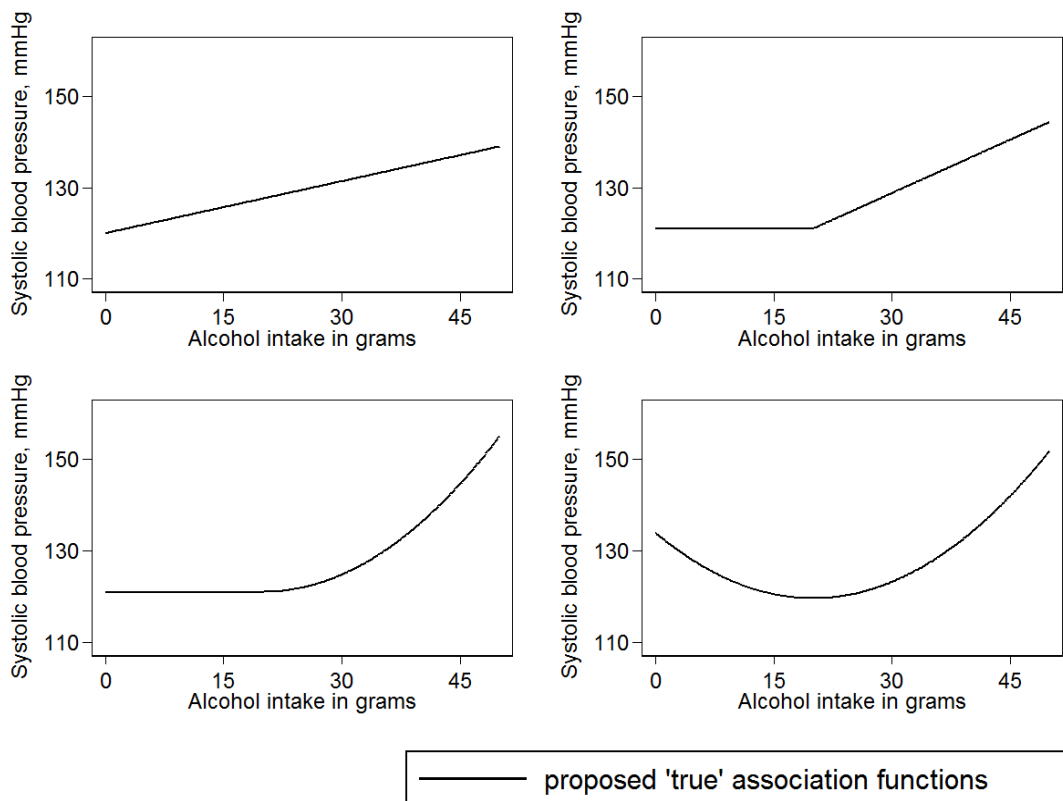


Figure 4.1: The proposed 'true' association functions used in the simulations to compare the properties and performances of fractional polynomials, restricted cubic spline, categorisation and linear regression techniques

Briefly from Figure 4.1; (i) the linear association function assume an increasing blood pressure for any unit of alcohol consumed. (ii) The piecewise association function was characterised by the absence of harmful effects at certain levels of consumption however after the threshold, the effect was large and harmful. Two types of piecewise associations were investigated; (a) linear and (b) nonlinear piecewise functions. The latter depicts the increasing nonlinear effect of alcohol on blood pressure after the threshold whilst the linear piecewise shows an increasing linear effect. (iii) For U-shaped or quadratic association function, the effects of alcohol intake on blood pressure were greater at minimum and maximum units of consumption. In the quadratic association function, the optimal intake favourable or beneficial to consumer's blood pressure was found between these extreme units (see Figure 4.1).

4.3 Methods evaluating the performance of statistical models

This section begins by briefly describing the implementation of categorical, linear, fractional polynomials (FP), and restricted cubic splines (RCS) regression models in the simulation. For additional information, the methods of categorisation, linearisation, FP, and RCS were broadly described in Chapter 2. This section concludes by explaining the performance measures used to evaluate the regression models under investigation.

In the simulations, the method of categorisation was implemented by transforming the continuous exposure variable (alcohol intake, g/day) in the following ways. First, the alcohol intake measures were categorised into three categories (CAT3) as commonly reported in alcohol studies (Higashiyama et al., 2013). Second, the alcohol intake measures were transformed into five categories (CAT5) - to assess the influence of number of categories in the simulations. The categorical boundaries in

CAT3 were established using the tertile distribution of the exposure. In contrast, quintiles were used to establish categorical boundaries under the CAT5 approach. Tertiles and quintiles produce ordered categories preferred by medical researchers in studies categorising continuous variables (see Chapter 3). Alternatively, the relationship between continuous exposure and the outcome variables were analysed by keeping both the exposure and outcome variables continuous in the simulations. The simplest approach of keeping these variables continuous in the analysis was assuming linearity on the data. This process was achieved through the method of linearisation.

For complicated functions covering both linear and nonlinear relationships, the FP and RCS were fitted. The FP model with $m \leq 2$ degrees was considered sufficient for the simulated alcohol-blood pressure relationships datasets. The FP family with $m \leq 2$ degrees offers a wide range of association shapes that generally improves the fits covered by conventional polynomials (Royston and Altman, 1994, Royston and Sauerbrei, 2008). For a permissible set of FP powers, $p_j \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$ there exist 44 possible combinations of models for the FP family with $m \leq 2$ degrees. Furthermore, the second degree FP models (with $m = 2$) allows the estimation of at most one possible turning point (Royston et al., 1999). Hence, the proposed FP approach should be adequate for any nonlinear relationship assumed in the simulations since few turning points (<2) are considered.

The RCS method was implemented by using 3 knots located at the 10th, 50th and 90th percentiles distribution of the exposure (alcohol intake, g/day). The RCS functions with 3 knots should provide adequate fits since the simulated datasets are less complex with few turning points (<2). The RCS functions with more than 3 knots were likely to be less parsimonious and over fitted because the curves under investigation were not characterised by multiple sudden changes over the exposure space (Durrleman and

Simon, 1989, Rutherford et al., 2015). Finally, the fractional polynomials (FP), restricted cubic splines (RCS) models in this simulation study were implemented using standard procedures found in most available statistical programs.

The final regression models attained with the methods described above were then evaluated and compared using the following performance measures:

4.3.1 Goodness of fit

The root mean square error (RMSE) was used to measure the goodness of fit for predicted models. The RMSE assessed the goodness of fit in the simulations by considering the distance between the predicted association curves and the ‘true’ association functions. In 1000 simulations (replicates), the RMSE was obtained by taking the difference between the predicted curves and true association functions (for each iteration). This was done at each observed data points along the exposure scale (x-axis). Thus, for simulations with different sample sizes $n = 200, 500, 1000, 5000, 10000$ replicated 1000 times ($R = 1000$) the RMSE was calculated as follows:

$$RMSE(\hat{f}_i(x_j)) = \sqrt{\frac{1}{n} \sum_{v_{ij}} (\hat{f}_i(x_j) - f_i(x_j))^2}, \quad \text{Eq. 4. 1}$$

where $\hat{f}_i(x_j)$ is the estimated association curve for the i^{th} simulated dataset evaluated at the j^{th} exposure value (x_j) and $f_i(x_j)$ is the ‘true’ curve at x_j .

To visualise and compare the performance of different methods fitted in various exposure-outcome shapes in the simulation, the median of all the RMSEs for each set of simulations (in 1000 replicates) were reported and presented using bar graphs. A regression model with the smallest median RMSEs had the better fit suggesting a closer relationship between the predicted and ‘true’ association shapes. The 95% confidence intervals (CI) for the median RMSE were also provided in summary graphs for

visualisation and also to convey the effects of sampling variation in the simulated datasets. To estimate the 95% confidence intervals for the median RMSEs the lower and upper boundaries were first calculated based on the following percentiles:

$$lb = \frac{R}{2} - \left(Z_{1-\alpha/2} \frac{\sqrt{R}}{2} \right) \text{ and } ub = 1 + \frac{R}{2} + \left(Z_{1-\alpha/2} \frac{\sqrt{R}}{2} \right) \quad \text{Eq. 4. 2}$$

where $R = 1000$ iterations and $Z_{1-\alpha/2}$ is the appropriate value from the standard normal distribution. At $\alpha = 0.05$ the standard normal value is 1.96.

When R replicates are ranked in an increasing order of magnitude, the 47th and 53th percentiles points (rounded to the nearest integer) gives the lower and upper bound respectively for the median RMSE estimates. This approximation method for median confidence intervals is acceptable in most sampling scenarios (Campbell and Gardner, 1988). In the simulations, the median confidence intervals have attractive properties compared to when using the general CIs from the means. For any estimator, they are easy and simple to obtain whereas the mean CIs are generally difficult to compute. After ranking the estimators, one could easily get the median CI as stated above after ranking the estimators whilst for mean CIs special care would be needed to determine the variance of the estimators. Wider CIs of the median may also suggest insufficient R in the simulations (Strelen et al., 2001).

Going back to the RMSE, one distinct advantage of this quantity is that it provides a quadratic loss function (Makridakis and Hibon, 1995). The RMSE enable researchers to study and infer on necessary conditions that may be required to achieve minimum distance between the predicted and ‘true’ functions in the simulations. For example, large value of the loss function means optimisation may be required to achieve minimization. In such scenarios, optimisation could be achieved by either (1) increasing/decreasing the sample sizes (n) in the data and/or (2) working on datasets

with small/large noises in the data. Thus, the RMSE plays an important role as a statistical measure of variance/uncertainty around the 'true' association functions (Makridakis and Hibon, 1995, Chai and Draxler, 2014). In statistics, such measures of variance are pre-requisite for inference. They provide a complete picture of the error distribution (Makridakis and Hibon, 1995, Chai and Draxler, 2014). Contrary to these positives, the RMSE is greatly influenced by extreme values or outliers hence it can be a misleading measure of the average fit (Willmott and Matsuura, 2005, Chai and Draxler, 2014). Nonetheless, the RMSE have been recommend for model assessment in medical studies for continuous outcome measures (Harrell et al., 1996). Besides the RMSE, another competing measure of model fit is the mean absolute error (MAE). However, the MAE was excluded for evaluation since is only recommended for uniformly distributed error models (Chai and Draxler, 2014).

4.3.2 The type I error and power rates

4.3.2.1 Type I error rate

In this simulation, the type I error was defined as the proportion (or probability) of rejecting the test of linearity while it was actually true. This test was performed with alternative regression models (fractional polynomials and restricted cubic splines) to assess their susceptibility in fitting linear association datasets. Using the simulated linear association datasets obtained through the 'true' linear functions assumed in Table 4.1, the linearity tests were performed as follows:

- i. The null hypothesis using fractional polynomials models was set up such that, $H_0: m = 1, p = 1$ and the alternative occurred for any $m \leq 2$ and $p \neq 1$. Under the spline regressions, the null hypothesis was such that $H_0: \hat{\beta}_1 = 0.38$. Recall: In Table 4.1 the slope coefficient for the 'true' linear association function is 0.38. The alternative was any association function obtained when fitting the RCS model with 3 knots.

- ii. After setting the hypothesis, 1000 simulations (iterations) were performed and fitted with both FP and RCS regression models.
- iii. Based on the final FP and RCS regression models (obtained at each iteration), the linearity test was performed based on the likelihood ratio statistic with 3 degrees of freedom assuming $\alpha = 0.05$. The chi-square value of this test can be written as $\chi^2 = -2l(x) - [-2l(\hat{f}(x))]$, where $l(x)$ is the partial log-likelihood from the linear model and $l(\hat{f}(x))$ is the partial log-likelihood from the estimated model. This likelihood ratio test statistic was considered appropriate because, in both fractional polynomials and restricted cubic splines models, the linear association functions are nested within these methods.
- iv. Finally, after 1000 replications in the simulations, the proportions of times the null hypothesis was rejected (with $p - value < 0.05$) was computed and reported as type 1 error rates.

4.3.2.2 Power

The model power rates were defined as the proportion of times the FP and RCS models were able to correctly identify the existence of nonlinearity in the data and reject the hypothesised linear relationships. When the model failed to reject a false null hypothesis (existence of linearity) then the type II error was committed. Given the latter, the knowledge of type II rates (γ) was essential for quantifying the power rates in each modelling approach. The two measures (power and type II rates) complement each other. In the simulations, the model power rates were computed in all nonlinear association shapes proposed in Table 4.1 following steps (i) – (iii) outlined when investigating the type I error rates in section 4.3.2.1. In step (iv), after 1000 simulations, the model power rate was computed by subtracting the type II error rate from one (that is $power = 1 - \gamma$, where γ is the type II error rate).

4.3.3 Confidence intervals, turning points, and coverage probabilities

4.3.3.1 Confidence intervals

The confidence intervals for each method were obtained and graphed after fitting the estimated functions in 1000 simulations (replicates). This made it possible to visually assess the performance of each method against the ‘true’ association functions. To construct the 95% CI graphs for each method, the 2.5th & 97.5th percentile points of the predicted outcomes were obtained and plotted on the exposure scale to represent the CI regions. In the graphs, the 50th percentile points representing the median predicted association curves for each method were also shown. This approach of calculating the 95% CI region is known as the percentile method (Diciccio and Romano, 1988, Haukoos and Lewis, 2005). For comparisons, the ‘true’ association function curves were also presented in the graphs. The outcome values in the graphs were transformed in log scale to retain the same range of units to visually assess the differences in the fitted functions.

4.3.3.2 Turning points

In all nonlinear association datasets considered in the simulations, the ‘true’ turning points (or thresholds) occurred when the exposure was at 20 units. The outcome occurring at 20 units of the exposure in these datasets was known as the ‘true’ optimal outcome. At this position, the ‘true’ functions retain the minimum outcome values (i.e. lower BP values) in the simulations. For prediction, the optimal exposure was estimated where the predicted function retains the minimum outcome. This means 1000 optimal exposure were obtained in the simulations together with the corresponding optimal outcomes. For reporting, the 50th percentile points of the distribution representing the median estimates were provided together with their 95% CIs (estimated as the 2.5th and 97.5th percentile distribution of the predicted values). These estimates were summarised in tables for comparison with the ‘true’ optimal values in the simulations. The tables

also showed the predicted outcomes (together with the CIs) occurring at 20 units of exposure to assess if the fitted models were underestimating or overestimating the outcome at the ‘true’ optimal exposure.

4.3.3.3 Coverage probabilities

Finally, the coverage probabilities were also computed as the proportion of times the $100(1 - \alpha)\%$ confidence intervals, say $\hat{d}_i \pm Z_{1-\alpha/2}SE(\hat{d}_i)$ included the ‘true’ optimal outcome (d) assumed in the R simulations (*where* $i = 1, 2, \dots, R$) (Burton et al., 2006, White, 2010). Assuming a nominal 95% confidence interval, the coverage probability less than 95% level represents under-coverage in the simulation. In contrast, the estimated CI bounds were conservative in their coverage when the calculated proportions were greater than the 95% nominal level.

4.4 Results

This section presents the results of this chapter. Section 4.4.1 provides the results showing the goodness of fit for predicted models. The results of type I error rates associated with fitting linear association datasets using FP and RCS models were discussed in section 4.4.2. Section 4.4.2 also covers the results on statistical power of FP and RCS models. The final section 4.4.3 presents the results on model coverage; assessed by the graphs showing the predicted CI regions and reporting on coverage probabilities.

4.4.1 Goodness of fit

The goodness of fit for each regression model obtained under the four association shapes in Figure 4.1 was quantified using the RMSE as explained in section 4.3.1. The RMSE results were explained by noting the influence of noise and sample size variation in the simulation. Section 4.4.1.1 discusses the results showing the

effects of varying noises in the datasets. The effects of varying sample sizes on RMSEs were discussed in section 4.4.1.2.

4.4.1.1 Effects of noise on the RMSE

Figure 4.2 shows the effects of varying noise in a sample with 200 observations, replicated 1000 times. The noise in the datasets was varied at $\sigma = 2.5, 5.0$ and 7.5 to assess its effects on the RMSEs of different regression models. The heights of the bars in the graphs represent the medians of the estimated RMSE obtained across the simulations. The median RMSE estimates were presented using the same unit intervals to allow comparison between different models. The confidence interval limits of the median RMSE estimates were also provided at the top of each bar in the graph. However, the CI widths were narrow and not entirely visible on the graphs.

Additional results showing the estimated median RMSEs and their corresponding 95% CIs in the simulations were displayed in Appendix C (see Table 4.3).

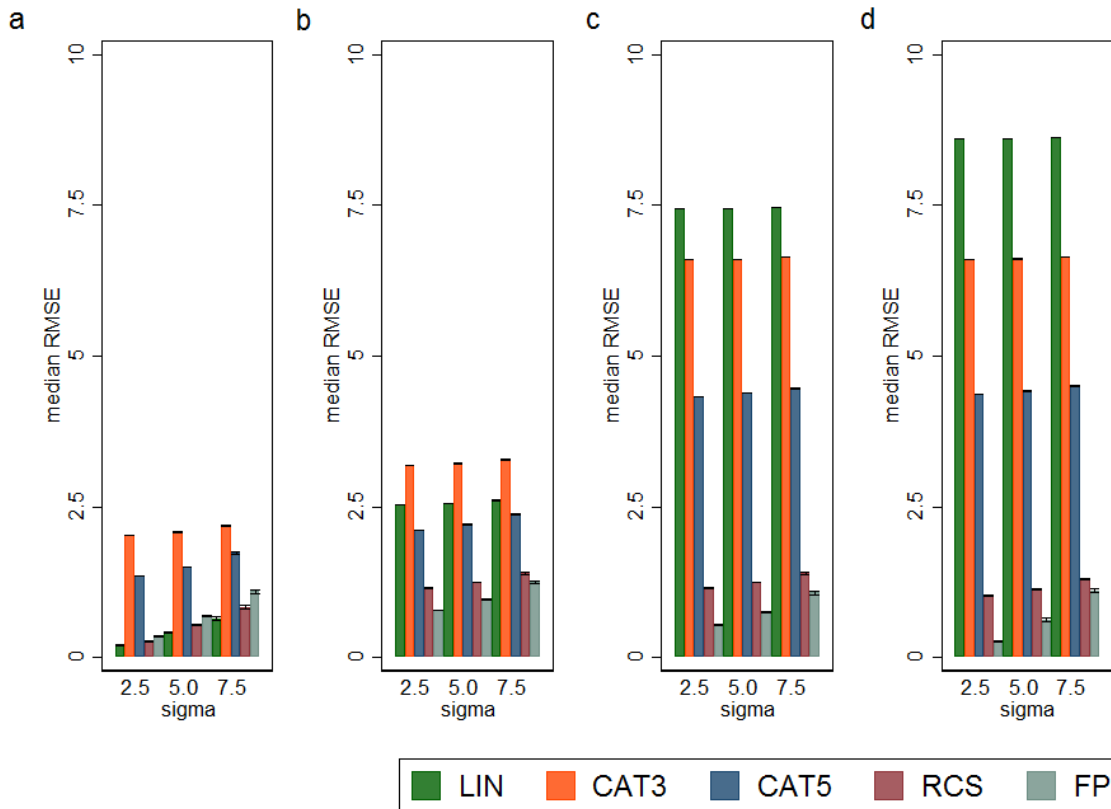


Figure 4.2: The estimated median RMSEs obtained when fitting (a) linear association shape (b) linear piecewise association shape (c) nonlinear piecewise association shape and (d) quadratic or U association shape using the linearisation, categorisation (CAT3 and CAT5), RCS, and FP models in a simulation study with 1000 replicates. Various noises (σ) were considered in the simulation with 200 observations (sample size). The 95% CI of each median RMSE are provided at the top of each bar.

In Figure 4.2 (a) where the ‘true’ association shape was linear; the linearisation method retained smaller RMSE values compared to when fitting the fractional polynomials, restricted cubic splines, and categorisation models in the same dataset. The latter holds when varying the noise (σ) in the data. Increasing σ from 2.5 to 5.0 doubled the median RMSE from 0.21 (CI=0.20, 0.22) to 0.42 (CI=0.41, 0.44) when fitting the linear regression model. When σ was increased from 2.5 to 7.5 the median

RMSE was almost three fold, 0.64 (CI=0.61, 0.68). The same pattern was also observed when fitting the same dataset with fractional polynomials and restricted cubic spline models. However, the RCS models retained smaller RMSE values than the FP models. The RMSE estimates under the CAT3 method were not adversely affected by σ in the same dataset. The CAT3 approach produced almost constant but large RMSEs when varying σ in the data. However, there was some slightest improvement when fitting the CAT5 with more categories. The CAT5 method retained the fit with the next largest RMSEs (that increased slightly with σ). Overall, these results suggest that increasing σ in linear association dataset contribute more change on the RMSEs of the linear, fractional polynomial, and restricted cubic spline regression models than when applying the methods of categorisation.

Under the linear piecewise thresholds, nonlinear piecewise thresholds and quadratic or U association datasets in Figure 4.2 (b)-(d); the fractional polynomial regression models produced the smallest RMSE quantities than the methods of linearisation, CAT3, CAT5 and restricted cubic splines. The RCS model followed the FPs with the next smallest RMSE estimates. The CAT3, CAT5 and linearisation models generally had larger RMSE estimates in these datasets. But, the CAT5 approach performed better than the CAT3 and linearisation methods. In addition, the three methods (CAT3, CAT5 and linearisation) were not affected by varying noise in these datasets. For any σ considered in the simulations, the three methods consistently produced larger constant RMSEs. In contrast, varying σ in the simulation affected the magnitudes of RMSE under the FP and RCS models (see Figure 4.2 (b)-(d)). Although the RMSE estimates under the RCS models were generally larger than those obtained with the FP models; varying σ contributed greater changes when fitting the FP models than when using the RCS methods (see Figure 4.2 (b)-(d)). The latter suggest the FP model was more susceptible to variation of noise in the data than the other methods of

analysis. This occurred despite the FP method retaining the least RMSE scores in the analysis. The RCS was more adaptive, its RMSE estimates were not adversely affected by noise variation - the RCS models neither retained larger or least RMSE estimates.

4.4.1.2 Effects of sample sizes on the RMSEs of methods being studied

The results showing the effects of sample size variation in the simulations are summarised in Figure 4.3. The simulations with different sample sizes taken in various association datasets where the noise was moderate, $\sigma = 5.0$ are shown.

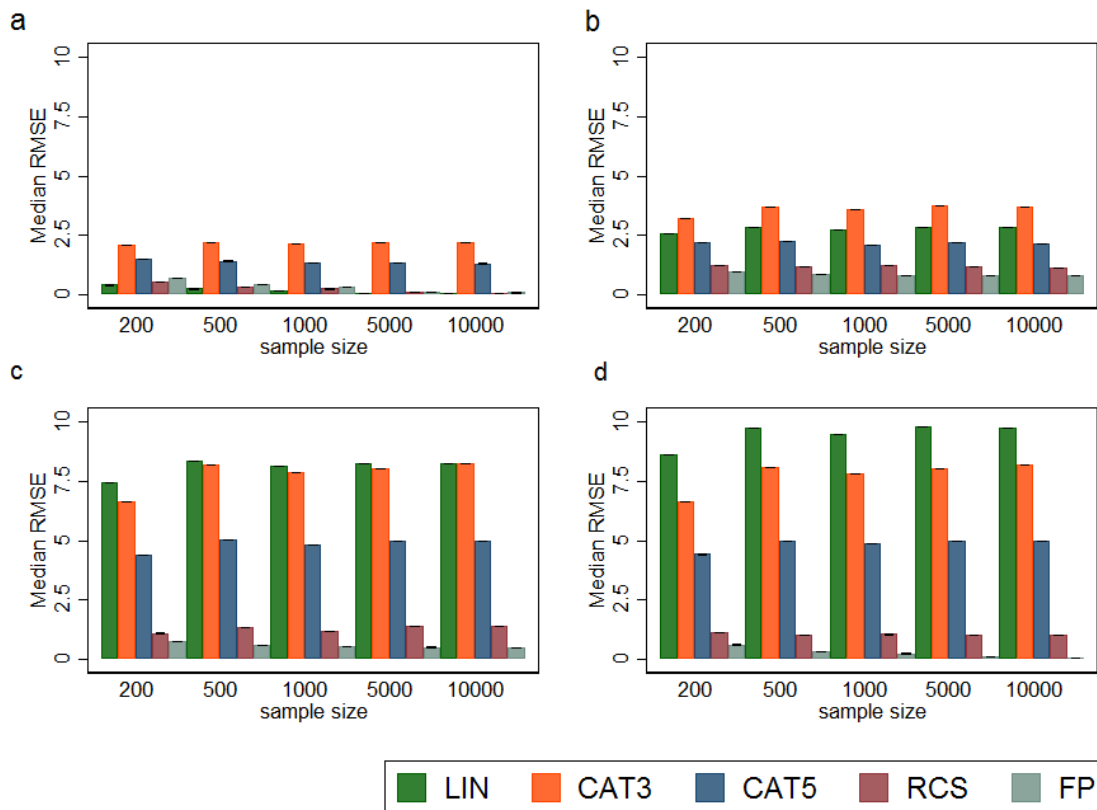


Figure 4.3: The estimated median RMSEs obtained when fitting (a) linear association shape (b) linear piecewise association shape (c) nonlinear piecewise association shape and (d) quadratic or U association shape using the linearisation, categorisation (CAT3 and CAT5), RCS, and FP models in a simulation study with 1000 replicates. Various sample sizes were considered in the simulations where $\sigma=5.0$.

The results shown in Figure 4.3 (a) were attained from the linear association datasets fitted using the linearisation, CAT3, CAT5, FP and RCS regression models. Varying and increasing the sample sizes from $n = 200$ to 10000 yielded almost constant but large RMSEs when applying the CAT3 and CAT5 in the data. But, the RMSEs under the CAT5 method were smaller compared to the CAT3 models. When comparing the two methods of categorisation (CAT3 and CAT5) with other methods; the linear, FP, and RCS regression models had smaller RMSEs that decreased with sample size increase. However, the linear models had the least RMSE estimates across the samples assumed in the simulation (see Figure 4.3 (a)). Overall, these results suggest sample size increase do not improve the RMSEs under categorical methods when fitting linear association datasets. In contrast, the RMSEs under the linear, FP and RCS models could be minimised by increasing samples in the data.

Under the linear and nonlinear threshold functions (shown in Figure 4.3 (b)-(c)); the CAT3 and linear models had smaller RMSEs in small sample datasets ($n = 200$). However, varying the sample to $n = 500, 1000, 5000, 10000$ produced the RMSEs that were large and steady (see Figure 4.3 (b)-(c)). In contrast, the FP and RCS models retained smaller but consistent RMSEs in the simulation (when the samples were varied between $n = 200$ and $n = 10000$). Reflecting on these results, larger RMSEs under the methods of linearisation and CAT3 suggest lack of flexibility and inability of these models to adapt when applied in threshold functions. The CAT5 method produced steady, moderately large RMSEs in the two datasets (see Figure 4.3 (b)-(c)).

Under the quadratic or U association datasets in Figure 4.3 (d), the FP models produced smaller RMSEs that were decreasing with the samples in the simulation. The RCS, CAT5 and CAT3 followed subsequently with the next smallest RMSE estimates. The linear regression models retained the largest RMSEs in the same datasets.

However, unlike under the FP models; the RMSEs obtained after fitting the CAT3, CAT5, linear and RCS methods remained steady across all the samples assumed in the simulation (see Figure 4.3 (d)).

Further results showing the estimated median RMSEs and their corresponding 95% CIs in other simulation conditions not presented here could be found in Appendix C (see Table 4.3). However, Figure 4.2 & Figure 4.3 provides a comprehensive summary on the pattern of RMSE quantities for the five methods across different association functions studied. In general, similar patterns were observed across other combinations of noise and sample sizes not presented in this section.

4.4.2 Estimated type I errors and statistical power

The results investigating type I error rates and statistical power associated with fitting fractional polynomials and restricted cubic spline models are presented in this section. The type I error and power tests were performed in relation to the null hypothesis that the association between the outcome variable (blood pressure, mmHg) and the exposure (alcohol intake, g/day) was linear. To perform the tests, the linear and nonlinear associations functions provided in Table 4.1 were simulated and the likelihood ratio tests described in section 4.3.2 was executed. The proportion of times the linear association models were rejected under each function based on 1000 simulations were noted and reported in Table 4.2 below.

Table 4.2: The proportion of times the test of linearity was rejected in 1000 simulations under linear association datasets fitted using FPs and RCS models (assuming different number of observations and random error (σ)).

True association functions	Methods of Analysis	Sigma (σ)	Number of observations				
			200	500	1000	5000	10000
Linear	FP	2.5	0.196	0.150	0.173	0.159	0.135
	RCS		0.044	0.042	0.051	0.046	0.051
	FP	5.0	0.210	0.187	0.228	0.151	0.146
	RCS		0.046	0.058	0.073	0.044	0.045
	FP	7.5	0.245	0.230	0.206	0.196	0.162
	RCS		0.064	0.048	0.051	0.044	0.055

The results in Table 4.2 suggest the type I error rates were increasing with noise (σ) when fitting the FP regression models in linear association datasets. However, whenever the observations in the datasets were increased, the type I error rates improved. For example, under the noisy datasets ($\sigma = 7.5$), when the sample size was small ($n = 200$), the type I error rate was 25% but this proportion decreased for larger samples. About 16% of type I error rate was observed under the FP regression model when the sample size was large ($n = 10000$). Compared to fractional polynomials, the type I error rates observed under the RCS regression models were smaller (considering all scenarios with different numbers of observations and σ in the simulations). The RCS methods produced type I error rates close to the nominal level of 5%.

The power rates obtained when fitting nonlinear association datasets with FPs and RCS models are shown in Table 4.4 (see Appendix C). As expected, the FP and RCS methods rejected the hypothesis of linearity majority of times when applied in

nonlinear association datasets. When the FP and RCS models were fitted in nonlinear threshold and quadratic datasets, the predicted power rates were 100% - suggesting that all the models fitted using linear regression were rejected (see Table 4.4 in the appendix).

For linear threshold associations, the FP and RCS models retained less than 100% power rates when the sample size was smaller ($n = 200$) and the data was noisy, $\sigma = 7.5$. The inability to reject the null hypothesis 100% times meant the type II errors were present. In such instances, the type II errors occurred when the test failed to reject the false null hypothesis in the simulation. The results in Table 4.4 (see appendix) under the linear threshold association datasets ($n = 200$, $\sigma = 7.5$) suggest the presence of 0.3% and 1.4% type II errors when fitting the FP and RCS models respectively.

4.4.3 Median predicted associations shapes, their confidence intervals, and turning points

This section presents the results showing the median predicted association shapes and their coverage regions. In the simulations, the median predicted association functions represent the average fit. Their coverage regions were reported using confidence intervals. For different association models produced using the methods under investigation, the aim was to determine whether the predicted median association shapes and their confidence intervals were able to identify and provide adequate coverage to the 'true' functions, and their turning or thresholds points. The results were summarised in two parts. Firstly, the median predicted association shapes and their confidence intervals were presented graphically – comparing predicted functions with 'true' associations assumed in Table 4.1. The second part summarised the results of the estimated outcomes and their confidence intervals predicted from the CAT3, CAT5, FP, and RCS methods at the threshold (that is at $c^* = 20$). These results were meant to infer on whether the methods under study provide reliable estimates at the 'true' thresholds

(turning points) or not. The findings on the coverage probabilities of optimal outcomes were also reported at the end of this section to validate the reliability of the estimates. Turning (or threshold) points and optimal outcome estimates were not reported under the linearisation method due to its limitations of predicting such features in nonlinear datasets.

4.4.3.1 The median predicted association shapes and their confidence intervals

To infer on the ability of the linearisation, CAT3, CAT5, FP and RCS models against the identification of true association functions in Table 4.1, the results were presented graphically in Figure 4.4 & Figure 4.5 for comparisons. The results in Figure 4.4 compared the median predicted association shapes against the ‘true’ linear and linear piecewise threshold association functions. The results for the nonlinear threshold and quadratic or U association function were provided in Figure 4.5. The 95% confidence intervals for the median predicted functions were also presented in these graphs. The summary results were obtained from the datasets with 200 observations and moderate noise, $\sigma = 5.0$ replicated 1000 times in the simulations. Identical graphs were obtained in other simulation conditions. However, narrow confidence intervals were observed in simulations with larger sample sizes. In addition, samples with large noise had wider confidence intervals (graphs not provided).

In Figure 4.4 under the linear association shapes, the linearisation, RCS, and FP regression models produced functions that lied entirely on the ‘true’ shape. However, the FP function was characterised by wider CI width at the lower tail of the exposure. For example, based on an antilog scale, at zero exposure, the FP model had an outcome of 120.74 (CI=110.63, 128.88). In contrast, the linearisation, CAT3, CAT5 and RCS models retained narrow CI width at zero units of the exposure with the predicted outcome of 119.98 (CI=118.58, 121.46), 124.66 (CI=123.54, 125.79), 123.13 (CI=121.64, 124.58) and 120.01 (CI=117.92, 122.05) respectively. The two methods of

categorisation (CAT3 and CAT5) produced step functions when fitted in linear association datasets – suggesting the categorical models do not accurately predict and identify the ‘true’ relationship assumed in the simulation. Likewise, the CIs produced under the categorical analyses provided insufficient coverage on the ‘true’ function (see Figure 4.4 – top row).

Under the linear piecewise threshold association shapes, none of the five methods entirely identified the ‘true’ function in the data. Through its lowest exposure group category, CAT3 was the only method of analyses that was able to identify the non-harmful effect on the ‘true’ function. The CAT3 method also captured the ‘true’ optimal outcome assumed in the simulations, however, this approach failed to identify the linear relationship in the dataset after the threshold. After the threshold, the CAT3 method produced a step function (see Figure 4.4 – bottom row). In contrast, the CAT5 method failed to identify the non-harmful effect occurring between 14-20 units of the exposure by overestimating the outcome as 122.73 units (instead of 121 units). At the upper exposure values (> 20 units), the CAT5 also suggested an increasing step function (just like the CAT3), however, the ‘true’ function was linear (see Figure 4.4 – bottom row). The linearisation method produced a linear fit that underestimated the outcomes at lower and upper tails of the exposure. For example, based on antilog estimates, when the exposure was zero, the outcome predicted using the linearisation method was at 113.19 units - an underestimation of the ‘true’ outcome (121 units) assumed in the simulations. In addition, the linear regression model overestimated the outcome at the actual threshold. In contrast, the FP and RCS methods provided near approximate shapes in the data, however, the estimated functions also struggled to provide sufficient coverage at the threshold. At the threshold, the FP and RCS models were overestimating the actual outcome in the simulation. Moreover, the FP function had wider CIs at the lower tail of the exposure distribution. For example, at zero unit of the exposure, the FP

regression model predicted 122.28 (CI=118.02, 129.87) units of the outcome. At the similar exposure point, the method of linearisation, CAT3, CAT5 and RCS predicted the outcome at 113.19 (CI=111.79, 114.66), 121.13 (CI=120.01, 122.26), 121.01 (CI=119.52, 122.47) and 118.36 (CI=116.27, 120.39) units respectively. (NOTE: In Figure 4.4, the estimates provided are on an antilog scale). For more details, see Figure 4.4 for comparison of median fits and their confidence interval regions.

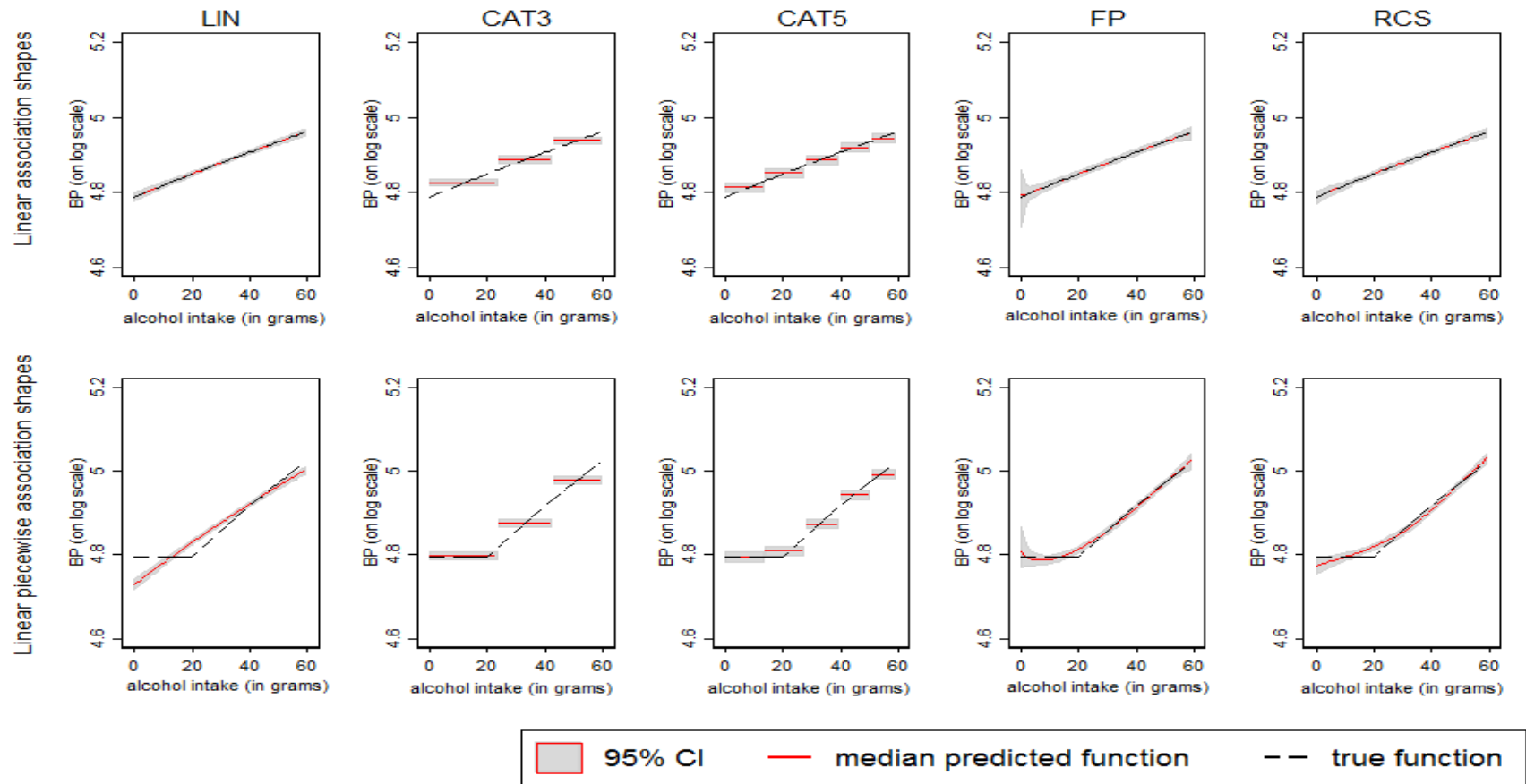


Figure 4.4: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations (replicates) after fitting the linear and linear piecewise association datasets using the methods of categorisation (CAT3 and CAT5), linearization (LIN), fractional polynomials (FP), and restricted cubic splines (RCS). The results were taken from a sample with 200 observations and moderate noise, $\sigma=5.0$

Figure 4.5 presents results from the nonlinear piecewise and quadratic or U association shaped datasets. Under the nonlinear piecewise association, the median predicted functions produced using the FP and RCS were visually indistinguishable. The two methods produced nearly similar association curves with their functions deviating a little from the 'true' at the lower values of the exposure. However, at the lowest values of the exposure, the FP model had wider CIs. Additionally, the two methods of FP and RCS slightly underestimated the 'true' optimal outcomes in the simulations. At the lower tail where no exposure effect was observed, the FP and RCS methods struggled to identify this effect producing nonlinear and linear functions respectively. However, their CI offered sufficient coverage on the 'true' effect. In contrast, the linearisation method failed to produce adequate fit depicting the 'true' nonlinear threshold function in the simulation. The two methods of categorisation also struggled to identify the 'true' relationship in the data after the threshold by producing misleading step functions. Furthermore, the CAT5 also overestimate the 'true' outcome at the threshold ($c^*=20$) (see Figure 4.5 – top row).

Like with the threshold functions, the CAT3, CAT5 and linearisation methods produced inadequate fits under the quadratic or U association datasets. The three methods produced median functions that do not lie on the 'true' association, so their estimated CI regions also failed to provide sufficient coverage on the actual fit. In contrast, the application of FP produced curves that lied entirely on the 'true' quadratic or U association functions. The RCS produced a fit that was very close to the 'true' quadratic shape. However, at the lower exposure values (< 5 units); the RCS model slightly underestimated the outcome value. The RCS function also predicted the linear relationship at the lower tail of the exposure (< 20 units) - failing to capture the actual nonlinearity assumed in the simulation (see Figure 4.5 bottom row).

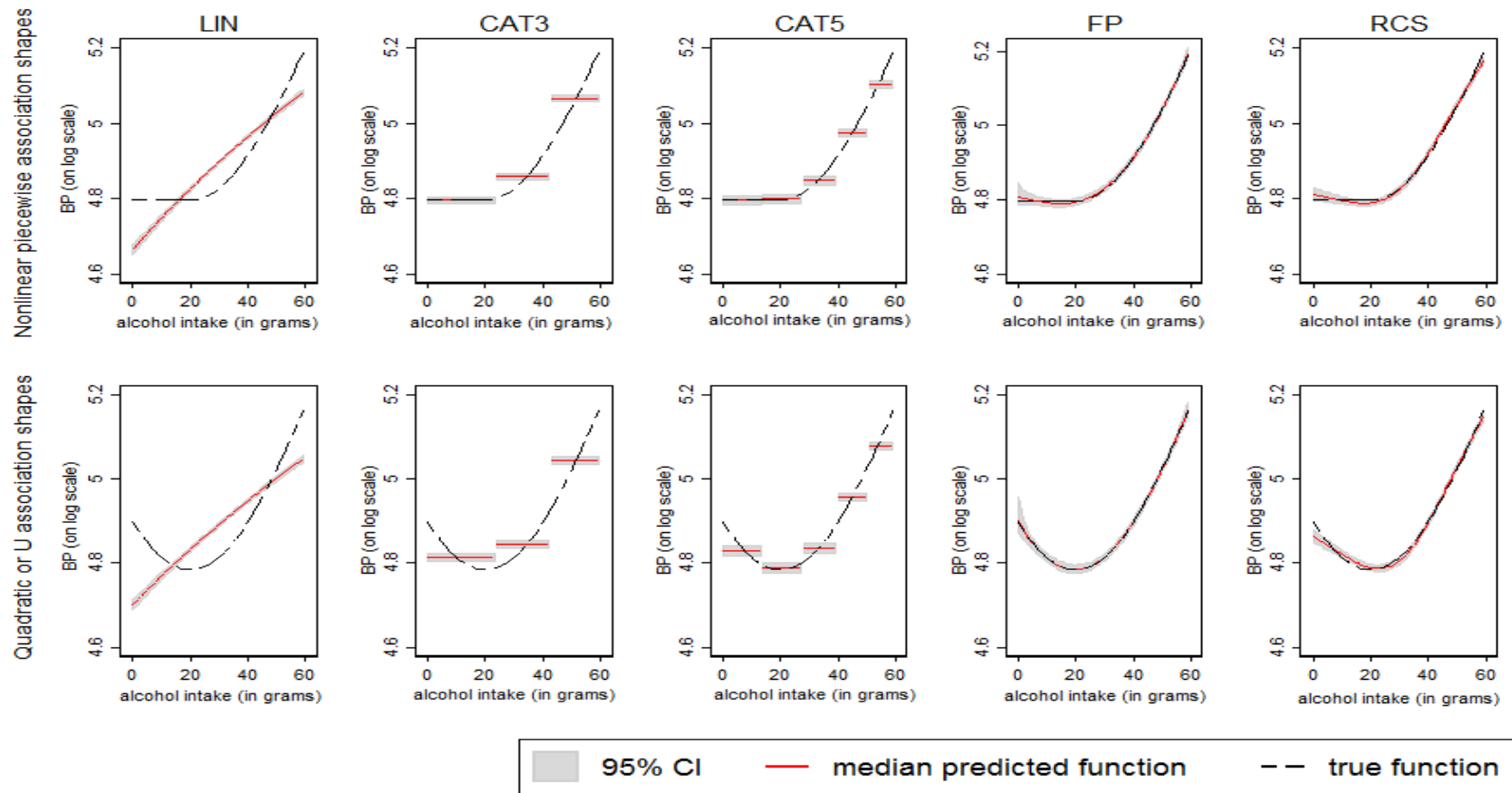


Figure 4.5: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations (replicates) after fitting the nonlinear piecewise and quadratic or U association datasets using the methods of categorisation (CAT3 and CAT5), linearisation, fractional polynomials (FP), and restricted cubic splines (RCS). The results were taken from a sample with 200 observations and moderate noise, $\sigma = 5.0$

4.4.3.2 Prediction of outcomes at the threshold ($c^*=20$) in different association shapes using categorisation, FP and RCS methods

Figure 4.6 & Figure 4.7 provides a summary showing the estimated outcomes and their confidence intervals predicted in the linear piecewise, nonlinear piecewise and quadratic or U association shapes using CAT3, CAT5, FP, and RCS models. The predicted estimates were obtained from the median predicted functions with 200, 1000 and 10000 observations respectively, assuming moderate noise, $\sigma = 5.0$ in the data.

In all the association shapes, the optimal outcome was assumed to occur at 20 units of the exposure (that is at $c^* = 20$). In the linear piecewise threshold association shapes, 20 units of exposure yield 121 units of the outcome. Based on these assumptions, CAT3 was the only method of analysis that provided accurate predictions on the ‘true’ optimal outcomes. The latter was true across all the sample sizes considered in the simulation (see Table 4.5 in the Appendix C). Increasing the sample sizes improved the estimated outcomes and also narrowed their CIs (see Figure 4.6). At $n = 200$, the outcome predicted when $c^* = 20$ was 121.13 (CI = 120.01, 122.26) units and this improved to 121.00 (CI = 120.82, 121.17) when n was large ($n = 10000$). However, the performance of this method and its ability to predict the ‘true’ optimal outcome was coincidental. The CAT3 method performed better because when the simulations were initially set-up, the ‘true’ optimal outcome was placed at the lower exposure category. In contrast, the CAT5, FP and RCS models overestimated the ‘true’ optimal outcome when $c^* = 20$ units – producing CIs with insufficient coverage on the actual outcome (see Figure 4.6). Instead, the ideal turning points under the FP and RCS methods were observed when the exposure was lower (see Table 4.5 in the Appendix C and Figure 4.4). For instance, when $n = 200$ and $\sigma = 5.0$, the optimal exposure was shifted to the left – reducing from 20 units (assumed) to 6 (CI=0, 11) units when

applying the FPs. At 6 units of the exposure, the corresponding outcome occurred at 119.73 (CI = 117.95, 121.28). Similarly, the RCS model attained its optimal outcome of 118.36 (CI = 116.27, 120.39) when the exposure was at zero. Apart from these results, an optimal outcome of 121.00 (CI=119.52, 122.39) was attained within 0-23 level of the exposure under the CAT5 method (see Table 4.5 in Appendices).

Under the nonlinear piecewise association function it was also assumed that 20 units of exposure attain an optimum outcome of 121 units. Except when the sample was small ($n = 200$), the RCS model was the only method of analyses that struggled to estimate the 'true' optimal outcome when the exposure was at 20 units (see Figure 4.6). In contrast, when $c^*=20$ units; the CAT3, CAT5 and FP methods produced estimates that were closer to the 'true' outcome (with CIs that offer sufficient coverage on the actual value). The latter was more noticeable in larger samples where more than 200 observations were considered (see Figure 4.6 for a summary of these results). Due to the same reason provided earlier, the CAT3 method was able to accurately predict the optimal outcome - the estimate coincided with the lower category of the exposure where the optimal value was placed during the simulation design. Contrary to the assumptions made in the simulations, the FP and RCS models predicted their optimal turning points at the lower exposure. For example, when $n = 200$ & $\sigma = 5.0$, the FP regression model attained its optimal outcome at 120.20 (CI = 119.00, 121.43) when the exposure was 15 (CI = 6, 18) units. For any simulation conditions, the RCS models predicted the optimal exposure at 18 units (with varying confidence intervals). For $n = 200$ & $\sigma = 5.0$, the suggested optimal outcome corresponding to 18 (CI = 15, 20) units of exposure was at 119.98 (CI = 119.06, 120.88) when fitting the RCS model. In comparison to the latter results, the CAT5 produced an optimal outcome at 120.76 (CI=119.43, 122.01) within the exposure category of 0-27 (see Table 4.6 in the Appendix C).

Overall, the FP and RCS models were underestimating the optimal outcomes in threshold datasets - shifting the position of the exposure to the left.

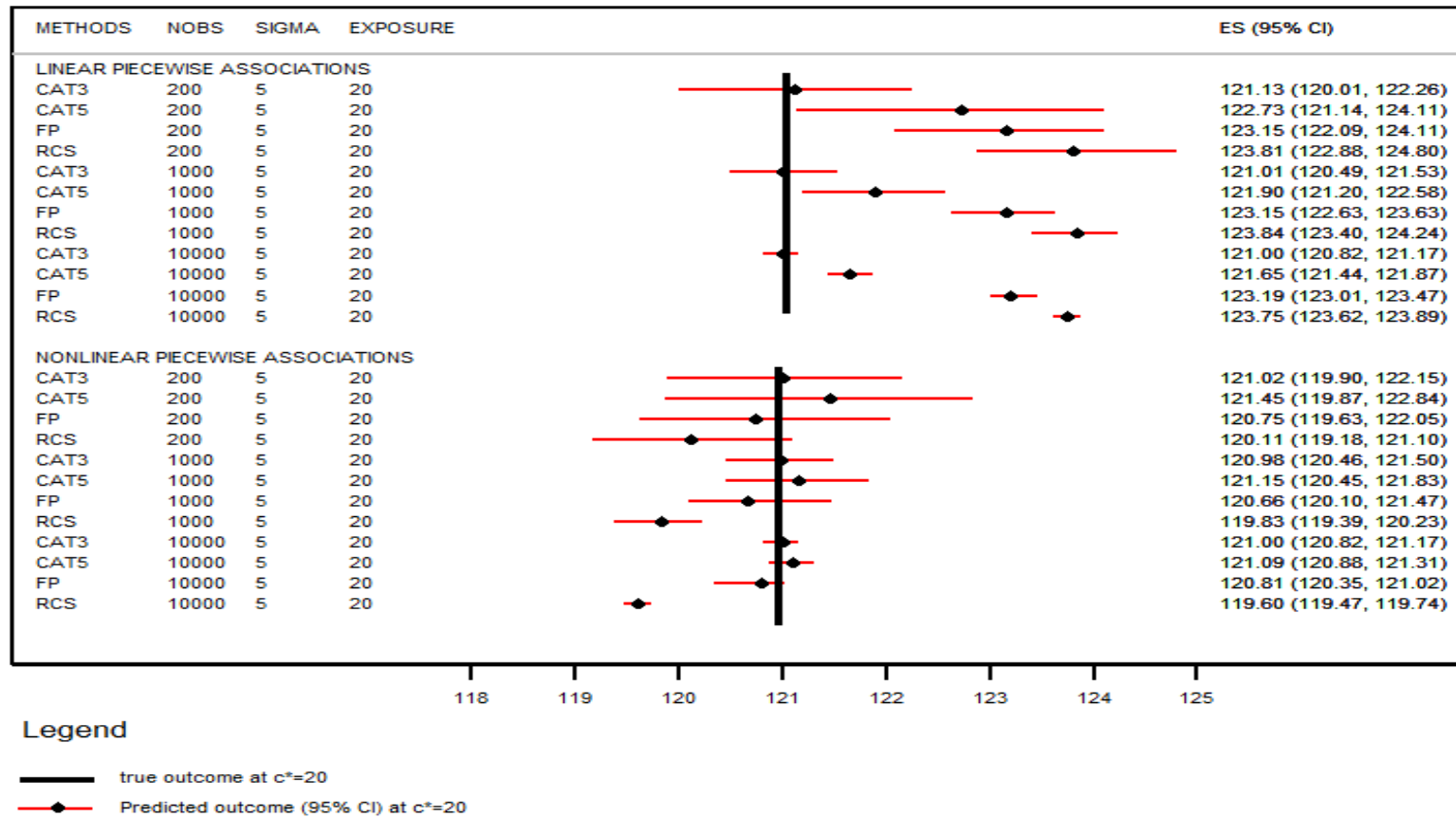


Figure 4.6: The outcome (at $c^*=20$) predicted by fitting the CAT3, CAT5, FP and RCS regression models in linear and nonlinear threshold datasets¹.

¹ True outcome equals to 121 units in the two threshold datasets (when $c^*=20$).

Figure 4.7 below compares the optimal outcomes predicted under the quadratic or U association shapes using the CAT3, CAT5, FP and RCS models. When the sample was small ($n = 200$) and the noise was moderate ($\sigma = 5.0$), the CAT3 method predicted the optimal outcome at 123.05 (CI = 121.93, 124.18) units at the lower exposure group (0-23 units) (see Table 4.7 in the Appendix C). In Table 4.1, the assumption was that 20 units ($c^* = 20$) of the exposure yield 119.6 units of the outcome (after first order differentiation). Given the latter, the CAT3 method overestimated the optimal outcome - producing CIs that do not cover the actual outcome in the simulation. When the sample was increased to $n = 1000$ or $n = 10000$, the predicted optimal outcome increased further away from the true value of 119.6 units resulting in 123.72 (CI = 123.19, 124.23) and 124.32 (CI = 124.14, 124.49) units respectively (see Figure 4.7). These results suggest the inability of the CAT3 method to produce reliable optimal estimates under the U or quadratic association datasets. However, there were considerable improvements when the number of categories was increased in the analysis. The CAT5 method predicted optimal outcomes closer to the 'true' in the simulations. For example, when $n = 200$ & $\sigma = 5.0$, the CAT5 method predicted the optimum outcome at 120.33 (CI=117.83, 122.58) units at the second exposure category (14-27 units). Similarly, large samples yielded closer estimates but with narrow CIs that do not include the true outcome assumed in the simulation. See Figure 4.7 and Table 4.7 (in Appendix C) for further details. In contrast, the FP model offered precise optimal estimates in the quadratic association datasets. The FP models accurately predicted the optimal exposure at 20 units in the simulation. For example, when $n = 200, 1000, 10000$ and $\sigma = 5.0$ the optimal outcomes predicted at the optimal exposure of 20 units were 119.67 (CI=118.53, 121.08), 119.59 (CI=119.14, 120.07) and 119.60 (CI=119.47, 119.74) respectively (see Table 4.7 in Appendix C). These estimated optimal outcomes were very close to the true value of 119.60 units

(with CIs that provided adequate coverage regions). From these results, increasing the samples in the simulations improved the precision of the optimal outcome estimates - narrowing the width of the predicted CIs. Under the RCS models, the predicted optimal exposure occurred at 23 units for all sample sizes considered in the simulations (see Table 4.7 in the Appendix C). Based on the latter, the position of the optimal estimates was shifted to the right when fitting the RCS model. At 20 units of the exposure, the RCS models predicted CIs with adequate coverage for the 'true' optimal outcomes. However, this was observed in the datasets with large noise, $\sigma \geq 5.0$ (see Table 4.7 in the Appendix C). Although for large sample sizes ($n \geq 10000$), the RCS may still require large noise, greater than 5.0 to accurately capture the 'true' optimal outcome in the predicted CI region. This is because, whenever the sample sizes were increased in the simulations, the CI regions for predicted optimal outcomes were becoming narrow. For example in Figure 4.6, when $n = 10000$ the RCS function had narrow CIs that did not include (or provide sufficient coverage) the 'true' optimal outcome. In contrast, when the noise was large, $\sigma = 7.5$, the estimated CI regions provided adequate coverage for the 'true' optimal outcome (see Table 4.7 in the Appendix C).

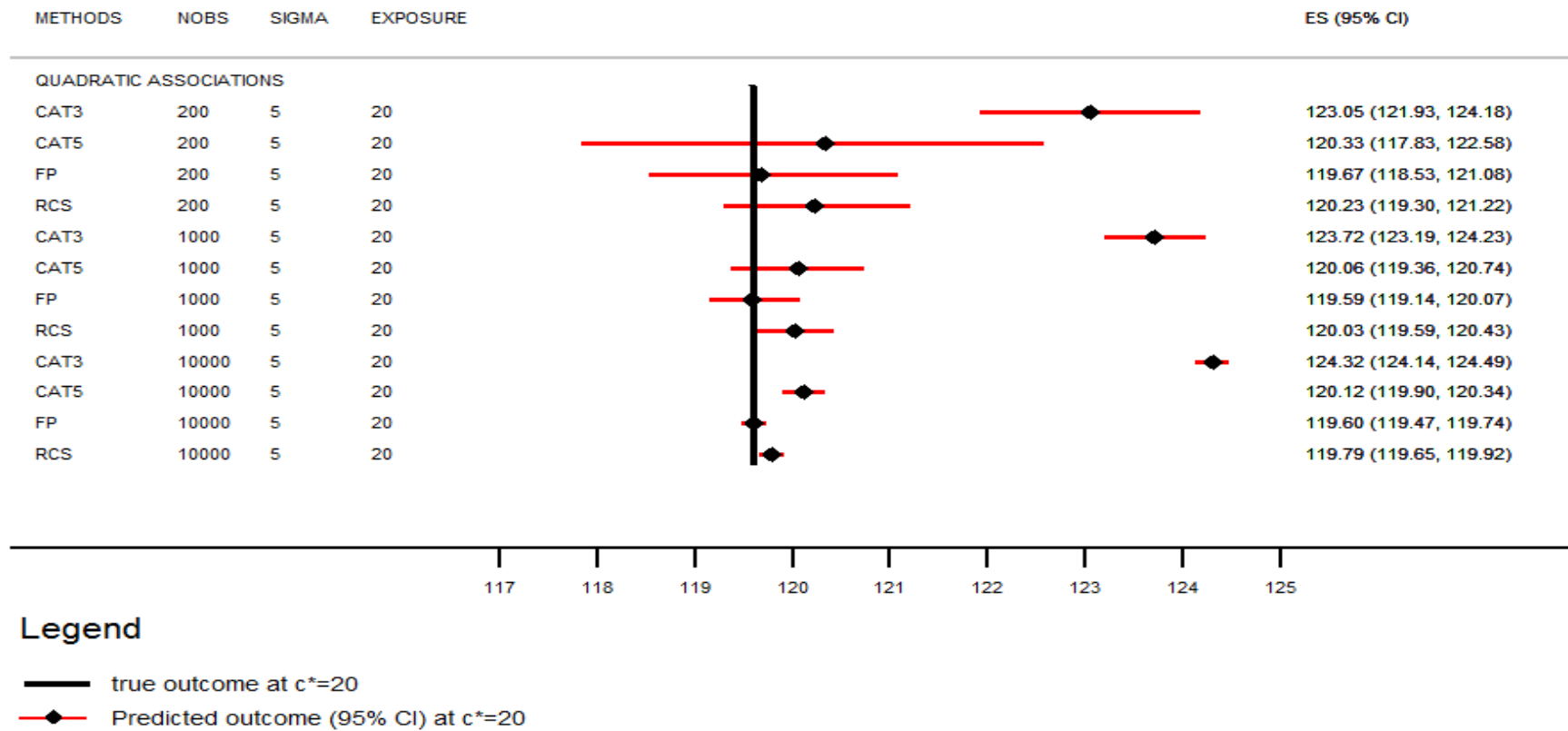


Figure 4.7: The outcome (at $c^*=20$) predicted by fitting the CAT3, CAT5, FP, and RCS regression models in Quadratic or U-shaped datasets¹.

¹ True outcome equal to 119.6 units (when $c^*=20$).

Finally, the results on coverage probabilities revealed conservative CI bounds for the optimal outcomes when fitting the CAT3, CAT5, FP and RCS models in the simulations. The calculated coverage probabilities attained with the four methods exceeded the 95% nominal levels assumed in the simulations. The latter was observed across all the nonlinear association shapes and simulation conditions considered in this chapter (results not shown).

4.5 Discussion

This section discusses the general approach, summarises the key results, challenges and limitations, strength and opportunities, novelty and future studies from this chapter. Section 4.5.1 discusses the general approach adopted in this chapter. Then, a summary of key results and how they compare with other studies are provided in section 4.5.2. The discussions of challenges and limitations, strength and opportunities, novelty and future work are provided in sections 4.5.3, 4.5.4, 4.5.5, and 4.5.6 respectively.

4.5.1 General approach

There exist few simulation studies comparing the performances of FP and RCS methods. This chapter was set-up to investigate the properties of these two models comparing them with the methods of categorisation (CAT3 and CAT5) and linearisation in the area of epidemiology. The focus was to establish the ability of these methods in (1) recovering the ‘true’ relationships assumed in the simulations and (2) estimating the ‘true’ turning points or thresholds in the data. To achieve this, several exposure-outcome relationships found in the area of epidemiology were simulated under the ‘normal error’ regression framework (i.e. assuming continuous outcome variables). The simulations were exemplified by using the alcohol-BP relationship scenarios in the literature. For simplicity, single predictor-outcome relationship datasets were generated

– assuming the exposure was measured as a continuous variable. A realistic data structure incorporating other covariates would be difficult to envision in the simulations since many variables influence each other in epidemiology. Chapter 6 of this thesis suggests an appropriate approach of adjusting for other covariates in single predictor models using real-application data.

4.5.2 Summary of main results

Firstly, the performances of the proposed regression methods were assessed based on the RMSE obtained after fitting different exposure-outcome relationship datasets. As expected, the linearisation model performed better (with smaller RMSEs) than the other methods when the exposure-outcome relationship was linear. The two methods of categorisation (CAT3 and CAT5) had the largest RMSEs when applied in the same dataset. Larger RMSEs were attributed to the step-functions produced when the two methods were employed. Under the linear and nonlinear piecewise threshold functions, the fractional polynomial regression models retained smaller RMSE estimates and the restricted cubic spline models followed. However, the two methods struggled to fully identify the true association curves. The CAT3, CAT5 and linearisation approaches produced the largest RMSEs under these curves. The large RMSEs attained using the linear regression and categorical analyses were due to the methods inability to adapt and lack of flexibility in fitting complex nonlinear functions. For instance, both the CAT3 and CAT5 were limited to step-functions whilst linearisation was restricted to linear functions only. The FP regression models also retained smaller RMSE estimates under the quadratic or U-shaped association. The RCS models followed the FP with the next smallest estimates. In contrast, assuming linearity under the quadratic association datasets produced functions with the largest RMSEs. Although the study cannot directly be compared to the findings in this chapter, Govindarajulu et al (2009) conducted a simulation study examining several smoothing

methods for estimating nonlinear exposure-outcome curves that included the FPs and RCS (with 5 knots). Amongst the investigated shapes that included the threshold and quadratic functions, both the FP and RCS functions performed well when fitted in quadratic datasets (with smaller RMSE quantities). In contrast, under the threshold dataset, the RCS performed better (with small RMSE estimates) than the FP function (Govindarajulu et al., 2009). However, these results were not directly comparable with findings in this chapter because the simulations differed in their framework. The simulations by Govindarajulu and colleagues were based on time to event outcomes with rightly skewed exposure variables whilst in this chapter, the outcome was continuous and the exposure was from the uniform distribution.

The RMSE estimates under the CAT3 and CAT5 models were not adversely affected by sample size variation. The two methods of categorisation retained steady RMSEs across the four association shapes when varying the samples in the simulations. For instance, the two methods of categorisation had large steady RMSEs across different samples in linear dataset. In contrast, the linear, FP and RCS models had smaller RMSEs that decreased with the samples in the same data. Apart from that, the linear, FP and RCS methods also had steady RMSEs (like CAT3 and CAT5) when fitted in thresholds (linear & nonlinear) datasets. In quadratic datasets, steady RMSEs were obtained when fitting the linear, CAT3, CAT5 and RCS models. In contrast, FPs produced RMSE estimates that decreased with larger samples in quadratic datasets - a suggestion of good fit. The theory of large numbers suggests a good model should improve fit whenever the sample is increased in the simulation (Siegmund, 2005).

Secondly, the type I error rates in FP and RCS models were assessed based on the assumption that the exposure-outcome relationship was linear (null hypothesis). The simulation results suggested that the FP functions were susceptible to noisy datasets

than the RCS models. Increasing the noise in small datasets produced higher type I error rates (maximum of 25%) amongst the FP models. In contrast, the RCS models produced error rates that were closer to the nominal level of 5%. Based on these findings, the FP method was more likely to produce over-fitted functions - rejecting the 'true' linear relationship more frequently than the RCS approach. In addition, fractional polynomials were too flexible compared to the RCS functions. When the sample sizes were large and there was less noise in the datasets, the type I error rates in FPs were minimal - closer to 10%. This means that the FP models require studies with large samples ($n \geq 10000$) and small noise ($\sigma \leq 2.5$) to improve their statistical power of identifying linearity in exposure-outcome investigations. The findings of high type I error rates under the FP models were also reported by Amber and Royston (Amblar and Royston, 2001). Amber and Royston (2001), found that the FP models ($m = 2$) were anti-conservative (with the maximum type I error of 15%) when the 'true' function was linear.

Under the nonlinear functions, both the FP and RCS approaches performed well to identify the non-existence of linearity in the datasets. There were no type II errors (accepting the null hypothesis of linearity when the actual relationship was nonlinear) committed when fitting the FP and RCS models in threshold and quadratic association datasets. This finding confirms the ability of the FP and RCS models in detecting the presence of nonlinearity in datasets with similar association shapes studied in this chapter.

Thirdly, the performances of fitted models were evaluated against predetermined 'true' exposure-outcome associations using the 95% CI regions. From $R = 1000$ simulations, the interest was establishing whether the mean predicted functions produced using the categorisation, linearisation, FP and RCS approaches were sufficient and provided enough coverage around the 'true' fit. Generally, narrow CI widths of

predicted mean functions were observed when the samples (n) were increased in the simulations. In contrast, increasing the noise (σ) widened the CI width of the predicted curves. The latter was more visible in datasets with small samples. Under the linear exposure-outcome association datasets, the linearisation, FP and RCS approaches produced functions that lied on the ‘true’ functions. In addition, their predicted 95% CIs provided sufficient coverage on the actual fit. However, the FP functions were characterised by wider CIs at the lower tail of the exposure. In application studies, researchers may find this behaviour unattractive forcing them to ignore the use of FP in their investigation since such wider CIs are biologically implausible. However, it has been advised that analysts ignore the portion where the estimated curve have wider CIs – since its unlike to change the interpretation of the fitted curves (Lorenz et al., 2017). Alternatively, the FPs could be fitted at high/large exposure values above zero - by shifting the origin of the data, adding a small constant value to exposure values (Royston and Sauerbrei, 2008). In contrast, the categorisation method produced step-functions that were inadequate for the ‘true’ linear associations in the simulation.

Under the linear and nonlinear piecewise threshold associations, the mean predicted curves obtained using the five methods of analysis did not lend themselves on the ‘true’ association shapes. Also, their CIs produced insufficient coverage on the actual curves - reflecting lack of fit. The FP and RCS struggled to estimate the predetermined ‘true’ threshold or turning points in these datasets. On the other hand, the CAT3, CAT5 and linearisation approaches produced fits that completely missed the true shapes yielding step and linear functions. Under the quadratic association datasets, the FP method was the only approach that produced mean functions that lied entirely on the ‘true’ curve. Similarly, their CIs provided adequate coverage for the actual fit. In contrast, the other methods provided imprecise fits with CIs that did not cover the ‘true’ curve sufficiently. Although the RCS predicted a near approximation curve, its function

struggled to lend itself on the ‘true’ fit at the lower exposures. The results similar to those reported with FP fit under the quadratic association datasets were also reported elsewhere (Strasak et al., 2011).

Finally, the performances of fitted regression models were evaluated against the turning points (or thresholds) in the simulations. Both the FP and RCS regression models produced misleading turning points in linear threshold association datasets. The ‘true’ position of the optimal exposure was underestimated and shifted to the left in both methods – suggesting protective effects at the lower exposure. Moreover, the estimated optimal exposure varied in the two methods; FP models were likely to estimate higher optimal exposure compared to the RCS functions. This has a huge implication in applied health research studies; different optimal exposure values often reported in application studies such as alcohol and blood pressure outcomes may be due to these methodological variations. Therefore, researchers are advised to carefully consider the properties and behaviour of these modelling techniques before investigating the turning points in exposure-outcome studies.

In nonlinear piecewise threshold datasets, the predicted optimal exposure under the RCS models was closer to the ‘true’ value than when fitting the FPs. However, the predicted optimal exposure was shifted to the left of the actual estimate in both models (appears at the lower exposure values) also underestimating the optimal outcomes.

For the quadratic or U-shaped association, the FP model provided precise estimates of the threshold in the simulations. The RCS misspecified the actual turning point of the exposure shifting it to the right.

Both the CAT3 and CAT5 were inefficient for estimating the exact positions of ‘true’ turning points in the simulations. The application of these methods forced the predicted outcomes to remain invariant within categories of the exposure producing step

functions - where adjacent categories of the exposure determined the changes in the predicted outcomes.

4.5.2.1 The difference between CAT3 and CAT5 approaches

The features of exposure-outcome relationships were masked when applying the CAT3 and CAT5 models in the simulation. The outcomes predicted in the two models were submerged and varied according to exposure categories. Thus, the number of exposure categories influenced how the predicted outcomes varied in such models. For instance, the CAT3 allowed the outcomes to vary within few or wide categories whilst the CAT5 had more or narrow categories. From the latter, the CAT5 also needed more parameter estimates (or degrees of freedom) than the CAT3 or other methods in the simulations. Hence, performance improvements from the categorical models with large number of categories come as a trade-off for more complex functions which are generally inefficient or unstable, e.g. using >10 categories to approximate a non-linear relationship.

4.5.3 Challenges and limitations

This section discusses the challenges and limitations of this simulation chapter.

4.5.3.1 The simulations not entirely inclusive

It was beyond the scope of this chapter to assess the performances of CAT3, CAT5, linearisation, FP, and RCS models in all the relationships functions found in epidemiology. There exist other plausible association shapes that were not investigated in this simulation chapter. Some example of the association shapes (with turning points) found in the literature include J-shapes (Lawlor et al., 2003, Jee et al., 2006, de Gonzalez et al., 2010), asymptotic shapes (Vesey et al., 1982) and sigmoidal/S-shaped/logistic functions (Pinheiro and Bates, 2000). However, I expect the properties and performances of these methods not to vary much when applied in some of these functions. For example, the RCS model is likely to struggle to lend itself (or fail to

accurately predict the position of turning points) on the J-shaped function at the lower exposure since the RCS model predict linear fits at the tails. In contrast, FP models are likely to lend themselves on the J-shaped functions and accurately predict the locations of the turning point in the datasets since both the J and U-shaped curves belong to quadratic functions - easily be fitted by the FP functions.

4.5.3.2 The normal error distribution

Binary and time to event outcomes are more common in epidemiological studies than continuous outcomes (Bender, 2009). Thus, the findings in this chapter may not be too generalisable. A future work focusing on binary outcomes is needed for generalisability. The existing simulation studies investigating the performances of FP and RCS in different exposure-outcome relationships have mostly focused on time to event outcomes using Cox regression models (Hollander and Schumacher, 2006, Govindarajulu et al., 2009, Keogh et al., 2012). However, it was not possible to make direct comparisons between studies and findings in this chapter since the simulation framework or settings were different.

The assumption of normality on the outcome variable was also likely to enhance the performances of predicted models in the simulations. According to Suissa (1991), continuous outcome models are disadvantaged by their complete reliance on the assumption of normality for the data. Hence, some deviations away from the Gaussian distribution would likely cause some uncertainties in the simulations.

4.5.4 Strengths and opportunities

Assessing the performances of CAT3, CAT5, linearisation, FP, and RCS models based on different exposure-outcome scenarios was possibly not going to be achievable with real data alone. The simulations in this chapter offered that opportunity – covering several scenarios that are usually difficult to evaluate with real application studies. The simulations evaluated the performances of these methods based on several exposure-

outcome relationship datasets characterised by (1) varying noises and sample sizes and (2) pre-determined turning points (or thresholds) quantities. Overall, the settings in this simulation chapter offer the following:

- i. A simple guide to set-up similar simulation studies in the area of epidemiology
- ii. An insight on the properties of CAT3, CAT5, linearisation, FP and RCS models when fitted in linear, thresholds and quadratic association datasets. The properties of these methods were evaluated based on several measures suitable for normal error models including the RMSE, type 1 errors, and coverage probabilities.
- iii. An inferential guide on the precision of the FP and RCS models when estimating the positions of thresholds or turning points in nonlinear association datasets.
- iv. A guide on the appropriate or suitable models for fitting different association shapes considered in the simulations.

4.5.5 Novelty

Despite the practical relevance of this topic, the researcher was not aware of any simulation study that assessed the performances of FP, RCS, CAT3, CAT5 and linearisation methods against thresholds or turning points in exposure-outcome relationships. Pastor and Guallar (Pastor and Guallar, 1998) have argued that often, the researchers visualise and approximate the location of the thresholds or turning points based on their predicted exposure-outcome shapes. The practice is largely subjective and may lead to inconsistent estimates for thresholds or turning points. Evidence in this chapter suggests that FPs and RCS methods (used to investigate exposure-outcomes relationships) may generally be useful for revealing the ‘true’ association shapes in the data and still fail to detect the positions of ‘true’ threshold points (or fail to achieve both). These are key finding that may be useful to researchers interested in exposure-outcome studies and to those reporting thresholds. In addition, this chapter

demonstrated the properties of CAT3, CAT5, linearisation, FP and RCS methods, exhibiting the ability of simulation modelling in several exposure-outcome relationships. In reality, investigating the properties of these methods with real data alone would generally be hard. Hence, the simulation approach was a novel idea adopted by the researcher to achieve the objectives in this chapter. Plausible exposure-outcome relationships often observed in epidemiology were simulated to ensure the findings in this chapter are practically relevant.

4.5.6 Future work

Although the present simulations were based on continuous outcomes, epidemiologists favour binary outcome models for reporting the occurrence of diseases or events in their studies. Often, binary outcome models are achieved by dichotomising continuous outcome variables when analysing the data. For example, serum creatinine above/below 1.4 mg/dL may define the presence/absence of abnormal renal function (Culleton et al., 1999). Taking the latter scenario into consideration, the present work is incomplete without the assessment of binary outcome models. Further simulation work is proposed in Chapter 5 to assess the performance and choice of categorisation, linearisation, FP and RCS approaches for handling continuous predictors in prognostic models under the binary outcome setup. The potential limitation with the results from the normal error models is that they are completely dependent on the Gaussian distribution assumed in the data. Thus, any deviations from the assumption of normality could cause more uncertainties in the estimates (Suissa, 1991).

4.6 Conclusions

Categorisation and linearisation methods performed poorly when nonlinearity was present in exposure-outcome relationships. The categorisation methods distorted trends in the data producing step functions with large RMSE estimates. The latter

occurred in models with few and more categories hence these findings are generalisable. The linearisation approach only worked well in linear association datasets. Fitting linear regression models when the associations were nonlinear resulted in misleading functions. Thus, these approaches are not appropriate for modelling complex nonlinear exposure-outcome associations. The results could worsen when clinical thresholds are required for decision making. The two methods have no ability for detecting features such as turning points in the datasets. When investigating their properties against nonlinearity and thresholds functions, the two approaches failed to identify the presence of nonlinearity and ‘true’ thresholds in the simulations.

Alternative methods such as FP and RCS have been suggested for modelling nonlinear exposure-outcome relationships. The FP produced high type I errors when applied in linear association datasets due to its flexibility and being more susceptible to fluctuations. Compared to FPs, the RCS models were generally more conservative and adaptive. The latter was not adversely affected by variation of noises and samples in the data.

The FP and RCS models also failed to accurately predict ‘true’ turning points (or thresholds) present in some nonlinear association datasets considered in the simulation. Under the quadratic association datasets, the FP function performed better than the RCS regression model - producing fits that accurately predict the ‘true’ thresholds in the data. In contrast, the RCS models produced near approximation estimates. In the other nonlinear curves considered in the simulations, the two methods produced varying thresholds estimates when applied in similar datasets. Clinically, this has huge implications; researchers need to be cautioned about the inconsistencies of estimating clinical thresholds using these methods.

Finally, as a lesson learned, it was important to recognise that the application of flexible regression models such as the FP and RCS would not always yield accurate results when estimating the actual turning or threshold points in the datasets (especially when the association shapes are unknown). Therefore, as a minimum check, it is recommended that these regression models be used together with the traditional methods such as linearisation and categorisation approaches to verify the existence of nonlinearity in the datasets. If both RCS and FP analyses provide evidence of abrupt changes in risk then there could be reasons to suspect the existence of nonlinearity and turning points in the datasets. In addition, if the FP model suggests U or J-shaped associations (or there are a priori reasons), then the FP method could be used to estimate the turning points together with their standard errors.

Chapter 5

Extensions to prognostic models with binary outcomes – a simulation study

5.1 Introduction

In Chapter 4, a simulation study was performed to study the performance of categorisation, linearisation, FP, and RCS approaches in normal error models - assuming continuous variables in predictor-outcome relationships. This chapter focuses on the situations with binary outcomes (e.g. disease recurred or did not; patient lived or died). The survey research in Chapter 3 suggests medical researchers favour binary outcomes in their studies to explain causal relations with other variables. Besides explaining causal relationships, binary outcome models could also be used for prediction purposes. Although similar in structure, predictive and explanatory models are different. Predictive models inform clinicians about the patient's health outcome (or prognosis) whilst explanatory models are mainly focused on explaining the cause of an event outcome (González-Ferrer et al., 2017). In epidemiology, predictive models are rarely investigated compared to explanatory models. Thus, this chapter concentrates on predictive models with binary outcomes – also known as ‘prognostic models’ in the literature (Steyerberg et al., 2013).

One issue of concern when developing prognostic models is the assessment of their predictive accuracy. The ability to accurately predict the occurrence of the event or disease outcomes on the basis of continuous risk factors is an important modelling step often overlooked by analysts (Collins et al., 2016). Traditionally, continuous predictor variables are treated as dummy variables (after categorisation or grouping) or linear terms when developing such models (Sauerbrei and Royston, 1999). However, the two

approaches of categorisation and linearisation may be inadequate or limited in characterising the unknown relationships – producing inaccurate predictions (Rosenberg et al., 2003). For example, if nonlinearity is present in the data, these practices could simplify relationships in the data restricting analysts to work with functions that are inappropriate for the final model. Additional problems associated with the method of categorising continuous predictor variables in medicine has been discussed in many research articles (Richardson and Loomis, 2004, Royston et al., 2006, Froslic et al., 2010, Baneshi and Talei, 2011). Unfortunately, there exist few studies examining the performance of prognostic models and explaining the effects and choices of handling continuous predictor variables in medical research to guide non-statisticians (Collins et al., 2016).

The two recent studies comparing the performance of prognostic models using the methods of categorisation and linearisation against alternative approaches involving fractional polynomials (FP) and restricted cubic splines (RCS) were performed by Nieboer (Nieboer et al., 2015) and Collins (Collins et al., 2016). Nieboer and colleagues focused on logistic prognostic models comparing FP and RCS against the method of linearising the predictor in four nonlinear association datasets. The researchers did not examine the models categorising the continuous predictor variables. Going further, this thesis chapter aims to compare the performance of FP and RCS models against the methods of categorisation and linearisation using a simulation study. To maintain consistency and continuity in the thesis, similar nonlinear shapes considered in Chapter 4 were also investigated here - focusing on logistic prognostic models. Collins and colleagues compared these four approaches using Cox regression based prognostic models predicting 10-year risk of cardiovascular disease and hip fracture in two cohort datasets. The authors used a resampling strategy (random sampling with replacement) to examine and validate the performance of FP, RCS, categorisation, and linearisation for

handling continuous predictors in prognostic models. However in the two studies of Nieboer et al., (2015) and Collins et al., (2016), the true association between the continuous predictors and the outcomes were unknown. Moreover, the properties of these methods against clinical features such as turning points and thresholds were omitted. In this chapter, the assessments and comparisons of FP, RCS, linearisation, and categorisation methods were performed assuming the log odds functions or shapes are known. This procedure allowed this research to infer on predicted models against the overall underlying ‘true’ functional relations. In the process, the clinical features such as turning or threshold points attained using these models were reported and evaluated for precision. The predictive ability of prognostic models developed using FP, RCS, linearisation, and categorisation methods were assessed and compared using two key measures of discrimination and calibration recommended for reporting in all prediction models (diagnostic or prognostic) (Moons et al., 2015). Net benefits (Vickers and Elkin, 2006, Kerr et al., 2016) curves were also used as another useful measure in the simulations. The details and explanation of these performance measures are provided in sections 5.3.1.2 to 5.3.1.4.

The specific objectives of this chapter are outlined and summarised in section 5.1.1. Sections 5.2, 5.3, 5.4, 5.5, and 5.6 describe the simulation framework, methods, results, discussion, and conclusions respectively of this chapter.

5.1.1 Aims and objectives

The main aim of this chapter was to investigate and compare through simulations the performance and choice of approaches including FP, RCS, linearisation, and categorisation for handling continuous predictors in prognostic models often reported in epidemiology. The research focused on binary outcome models fitted using logistic functions. The following specific plots and measures were quantified and

presented to evaluate and compare the performances of various prognostic models in the simulation:

- i. The median predicted prognostic plots/curves showing the relation between a continuous predictor and the occurrence of an outcome from various approaches were summarised for comparison based on 1000 simulations. Further, the precision of median predicted prognostic models in estimating the actual turning points or thresholds were summarised and reported with their corresponding 95% confidence intervals estimates.
- ii. The ability of the models to differentiate between patients with an outcome and those without an outcome were evaluated using the median c-index scores from the area under the ROC (Receiver operating characteristic) curve (AUC). Median estimates of the c-index together with their 95% confidence intervals were reported from 1000 simulations to summarise model performance in various datasets.
- iii. The influence of various approaches used to handle continuous predictors was evaluated through calibration plots. The median observed probabilities in 1000 simulations were plotted against predicted probabilities to assess their agreement.
- iv. Finally, the prognostic models attained using the FP, RCS, linearisation, and categorisation approaches were assessed using the decision analysis curves for clinical usefulness (measured using median net benefits curves) at various probability thresholds.

5.2 Monte Carlo simulation framework

A single predictor-outcome relationship data structure was proposed in the simulation to investigate and compare the properties and performances of FP, RCS,

categorisation and linearisation methods suggested for handling continuous predictors when developing prognostic models. The outcome variable (y) was assumed to be a binary event and various prognostic models were developed using these methods for comparison by applying logistic regression models. The aim was to investigate these methods assuming the existence of linear and nonlinear relationships between the continuous predictor variable (x) and the $logit(\pi_i) = f_i(x)$. To be practically relevant, the descriptions of the simulation set-up guided by example scenarios in the field of epidemiology are provided in sections 5.2.1 and 5.2.2 respectively.

5.2.1 Simulation set-up

For continuity from the previous chapter, let's assume the continuous predictor variable (x) was drawn from a uniform distribution with a range of values between 0 and $\max(x)$ and the outcome variable (y) was generated from a random binomial distribution such that the event of interest takes values 1 and 0 otherwise. Let $y = 1$ represent the presence of an event and $y = 0$ its non-existence. Then for each individual in the simulated datasets, $y_i \sim binomial(1, \pi_i)$ where π_i are observed mean probabilities of an event, $y = 1$ written as $P(y = 1|x)$. The observed mean probabilities (π_i) relate to \mathbf{X} through the canonical link function known as the logit expressed as follows:

$$logit(\pi_i) = \log_e \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \mathbf{X}\beta = \beta_0 + \sum_{j=1}^k \beta_j x_j = f_i(x), \quad \text{Eq. 5. 1}$$

where $\left(\frac{\pi_i}{1-\pi_i} \right)$ is the odds of an event with $0 \leq \pi_i \leq 1$ and $f_i(x)$ is the mean function relating the continuous predictor variable (x) to the logit. Possible mean functions include the alcohol-hypertension relationship example scenario provided in section 5.2.2.

5.2.1.1 Computation of estimates

5.2.1.1.1 Observed probabilities

Assuming the logit model in Eq.5.1 is known, the observed mean probabilities can be calculated as follows:

$$\pi_i = P(y = 1|x) = \exp(f_i(x)) / (1 + \exp(f_i(x))), \quad \text{Eq. 5. 2}$$

Such that for n independent random observations corresponding to y_1, y_2, \dots, y_n , the probability function of y_i is given by:

$$f_i(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \text{ where } y_i = 0 \text{ or } 1, i = 1, 2, \dots, n \quad \text{Eq. 5. 3}$$

with the probability of having an event given as π_i and variance as $\pi_i(1 - \pi_i)$ (McConnell and Vera-Hernández, 2015).

5.2.1.1.2 Observed odds ratio curves

To construct the observed odds ratio curves the $f_i(x)$ are transformed such that

$$\pi_i(x) = \exp(f_i(x)), \quad \text{Eq. 5. 4}$$

where x is the predictor value for a single continuous variable considered in the simulations.

The observed odds ratio curves are estimated using FP, RCS, categorisation, and linearisation models and plotted point by point of the predictor with a range of values between 0 to $\max(x)$ for R simulation (replication). After R simulations, the 50th percentile curve data was obtained to represent median predicted functions under each method. The data for predicted median functions included their 95% confidence intervals obtained using the 2.5th and 97.5th percentile points of the empirical distribution in the simulations. Note: The R simulations should be sufficient with minimum Monte Carlo error (MCE) (Koehler et al., 2009). The simulation was

performed using a Stata program assuming similar procedures in section 4.2.2.1 of Chapter 4 (varying observations inside the loop). See Appendix B for Stata codes.

5.2.2 Extension of the alcohol-blood pressure example

Based on the alcohol-blood pressure relationship example studies reported in Chapter 4 (section 4.2.2.2), suppose the interest was on developing prognostic models for hypertension patients treating alcohol consumption as a predictor variable (x). In this scenario, let the outcome (y) be a binary variable taking value 1 for patients with the disease (hypertension) and 0 for non-diseased patients (no hypertension). Furthermore, suppose for each individual in the simulation, a binary outcome was generated through a random binomial distribution with observed mean probability π_i given in Eq. 5.2, where $\text{logit}(\pi_i) = f_i(x)$ and x was drawn from a uniform distribution with a range of values between 0 and 60 grams for alcohol consumption. That is, $y_i \sim \text{binomial}(1, \pi_i)$ where $\text{logit}(\pi_i) = f_i(x)$ are logit functions. The examples of logit functions $f_i(x)$ could include those in Table 5.1 representing alcohol-hypertension relationship shapes reported in epidemiological studies.

In the simulation, the alcohol-hypertension association datasets in Table 5.1 were generated assuming $N = 1000$ individuals (observations). Furthermore, the datasets were replicated $R = 1000$ times to compare various logistic prognostic models attained with FP, RCS, categorisation, and linearisation approaches. Replication of the samples 1000 times yield MCE of approximately 3% (Koehler et al., 2009).

To attain reasonable prediction functions, the values of β_0 and β_i 's in Table 5.1 were chosen such that the disease outcome was approximately 10% in $N = 1000$ observations. The reported prevalence of hypertension varies across the world. A systematic review has previously reported the lowest prevalence of hypertension in rural India and the highest in Poland at 3.2% and 72.5% respectively (Kearney et al., 2004).

Table 5.1: Proposed linear and nonlinear logit functions used in the simulation to compare various approaches of handling the continuous predictor (x) when developing prognostic models.

Type of associations	Logit function equations	Logit functions
Linear	$\log_e \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 x = f_1(x)$	$f_1(x) = -2.5 + 0.01x$
Linear piecewise threshold	$\log_e \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1(x - c^*) = f_2(x)$	$f_2(x) = \begin{cases} -2.9 & \text{if } x \leq 20 \\ -2.9 + 0.045(x - 20) & \text{if } x > 20 \end{cases}$
Nonlinear piecewise threshold	$\log_e \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1(x - c^*)^2 = f_3(x)$	$f_3(x) = \begin{cases} -2.9 & \text{if } x \leq 20 \\ -2.9 + 0.0015(x - 20)^2 & \text{if } x > 20 \end{cases}$
U-shaped or Quadratic	$\log_e \left(\frac{\pi_i}{1-\pi_i} \right) = \beta_0 + \beta_1 x + \beta_2 x^2 = f_4(x)$	$f_4(x) = -2.2 - 0.0128x + 0.00032x^2$

Using the example simulated datasets proposed above, the next section 5.3 describes various approaches of handling alcohol consumption (measured in grams) when developing logistic prognostic models for hypertension patients. Various performance measures used to evaluate these prognostic models were also described in section 5.3.1.

5.3 Approaches for handling continuous predictors

The common approaches of handling continuous predictor variables when developing predictor-outcome prognostic models include the methods of categorisation and linearisation. These methods are compared to alternative prognostic models attained with fractional polynomials and restricted cubic spline approaches.

In the simulation, the method of categorisation was implemented by using tertile and quintile values of the predictor (alcohol consumption) to form models with three (CAT3) and five (CAT5) categories respectively. Tertiles and quintiles are common way of establishing categories in medical studies (see Chapter 3). Apart from the latter, prognostic models were developed by assuming linearity between alcohol consumption and the disease outcomes. This simple approach is known as linearisation in this chapter.

Alternatively, the method of fractional polynomials that involves limited but flexible sets of transformations defining the relationship between alcohol consumption and hypertension was applied for model development. In the simulation, FP modelling was performed using the standard implementation procedures available in the Stata program. The second degree FPs ($m = 2$) that occurs as a default in Stata offers a wide range of shapes sufficient to cover the four association functions in the simulations. The final approach involves developing logistic prognostic models using restricted cubic

splines (RCS) (Desquilbet and Mariotti, 2010). The alcohol consumption was treated non-parametrically such that the developed model has 3 knots placed at 10th, 50th and 90th percentiles of the observed measurement distribution. Since the association datasets considered in the simulation has few turning points (< 2), RCS with 3 knots should provide adequate prognostic models. Moreover, the RCS models with 3 knots have smooth functions than the models with greater knots (> 3). Thus, the RCS functions with 3 knots were also suggested for use to achieve smoothness and avoid over-fitted models.

5.3.1 Model evaluation

The predictive accuracy and performances of predicted prognostic models attained with FP, RCS, categorisation and linearisation approaches were evaluated based on (1) the ability to discover ‘true’ relationship between binary outcomes and the predictor variable, (2) the ability of predicted models to differentiate between subjects with an outcome and those without an outcome, (3) the influence of various approaches in handling continuous predictor variables and (4) the clinical utility. To achieve (1) the performance of predicted prognostic models were assessed and compared using graphs examining the overall fit and their predicted turning points estimates. Discrimination, calibration, and decision curve analysis graphs were used to draw inference on point (2) to (4) above. The suggested methods of model performance including discrimination, calibration, and decision curve analysis are recommended for reporting and evaluation of prediction models in medicine for individual prognosis or diagnosis (Moons et al., 2015). The description of these measures and how they were presented is provided below.

5.3.1.1 The median predicted functions, optimal rates, and their 95% confidence intervals

To construct the median predicted functions from various modelling techniques after 1000 simulation (replications), the 50th percentile points of the event outcome (hypertension) probability distribution were obtained and plotted against predictor values (alcohol intake) to represent the average fit. The median predicted functions were then compared by overlaying each fit against the ‘true’ model in the simulations. With these plots, it was possible to make an assessment on whether the methods under investigation have the abilities to produce identical fits as the truth or not. The 95% confidence intervals for the median functions were also presented to assess how each modelling technique was affected or influenced by the distribution of the data. In 1000 simulations, the 95% CI region was represented by the 2.5th and 97.5th percentile points of the event probabilities plotted along the predictor scale.

Furthermore, to assess the accuracy and performance of predicted functions, the median optimal event (hypertension) probabilities and their 95% CIs were estimated for comparison with the true values assumed in the simulation. The optimal estimates were defined as points where the predicted functions attain the minimum event probability. In the simulation, 1000 estimated optimal event probabilities were obtained and the 50th percentile point of the distribution represented the median optimal rate. The 2.5th and 97.5th percentile distribution of the optimal probabilities represented their 95% CIs.

Another useful measure considered in the evaluation of predicted models was the coverage probability. The success rate of CIs was measured by their ability to provide coverage of the ‘true’ optimal estimates across the different logit models proposed in Table 5.1. From 1000 simulations, the coverage probability was the fraction of time the confidence interval contains the true optimal rate (White, 2010). Given this, the coverage probabilities based on nominal 95% confidence intervals for the median

optimal event (hypertension) rate (denoted by $\tilde{\pi}_i$) was calculated from the estimated functions as $\tilde{\pi}_i \pm Z_{\alpha/2} * \widehat{SE}(\pi_i)$ where $Z_{\alpha/2}$ is the critical value from the standard normal distribution and $\widehat{SE}(\pi_i)$ are the standard errors of the estimated probabilities of the event outcome in each iteration (Zhao and Kolonel, 1992, Dalen et al., 2009). The $\tilde{\pi}_i$ and $\widehat{SE}(\pi_i)$ were collected at each iteration. For a 95% CI, its coverage was said to be correct if it includes the true parameter 95% of the time. If the coverage probability was less than the 95%, then the CI was said to be narrow with small standard errors. Or else the CIs are too wide with large standard errors (White, 2010).

The descriptions of how other performance measures are applied in this Chapter are provided below. The proposed measures were considered for evaluation because they are recommended for reporting and evaluation of prediction models in medicine (Moons et al., 2015).

5.3.1.2 Discrimination

This is a key aspect of model performance when working with logistic regression. It is defined as the ability of the model to differentiate between patients with the event outcome and those without (Royston and Altman, 2010, Collins et al., 2016). It was evaluated through the c-index or statistic – a measure that summarises the area under the receiver operating characteristic curves (AUC) for binary outcomes (Harrell et al., 1996, Royston and Altman, 2010). For $R = 1000$ iterations, the c-index scores were collected from each regression model. Then, the median c-index estimates for each modelling approach were presented for comparison. Furthermore, the 2.5th and 97.5th percentile distribution of the c-index scores collected from 1000 simulations represented the 95% CI regions. For interpretations, the median c-index scores were judged on their values on the range between 0.5 and 1.0 (Royston and Altman, 2010). For example, a value of 0.5 indicates no predictive ability (or poor discrimination) whilst 1.0 suggests

perfect separation of patients with different outcomes. Generally, larger values above 0.7 indicate the model's ability to discriminate (Akobeng, 2007).

The disadvantage of the c-index measure is that it only focuses on the predictive accuracy of models under study, its estimate cannot tell us whether a model is worth using or not (Vickers and Elkin, 2006). In a situation where different prognostic models are being compared (as in this chapter), the c-index is unable to provide and guide on preferred models. To address this problem, the other measures including decision analysis curves (see section 5.3.1.4) were considered alongside the c-indexes to evaluate the clinical usefulness of predictive models in the simulation.

5.3.1.3 Calibration

Calibration is another useful and very popular measure of performance often used to evaluate prognostic models. It measures how close the predicted probabilities are to the observed rates of the event outcome (Giancristofaro and Salmaso, 2007, Royston and Altman, 2010). In the simulation, the predicted prognostic models obtained using the FP, RCS, categorisation and linearisation approaches were assessed by overlaying plots of observed probabilities against predicted probabilities (median) along the 45° diagonal line. Well-calibrated plots lie entirely on the 45° line (Altman et al., 2009, Collins et al., 2016). The challenge with this approach includes interpretation of deviations from the 45° line of identity which to some extent can be subjective (Austin and Steyerberg, 2014). Nonetheless, this method was still used to evaluate the performances of prognostic models used in the simulation. The alternative methods including the Hosmer-Lemeshow test are available. However, the test is directly influenced by sample sizes and discouraged for use in large samples (Paul et al., 2013). For example, the Hosmer-Lemeshow test will fail to identify small departures from the predicted models when the sample size is large. This is because the power of the test

increases with the sample sizes thus producing significant values. In contrast, graphical methods do not suffer from sample size limitation as the test-based approaches.

5.3.1.4 Clinical utility of predicted prognostic models

To decide whether the predicted prognostic models attained with the methods of categorisation, linearisation FP, and RCS are clinically useful or not, the decision curves were drawn and assessed to evaluate competing models (Vickers and Elkin, 2006). The decision curves took into consideration the clinical consequences of competing prognostic models. The net benefits attained in various prognostic models at different threshold probabilities were shown graphically to inform choices on the most appropriate model (Hunink et al., 2014). An illustrative example on how the model's net benefits and threshold probabilities could inform clinicians and patients to make the clinical decision for treatment is provided below:

Suppose an alcoholic patient was faced with a decision to undergo a hypertension treatment. If the decision was informed by the prognostic model - suggesting the probability of having the disease to be close to 1, the patient will request to be treated. If the probability of having hypertension was close to 0 in the prognostic model, the patient would unlikely opt for hypertension treatment. However, at values between 0 and 1, the patient would be unsure of whether to ask for treatment or not. In this case, the threshold probability (p_t) would be required to inform the clinicians or patient's decision. Given this scenario, the threshold probability (p_t) can be defined as where the expected benefit of hypertension treatment is equivalent to the expected benefit of avoiding the treatment.

Assuming the decision tree in Figure 5.1, the expression for threshold probability (p_t) could be written as follows:

$$p_t A + (1 - p_t) B = p_t C + (1 - p_t) D \quad \text{Eq. 5. 5}$$

Such that

$$p_t A - p_t C = (1 - p_t)D - (1 - p_t)B \Rightarrow \frac{A-C}{D-B} = \frac{1-p_t}{p_t} \quad \text{Eq. 5. 6}$$

From Eq. 5.6, $D - B$ quantity is the consequence of receiving hypertension treatment when it was not needed. The harm occurs due to false-positive results (see Figure 5.1).

The quantity associated with the consequence of avoiding hypertension treatment when it could have been beneficial is given as $A - C$. In $A - C$ quantity, comparing the true positive and false negative results, the harm is from the latter (see Figure 5.1). Clearly, from Eq. 5.6, the threshold probability at which a patient decides on the treatment depends on how one weighs the relative harm of false-positive and false negative results.

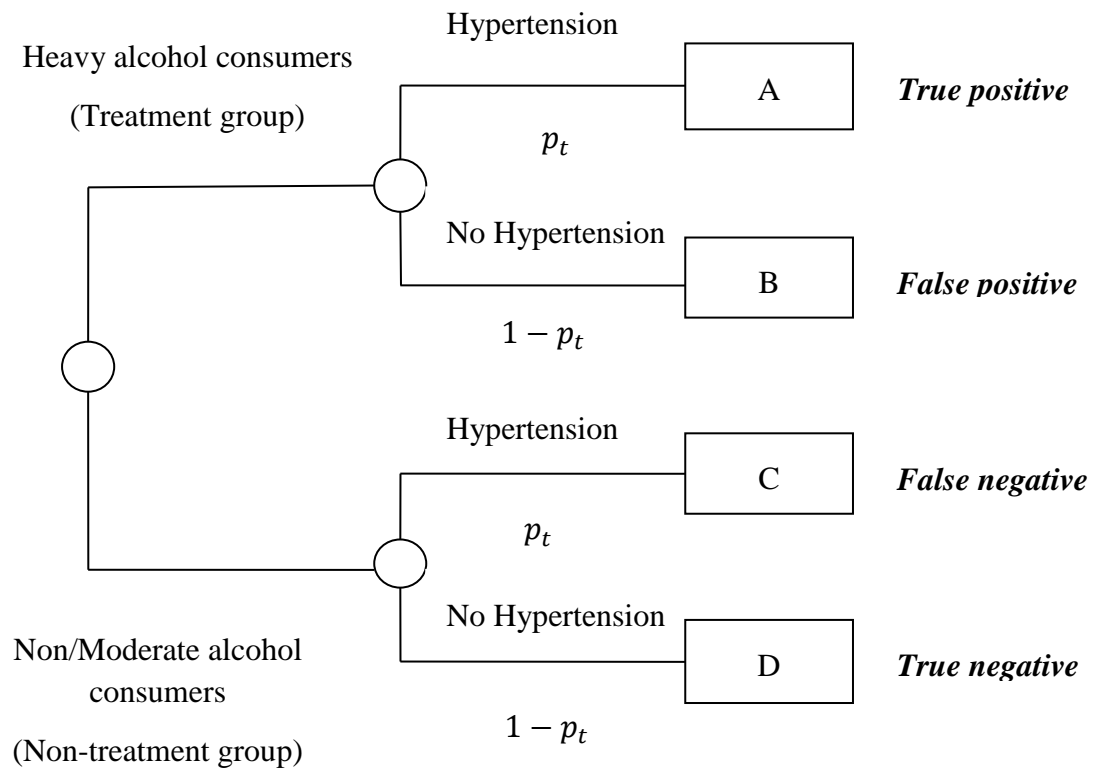


Figure 5.1: A decision tree diagram. The probabilities of disease and no disease are given by p and $1-p$ respectively. The values of true positive, false positive, false negative and true negative are given by A , B , C , and D respectively.

Based on the concept above, Vickers & Elkin (2006) suggested a method which allows for varying thresholds depending on the uncertainties associated with each outcome and individuals preferences. The proposed methods allow computation of the clinical net benefit in predicted prognostic models as follows:

$$NetBenefit = \frac{TruePositives}{n} - \frac{FalsePositives}{n} \left(\frac{p_t}{1-p_t} \right) \quad Eq. 5.7$$

where n is the number of observation (or sample size) considered in the simulations and p_t is the threshold defining risk given in the probability scale to weight the cost of false positive to false negative (from Eq.5.6).

In the results section, decision curves showing the net benefit (on the y-axis) against a range of selected p_t values (on the x-axis) were presented to compare various prognostic models in different linear and nonlinear datasets assumed in the simulations. In these curves, a useful prognostic model was the one that achieves the highest net benefits curve across the range of selected p_t .

There are disadvantages of working with the decision analysis curves. First, it is important to recognise that the computations of the net benefits require some weights on how individuals perceive the harms and benefits of a particular treatment. This information is not always readily available; thus defining the threshold (p_t) is difficult. The difficulties may occur at the population level where insufficient data on the harms and benefits is likely. Furthermore, the weights may be different between patients necessitating individual thresholds (Steyerberg et al., 2010). Hence, this is the reason why a range of thresholds (p_t) for the occurrence of the outcome are considered in the simulations.

5.4 Results

The results summarising and comparing various association shapes of the predictor in logit models fitted with fractional polynomials, restricted cubic splines, linearisation and categorisation (CAT3 & CAT5) approaches are presented in section 5.4.1. The results on discrimination and calibration are presented in section 5.4.2 and 0 respectively. The summary plots showing the clinical utility or net benefit results for applying the categorisation, and linearisation, FP, RCS models in the simulation are presented in section 5.4.4.

5.4.1 Comparison of various modelling techniques based on different association shapes of the predictor

Figure 5.2 compares FP, RCS, CAT3, CAT5 and linearisation methods assuming various association shapes between continuous predictor variables and the probabilities of an event outcome. The results were obtained through a simulation study comparing these methods in linear, thresholds and quadratic or U association shaped datasets. The simulations were replicated 1000 times and the median predicted functions were reported for comparison with the true shapes considering alcohol-hypertension relation as an example scenario.

Under the linear association datasets, the linear and restricted cubic spline approaches produced almost similar linear fits between the predictor and probabilities of the event outcome. The median predicted functions produced by fitting the linear and RCS models lied entirely on the ‘true’ fit when the predictor was equivalent to or greater than 20 units. For lower predictors (values below 20 units), the linear and RCS functions slightly deviated away from the ‘true’ fit - underestimating the ‘true’ probabilities of the event outcome (see Figure 5.2- top left). In contrast, fitting the fractional polynomial models produced the fit that lied on the ‘true’ function when the predictor was equivalent to or greater than 14 units. Compared to the linear and RCS models, the FP function greatly underestimated the outcome probabilities when the predictor was less than 14 units. The fitted FP function also had an artefact or ‘spike’ at the lower tail of predictor. The artefact was observed when the predictor was zero – an indication that zero values in the predictor data affect the behaviour of the FP models. Apart from the three models that keep continuous predictor values continuous in the analysis, the CAT3 and CAT5 methods produced step functions that suggest increasing probabilities of the event outcomes in the data (see Figure 5.2 - top left).

In linear threshold datasets, none of the five methods including CAT3, CAT5, linear, FP and RCS models were completely able to identify the true curve in the data. The FP and RCS partially identified the 'true' association at the upper range (> 28 units) of the predictor - where the actual relationship was linear (see Figure 5.2 – top right). Apart from that, the FP and RCS models underestimated the probabilities of the outcome at the lower predictor (< 10 units) and also overestimated the probabilities of the outcome when the predictor was between 10 - 28 units. Based on the latter, the two functions overestimate the probability of the outcome at the threshold (occurring at 20 units of the predictor). In Figure 5.2 (top right), the main dissimilarity between the FP and RCS models was an artefact observed in the FP fit when the predictor was zero. Other than that, the CAT3, CAT5 and linearisation methods revealed incorrect relationships in the linear threshold data. The CAT3 and CAT5 methods overlooked the actual curve features in the data by producing increasing step functions that assumes constant changes at different levels of the exposure (see Figure 5.2 (top right)). In contrast, the method of linearisation retained a linear function suggesting an increasing relationship between the predictor and the outcome. Overall, the median probability function attained by assuming linearity on the data underestimated the occurrence of the outcome at the lower and upper tails of the predictor distribution. Moreover, the method of linearisation overestimated the occurrence of the outcome at the threshold (occurring at 20 units of the predictor). A graphical comparison of CAT3, CAT5, linear, FP and RCS methods in the linear threshold dataset is provided in Figure 5.2 (top right).

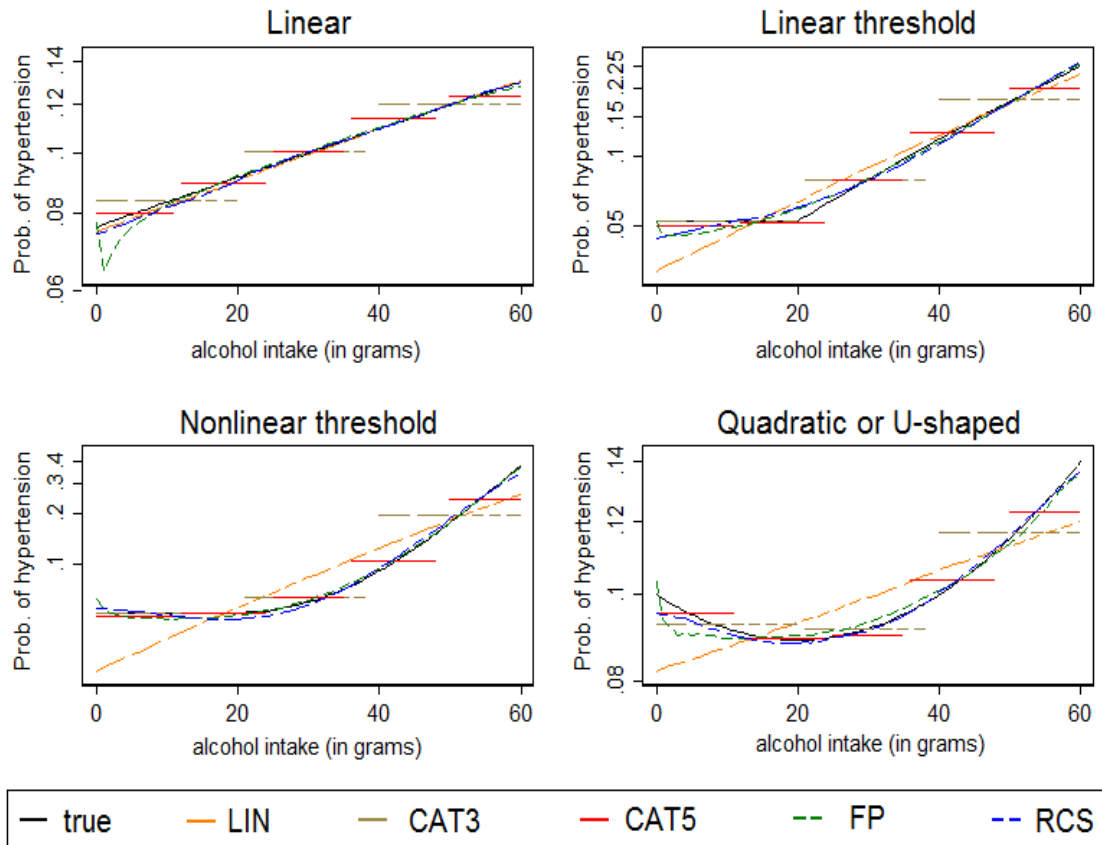


Figure 5.2: Comparison of FP (green), RCS (blue), CAT3 (brown), CAT5 (red) and linearisation (orange) methods in a simulation where continuous predictor variable assume various shapes for prediction of event outcome. Median probability functions obtained with these methods after 1000 simulations are presented to compare them against true shapes (black) in linear, thresholds and quadratic datasets.

Under the nonlinear threshold datasets, the methods of FP and RCS produced approximately near similar association functions that were close to the true relationship. However, none of the two approaches produced a fit that entirely lends itself on the true curve (see Figure 5.2 – bottom left). The two methods of categorisation produced increasing step functions that do not identify with true curve in the data. The linearisation approach produced a linear association function - showing an underestimation of the outcome probabilities at the predictor below 18 units and above

50 units. At the predictor range between 18 - 50 units, the probabilities of the outcome event would be overestimated when fitting the linear function (see Figure 5.2 – bottom left).

For quadratic or U association datasets (Figure 5.2 – bottom right), the FP and RCS revealed the existence of U associations in the data however the two models struggled to lend themselves on the true curve at the lower tail of the predictor. Fitting FPs produced logistic functions that underestimated the ‘true’ probabilities of the event outcome at the lower tail (≤ 15 units) and upper tail (≥ 44 units) of the predictor. When the predictor values were between 15 – 44 units, the FP function overestimated the ‘true’ probabilities of the event outcome. In addition, the FP function was characterised by some artefacts around the zero predictor values (see Figure 5.2 – bottom right). In contrast, fitting the RCS function underestimated the ‘true’ probabilities of the event outcome when the predictor was ≤ 38 units and ≥ 50 units. When the predictor was between 38 – 50 units, the RCS was overestimating the ‘true’ probabilities of the event outcome. Apart from the FP and RCS methods, categorisation and linearisation produced inadequate fits in the data - showing step functions and linear relationships respectively. See a graphical comparison of these methods in Figure 5.2 – (bottom right).

5.4.1.1 Confidence intervals of predicted functions

Figure 5.3 & Figure 5.4 shows the 95% CI regions corresponding to the median predicted functions in Figure 5.2.

The results under the linear association datasets (Figure 5.3 - top) show that the FP function had the widest CIs at the tails of the predictor compared to the linear, CAT3, CAT5 and RCS models. This behaviour was attributed to instability of the FP function at the tails of the predictor (see Figure 5.2 - top right). Apart from that, the

linear model produced a fit with the thinnest CIs across the range of the predictor distribution whilst the CAT3 and CAT5 methods retained step function CIs (see Figure 5.3 - top).

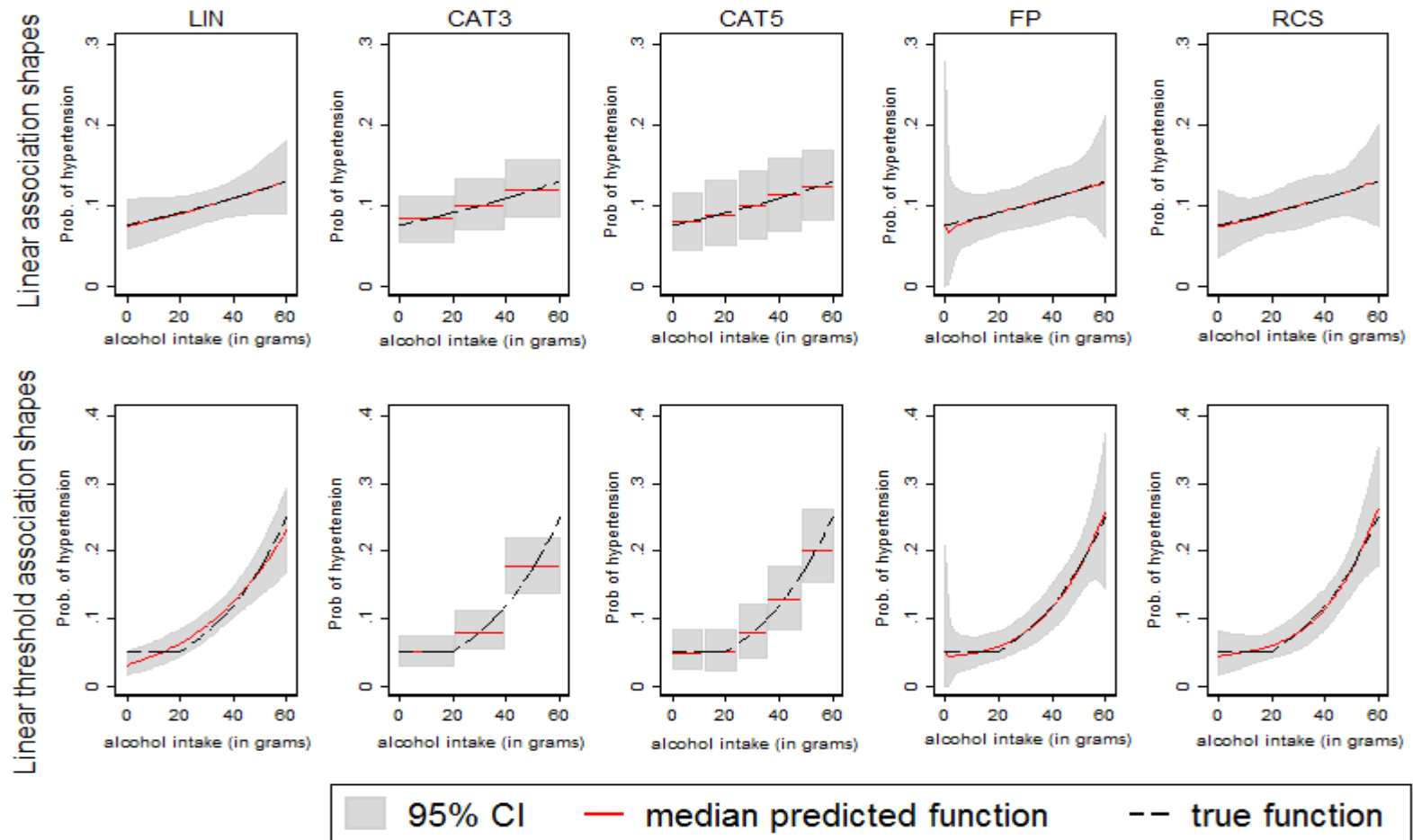


Figure 5.3: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations (replicates) after fitting linear and linear threshold association datasets using linearisation, categorisation, FP and RCS modelling approaches

Under the linear threshold datasets (Figure 5.3 - bottom), the CI regions produced using the CAT3 model failed to sufficiently cover the true function. In contrast, the CAT5, linear, FP and RCS models produced CIs that offer sufficient coverage region on the true function. However, the FP function had extremely wider CIs at the tails of the predictor distribution compared to the CAT5, linear and RCS fits. The FP function was unstable at the tails, thus the extremely wider CIs. Beside the FP function, the RCS model also had wider CIs at the upper tail of the predictor compared to the CAT5 and linear fits (see Figure 5.3 – bottom).

Under the nonlinear threshold datasets, the CAT3, CAT5 and linearisation methods produced inadequate fits with insufficient CI regions for the true functions. The linearisation and CAT3 methods struggled to adequately estimate and provide coverage of the true probability of an event outcome at the central distribution of the predictor. For example, at 40 units of the predictor, true probability of an event outcome was 0.09. However, the linear and CAT3 models overestimated the probabilities of the event outcome at 0.12 (CI = 0.10, 0.15) and 0.19 (CI = 0.15, 0.24) respectively. In contrast, the CAT5 struggled at the upper distribution of the predictor – underestimating the true event probabilities (see Figure 5.4 - top). Apart from the latter, the alternative methods of FP and RCS produced CIs that were adequate for the true functions (see Figure 5.4 - top). Like in the previous relationship functions, the FP model retained wider CIs at the lower and upper tails than the other methods. Detailed graphical comparison of median predicted functions and their CIs in the nonlinear threshold datasets are provided below in Figure 5.4 (top).

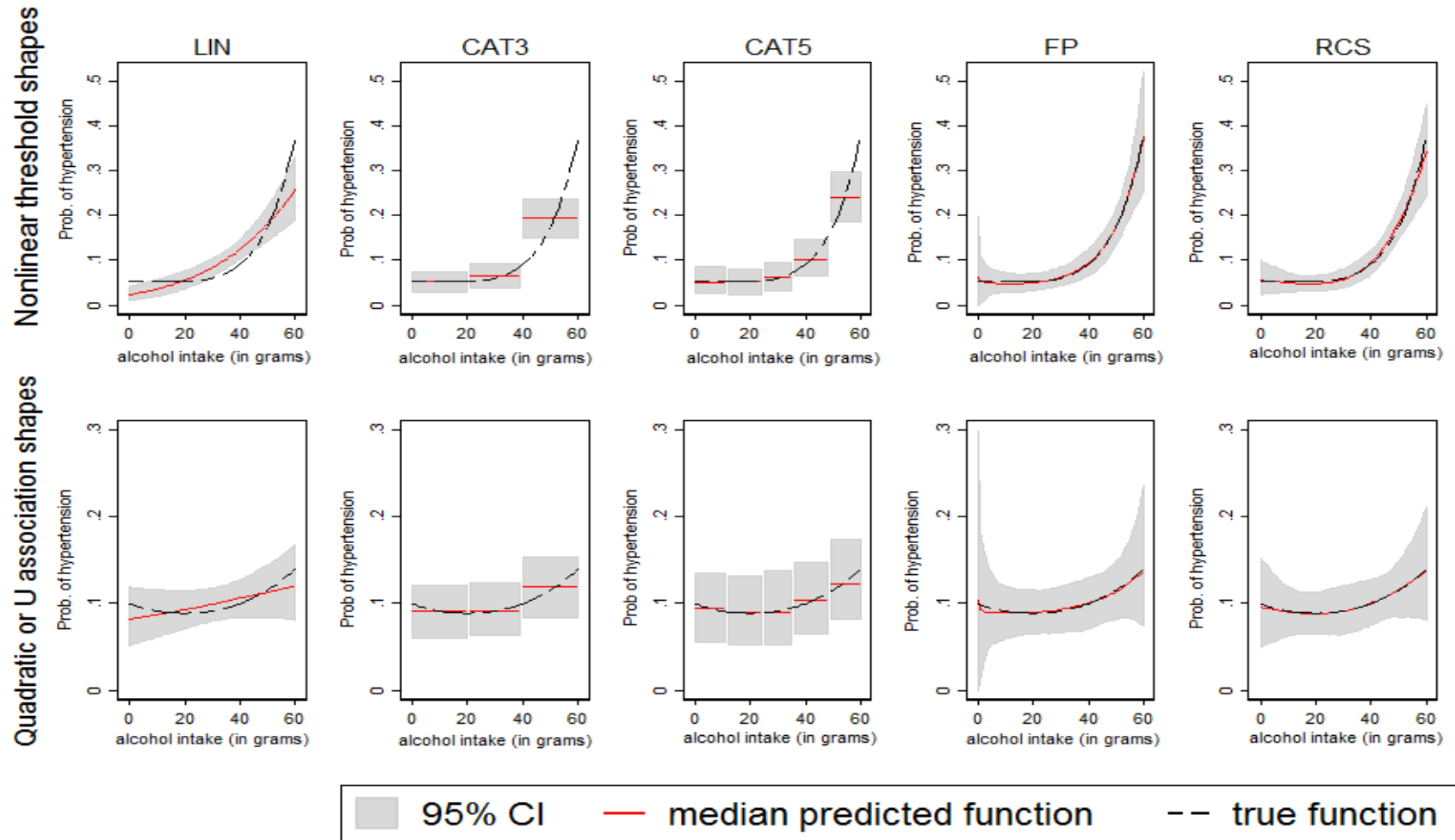


Figure 5.4: The median predicted functions and their 95% confidence interval regions obtained from 1000 simulations after fitting nonlinear threshold and quadratic or U association datasets using linearisation, categorisation, FP, and RCS modelling approaches

The 95% CIs attained using the quadratic or U association datasets are found in Figure 5.4 (bottom). Although the linear, CAT3 and CAT5 models do not adequately lend themselves on the ‘true’ quadratic function, they retained sufficient CIs on the data. Unsurprising, the FP and RCS models retained near approximation fits with adequate coverage for the actual fit (see Figure 5.4 – bottom). However, the 95% CI regions in the FP and RCS models were narrow at the centre of the predictor distribution and wide at the tails. In comparison, the FP function predicted wider CIs at the tails than the RCS models. For example, when the predictor was zero, the probabilities of the event outcome estimated from the FP and RCS functions were 0.10 (CI = 0.00, 0.30) and 0.10 (CI = 0.05, 0.15) respectively. At the upper tail, when the predictor was 59 units, the estimated probabilities of the event outcome was 0.13 (CI = 0.08, 0.22) in the FP model and 0.13 (0.08, 0.20) in the RCS function. These estimates were in comparison with the ‘true’ of probabilities of 0.10 and 0.14 occurring at 0 and 50 units of the predictor.

5.4.1.2 Estimated turning points or optimum probabilities

Table 5.2 below present the turning points or optimal probabilities of the event outcome from the threshold and quadratic datasets using the CAT3, CAT5, RCS and FP models. This section also summarises the results on coverage probabilities from the simulations. The coverage probabilities of ‘true’ optimal rate of an event outcome in 1000 simulation are summarised in Table 5.3 for comparison.

Table 5.2: Comparison of optimal predictor and probability estimates obtained across 1000 simulations after fitting thresholds and quadratic association datasets using different modelling approaches.

Association datasets under investigation	Modelling approaches	Estimates			The true optimal probability of an outcome at 20 units of the predictor
		Estimated optimal predictor	The estimated optimal probability of an outcome	The estimated probability of an outcome at 20 units of the predictor	
Linear threshold datasets	CAT3	-	0.05 (0.03, 0.07)	0.05 (0.05, 0.08)	0.05
	CAT5	-	0.04 (0.02, 0.07)	0.05 (0.02, 0.08)	
	RCS	0 (0, 22)	0.04 (0.02, 0.07)	0.06 (0.04, 0.08)	
	FP	0 (0, 18.5)	0.04 (0.00, 0.06)	0.06 (0.04, 0.08)	
Nonlinear threshold datasets	CAT3	-	0.05 (0.03, 0.07)	0.05 (0.03, 0.08)	0.05
	CAT5	-	0.04 (0.02, 0.06)	0.05 (0.02, 0.08)	
	RCS	18 (0, 25)	0.04 (0.02, 0.06)	0.05 (0.03, 0.07)	
	FP	9 (0, 27)	0.04 (0.00, 0.06)	0.05 (0.03, 0.07)	
Quadratic association datasets	CAT3	-	0.08 (0.06, 0.11)	0.09 (0.06, 0.12)	0.09
	CAT5	-	0.07 (0.05, 0.10)	0.09 (0.05, 0.13)	
	RCS	22 (0, 60)	0.08 (0.05, 0.10)	0.09 (0.06, 0.11)	
	FP	9 (0, 60)	0.07 (0.00, 0.10)	0.09 (0.07, 0.12)	

Under the linear threshold datasets, the RCS and FP models predicted the optimal predictor values at 0 (CI = 0, 22) and 0 (CI = 0, 18.5) units respectively. At the estimated optimal predictor of 0 units, the corresponding probabilities of the event outcome were 0.04 (CI = 0.02, 0.07) and 0.04 (0.00, 0.06) in the RCS and FP models respectively. At 20 units of the predictor where the ‘true’ probability of 0.05 for the event outcome was expected, both the RCS and FP models overestimated the event rates – suggesting the probability of 0.06 (CI = 0.04, 0.08) (see Table 5.3). These results suggest that fitting the RCS and FP logistic regression models in linear threshold datasets shift the ‘true’ position of the predictor producing an underestimation. As seen in Table 5.2, the location of optimal predictor was shifted to the left – producing smaller probabilities of the event outcome than at the true turning point. In 1000 replications, the coverage probabilities of the ‘true’ optimal outcome recorded when fitting the RCS

and FP models in linear threshold datasets were 1.000 and 0.721 respectively (see Table 5.3). The coverage probability attained using the FP function was far from the 95% nominal level than when applying the RCS method. These results implied some under-coverage probability in the FP models and conservative coverage in the RCS models.

Table 5.3: The coverage probabilities of 'true' optimal outcome events in 1000 simulations (replications) obtained in thresholds and quadratic datasets after fitting categorisation (CAT3 and CAT5), RCS, and FPs.

Types of association dataset under investigation	Modelling Approaches			
	CAT3	CAT5	RCS	FP
Linear threshold datasets	0.999	1.00	1.000	0.721
Nonlinear threshold datasets	1.000	1.00	1.000	0.755
Quadratic association datasets	0.907	0.957	0.867	0.658

Under the nonlinear threshold datasets, the RCS model predicted the optimal predictor at 18 (CI = 0, 25) units with the corresponding probability of 0.04 (CI = 0.02, 0.06) for the event outcome. In contrast, the optimal predictor attained by fitting the FP model was 9 (CI = 0, 27) units and the corresponding probability of the event outcome 0.04 (CI = 0.00, 0.06). When compared to the 'true' optimal probability of 0.05 attained at 20 units of the predictor, both the RCS and RCS resulted in underestimated probabilities of the event outcome and the optimal predictor (see Table 5.2). Apart from the latter, the FPs returned the lowest coverage rates of the 'true' optimum outcome than when using RCS models. In 1000 simulation using nonlinear threshold datasets, the coverage probability using FP models was 0.755 whilst in RCS was at 1.000. These

results suggest under-coverage probability in FP models and conservative coverage in the RCS models.

Under the quadratic datasets, the optimal predictor was overestimated when fitting RCS models. The RCS predicted the optimal predictor at 22 (CI = 0, 60) units with the corresponding probability of 0.08 (CI = 0.05, 0.10) for the event outcome. Contrary to the RCS method, fitting the FP models resulted in an underestimation of the 'true' optimal predictor. The optimal predictor under the FP method was estimated at 9 (CI = 0, 60) units. The corresponding optimal probability of the event outcome was also underestimated when fitting the FP function (see Table 5.2). Although there were contradictions in the predicted optimal/turning points from these models (RCS and FP), they both accurately estimated the 'true' probability of the outcome at 20 units of the predictor (see Table 5.2). In Table 5.3, when comparing the coverage probability, fitting the RCS functions retained greater proportion than the FP models (0.867 vs 0.654). However, under-coverage probabilities were evident in the two methods when fitted in quadratic or U association datasets. The estimated proportions that the interval contains the 'true' optimal outcome was less than the nominal 95% level assumed in the simulation for both models.

Although the categorisation produced inadequate fits (step functions) in the simulation, the CAT3 and CAT5 had greater coverage probabilities of the event outcome in thresholds and quadratic datasets (see Table 5.2). However, this was not surprising. The 'true' optimal points were placed in the lower category (under the CAT3) and second category (under the CAT5) hence these categories were always returned every time the minimum probabilities were recalled in the simulations. Therefore, this retained the coverage probabilities greater or closer to the nominal rate of 95% under the two methods (see Table 5.3).

5.4.2 Discrimination

Table 5.4 compares the differences in AUC of five models (linear, CAT3, CAT5, FP and RCS) in the simulation. The median estimates of the AUC and their 95% confidence intervals were reported from the four logit functions (including log linear, thresholds and quadratic datasets) for comparison.

There was discrimination failure when applying the methods of linearisation, CAT3, CAT5, RCS, and FP in log linear datasets. In Table 5.4, the logistic regression models (attained through the methods of linearisation, CAT3, CAT5, RCS, and FP) produced the c-index scores that were closer to 0.5 – suggesting the inability of the models to discriminate between outcomes with the event and those without. The logistic regression models (attained by linearising and categorising the continuous predictor into three groups) retained similar c-index scores - with little differences on their confidence intervals. For example, the c-index scores under the methods of linearisation and CAT3 were recorded as 0.55 (CI=0.50, 0.61) and 0.55 (CI=0.50, 0.60) respectively. In contrast, there was slight improvement on the c-index scores when fitting the FP, CAT5 and RCS models. The FP, CAT5 and RCS models yielded c-index scores of 0.56 (CI=0.50, 0.61), 0.57 (CI=0.53, 0.62) and 0.58 (CI=0.53, 0.71) respectively (see Table 5.4). Although the difference was not substantial, the latter results also suggested the CAT5 performs better than the FP approach.

Table 5.4: The median estimates for the area under the ROC curve (AUC) and their 95% confidence intervals obtained after fitting FP, RCS, categorisation, and linearisation models in a simulation study replicated 1000 times. The reported estimates were obtained after applying these methods in log linear, thresholds and quadratic datasets.

Types of association datasets under investigation	Methods of analysis				
	Linearisation	Categorisation (CAT3)	Categorisation (CAT5)	Restricted cubic spline	Fractional polynomial
Linear	0.55 (0.50, 0.61)	0.55 (0.51, 0.60)	0.57 (0.53, 0.62)	0.58 (0.53, 0.71)	0.56 (0.50, 0.61)
Linear threshold	0.69 (0.61, 0.73)	0.65 (0.60, 0.70)	0.67 (0.61, 0.72)	0.68 (0.62, 0.78)	0.67 (0.61, 0.73)
Nonlinear threshold	0.70 (0.64, 0.76)	0.67 (0.62, 0.72)	0.70 (0.64, 0.75)	0.71 (0.65, 0.81)	0.70 (0.64, 0.76)
Quadratic or U shaped	0.53 (0.50, 0.59)	0.54 (0.51, 0.59)	0.56 (0.53, 0.61)	0.57 (0.53, 0.69)	0.55 (0.50, 0.61)

There was an improvement in c-index scores when fitting the log linear, CAT3, CAT5, FP and RCS regression models in linear threshold datasets. However, discrimination remained poorer across the five methods of analysis. The CAT3 method had the lowest c-index score of 0.65 (CI=0.60, 0.70) followed by the CAT5 and FP that achieved similar c-index scores of 0.67 (with slightly different CIs). The RCS and linearisation approaches had the largest but poor discrimination with c-index scores of 0.68 (CI=0.62, 0.78) and 0.69 (CI=0.61, 0.73) respectively (see Table 5.4).

Fair discrimination was obtained with CAT5, FP, RCS and log linear models in nonlinear threshold datasets. In contrast, the CAT3 approach retained the least c-index of 0.67 (CI=0.62, 0.72) – suggesting poor discrimination in the data (see Table 5.4 for more details).

The five methods of analyses (including the linear, CAT3, CAT5, RCS and FP) had no predictive discrimination ability when fitted in quadratic datasets. The c-index scores attained when fitting these methods in quadratic datasets were closer to 0.5 – suggesting discrimination failure (see Table 5.4).

Overall, the RCS methods produced greater c-index scores amongst the five methods considered in the simulations whilst the CAT3 retained the least c-index scores. However, the difference between the five methods of analysis was not substantially large/worse (see Table 5.4). The CAT5, FP and linearisation competed fairly but the CAT5 slightly outperformed the linear and FP models in some scenarios.

5.4.3 Calibration plots

Figure 5.5 & Figure 5.6 show calibration plots in 1000 simulations obtained after fitting various datasets using the logistic regression models through the linearization, CAT3, CAT5, FP and RCS approaches. The plots summarise the agreement between observed probabilities against predicted probabilities. When the

observed and predicted probabilities are in agreement, the estimated calibration curves should produce an ideal 45° line with an intercept of zero and slope of one. Otherwise, the model is not well calibrated.

The calibration plots obtained in Figure 5.5 (top) after fitting different logistic regressions using linear relationship datasets showed no perfect calibration when applying any of the linear, CAT3, CAT5, FPs and RCS models. The linearisation and FP models produced better calibrated plots in lower prediction regions below 20% and 30% respectively. In high prediction region both the linear and FP models showed disagreements between the observed and predicted probabilities. The predicted probabilities were lower than the observed in the high prediction region when fitting the linear model and slightly higher under the FP function (see Figure 5.5 - top). In contrast, fitting the RCS models in linear association datasets produced disagreements between observed and predicted probabilities in lower ($< 10\%$) and high ($>17\%$) prediction regions. In the range between 10% and 17%, the predicted probabilities agreed with those observed and the RCS plot was well calibrated in this region. In other words, the RCS calibration plot showed a combination of agreement in the prediction region between 10% to 17% and miscalibration in the lower ($<10\%$) and upper ($>17\%$) regions of the predicted probabilities (see Figure 5.4 – top). Apart from that, CAT3 and CAT5 methods produced plots showing systematically lower predicted probabilities than those observed. The estimated calibration plots obtained under the two methods of categorisation lied above the 45° line – showing positive intercept coefficients and parallel slopes. However, the calibration plot under the CAT5 looked better than that under the CAT3 since it was closer to the 45° line (although the difference between the plots produced by the two functions was not substantial) (see Figure 5.4 (top)).

In linear threshold datasets, the linear model produced a calibration curve with an intercept greater than zero and a negative slope < 1 for prediction probabilities above 25%. In Figure 5.5 (bottom), the linear model was well calibrated in the prediction region between 5% - 15% and miscalibrated in the prediction region below 5% and above 15% - suggesting a combination of agreement and disagreement. Generally, such calibration plots are difficult to interpret. However, overall sum differences between predicted and observed probabilities along the calibration curve show that the predicted probabilities are generally too high indicating the presence of miscalibration. In contrast, the two methods of categorisation (CAT3 & CAT5) produced calibration plots showing systematically lower predicted probabilities than those observed. However, there were little differences between observed and predicted probabilities across the prediction range in both models. Under the methods of FP and RCS, the plots were well calibrated in the prediction region below 30%. In the prediction region above 30%, the predicted probabilities were lower than those observed when fitting the FP models. In contrast, the predicted probabilities exceeded those observed when fitting the RCS models in this region ($>30\%$) (See Figure 5.5 – bottom).

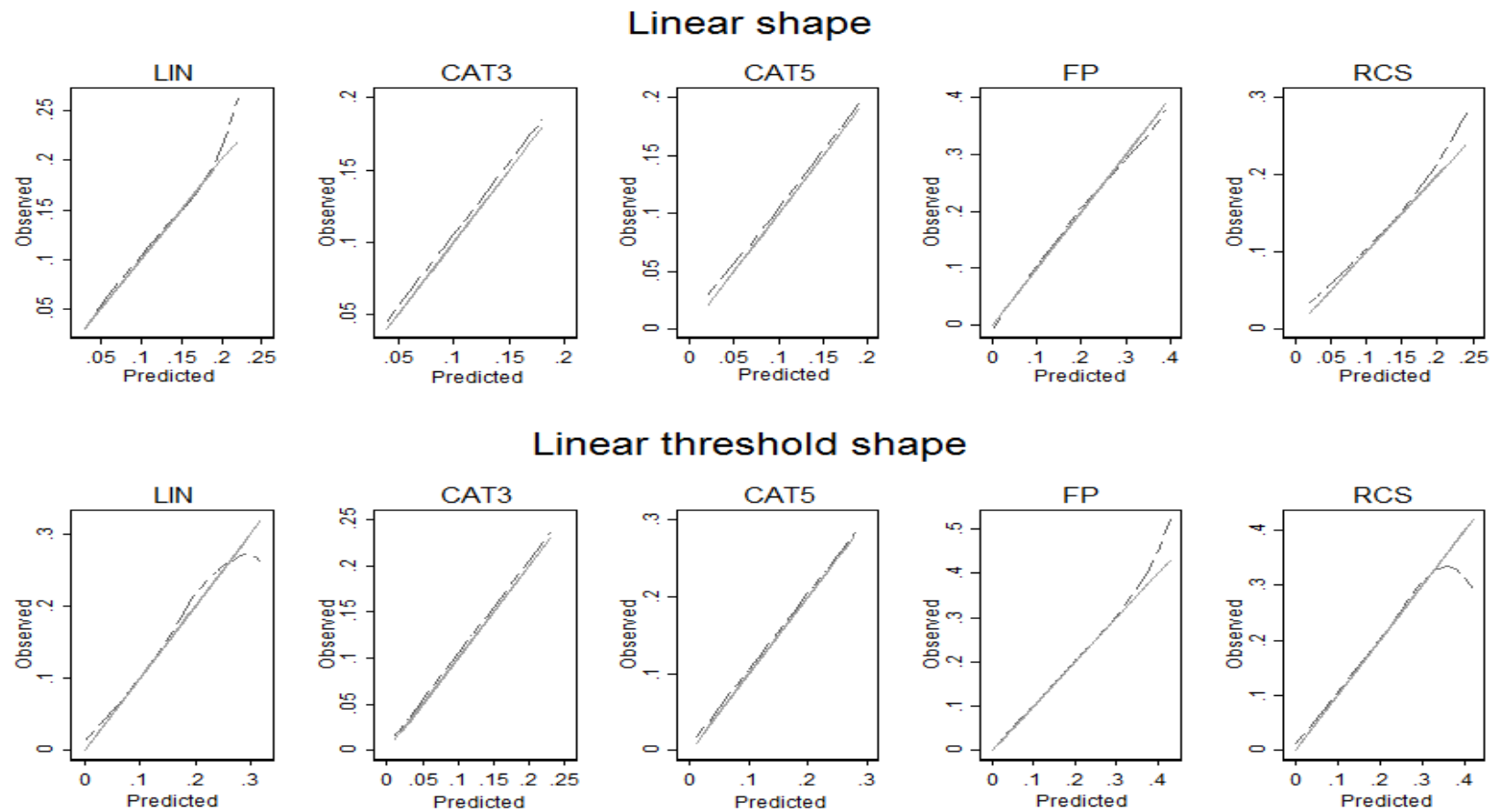


Figure 5.5: Calibration plots of the event probabilities obtained in log odds models. The plots were obtained in a simulation with 1000 replicates comparing linearisation, categorisation, FPs and RCS approaches in linear and linear threshold datasets respectively. For each approach, the median observed probabilities of an event were plotted against the predicted probabilities.

In nonlinear threshold datasets, better calibration plots occurred when fitting the CAT3, CAT5, FP and RCS models. However, there was small miscalibration in the RCS models at the extreme prediction region (>45%). In contrast, the CAT3, CAT5 and FP models produced calibration plots very close to the 45° line (across the prediction range). However, the CAT3 and CAT5 models still retained plots with slightly lower predicted probabilities than those observed in the true function (see Figure 5.6 - top). In the same datasets, the worst miscalibration plot occurred when applying the method of linearisation (see Figure 5.6 - top). The linear model retained a plot showing disagreement between the observed and predicted probabilities at extreme prediction regions <5% and > 15% (see Figure 5.6 - top).

In quadratic relationships, there was a combination of agreement and miscalibration when fitting the log linear, FP and RCS models in the datasets. The methods of linearisation, FP and RCS produced plots suggesting well calibration in lower prediction region and miscalibration in high prediction range (see Figure 5.6 – bottom). In contrast, fitting CAT3 and CAT5 models produced calibration curves with intercepts terms greater than zero and positive slopes - showing consistently lower predicted probabilities than the observed probabilities (see Figure 5.6 – bottom).

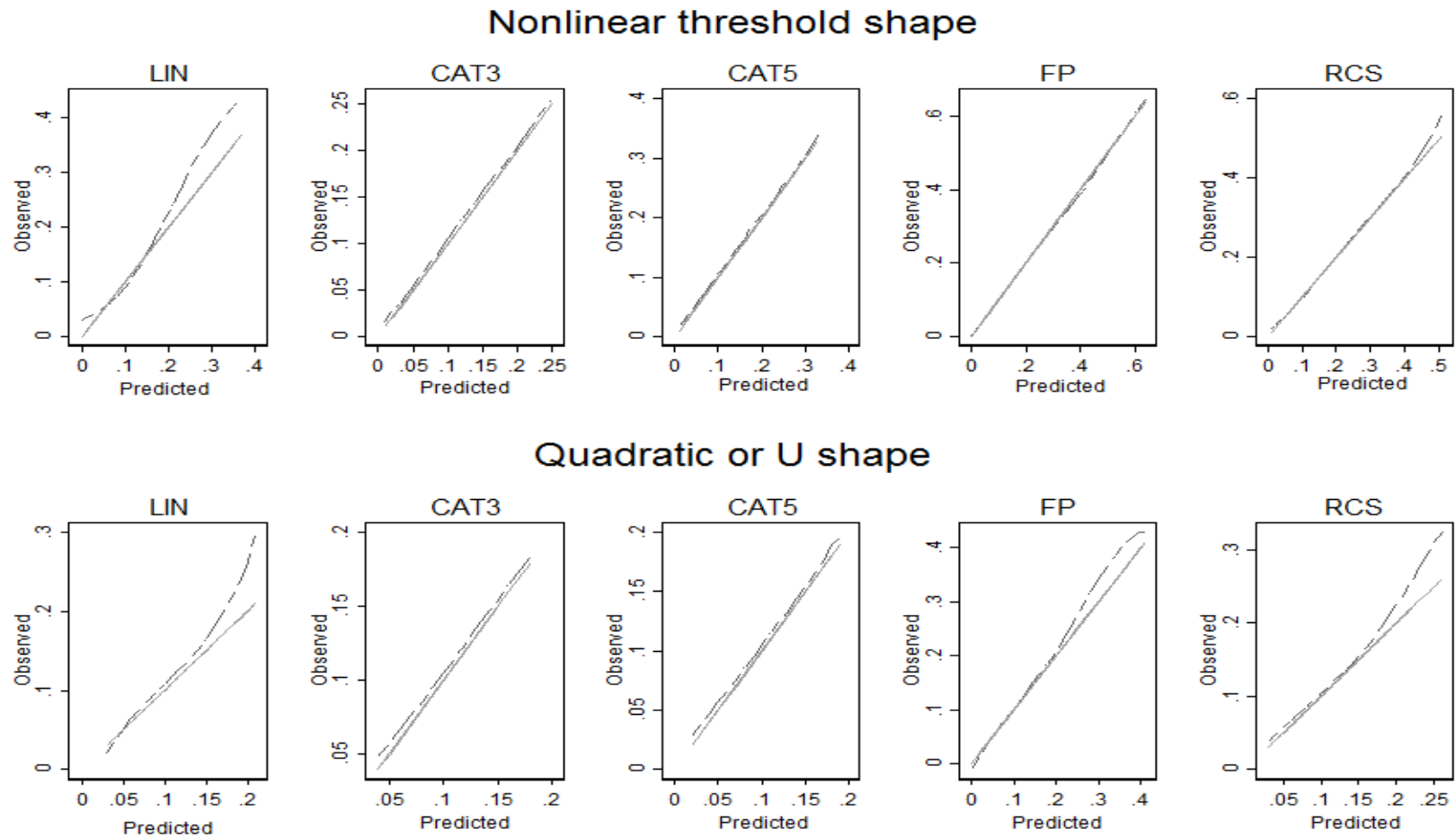


Figure 5.6: Calibration plots of the event probabilities obtained in log odds models. The plots were obtained in a simulation with 1000 replicates comparing linearisation, categorisation, FPs and RCS approaches in nonlinear thresholds and quadratic or U shaped datasets respectively. For each approach, the median observed probabilities of an event were plotted against predicted probabilities.

Overall, the results above suggest the two method of categorisation (CAT3 and CAT5) were likely to produce lower predicted probabilities than those observed in true functions assumed in this simulation study. However, the CAT3 and CAT5 retained better plots than the linear, FP and RCS models. The calibration plots attained by methods of linearisation, FP and RCS were characterised by combinations of agreement and disagreements (or miscalibrations) when applied in the same datasets.

5.4.4 Clinical usefulness of statistical models under investigation

The decision curves plotting the net benefits against varying threshold probabilities are presented in this section to evaluate the risk prediction models and their clinical utility. The plots identified the threshold probabilities at which the prediction models were of value, the magnitudes of net benefits and the overall optimal models in the four association datasets considered in the simulations.

In Figure 5.7, the strategy with large clinical benefits has the highest curve. The grey solid line assumes that all patients have the event outcome (hypertension) and are all treated. This means that any prediction model closer to the grey solid line (or with less net benefits) has negative clinical consequences. Furthermore, the black solid line assumes no patient has the disease and required for treatment. Given these scenarios, the prediction models attained by using the methods of linearisation, CAT3, CAT5, FP and RCS were similar to the strategy of treating all patients with the event outcome at the lower threshold probabilities, $p_t < 7\%$ than at higher thresholds. For $p_t < 7\%$, the five prediction models are inefficient and have no clinical use than the strategy of treating all patients with the event outcome. In Figure 5.7, the clinical usefulness of these prediction models was only realized when the thresholds probability, p_t was between 7% and 13%. When the threshold probabilities was between 7% and 13%, the net benefits derived from using the five methods were greater or better than the strategy of

not treating any patient (assuming no patients has the event outcome). In other words, when $p_t < 7\%$ or $p_t > 13\%$, the five prediction models have no clinical value since they are no better than a strategy of treating all the patients with the event outcome ($p_t < 7\%$) or not providing treatment for all patients without the event outcome ($p_t > 13\%$).

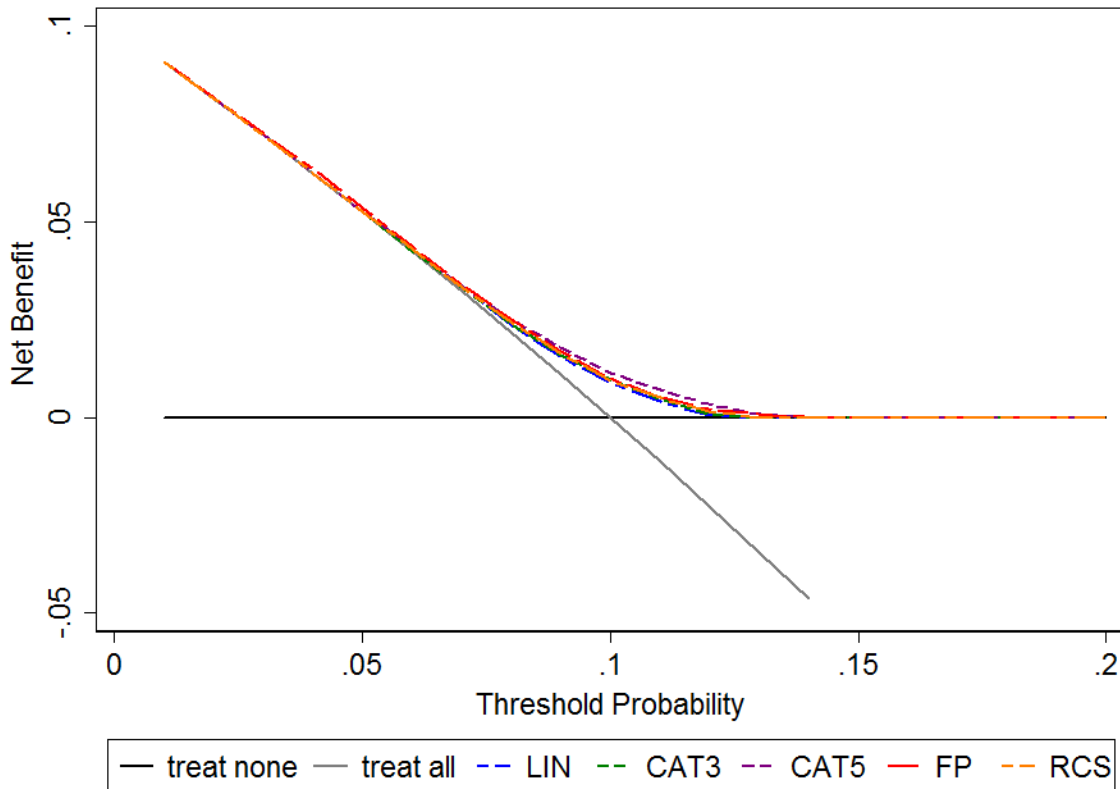


Figure 5.7: The median predicted curves attained from 1000 simulations showing the Net Benefits of applying various statistical models (FP, RCS, CAT3, CAT5 and linearisation approaches) in linear predictor-outcome relationship datasets

The advantages of applying the methods of linearisation, CAT3, CAT5, FP and RCS in the linear association datasets can be summarised by assessing the reduction of false positive results from each prediction model. This is because whenever treatment is guided by prediction models, the harm often occurs due to false-positives. Table 5.5 and Figure 5.8 compares the net benefit and reduction of false positive results per 100 patients at various threshold probabilities as obtained in the predicted models.

According to the results in Table 5.5, the net benefit of 0.05 (0.03, 0.07) was attained at p_t of 5% in all the prediction models. This means that compared to none treatment of all patients, the prediction models suggest 5 true positives per 100 patients without any increase in the numbers of false positives. At p_t of 8%, the FP and CAT5 models predicts 3 true positives per 100 patients compared to 2 obtained when fitting the CAT3, log linear and RCS models (not shown in the table). This agrees with what was observed in Figure 5.7 – the CAT5 had higher net benefits followed by the FP model when $7\% < p_t < 13\%$. This means that at the threshold probabilities (p_t) between 7% and 13% patients or clinicians may choose the results from the CAT5 model as their treatment strategy over the FP, CAT3, log linear and RCS results since it achieves higher clinical benefits. However, the results from the FP, CAT3, log linear and RCS methods were not substantially worse/different from those attained using the CAT5 (see Figure 5.7).

The results on false-positive reductions per 100 patients also suggested some agreement between the methods of CAT3, CAT5, log linear, FP and RCS when the threshold probability was $p_t < 7\%$ or $p_t > 13\%$ (see Figure 5.8). At the threshold probabilities, p_t between 7% and 13% there was disagreement on these methods. For example, at p_t of 10% the CAT5 model produced 10 fewer false positives per 100 patients compared to 9 (CAT3, FP and RCS) and 8 under the linearisation approach. This means that applying the CAT5 a treatment strategy will reduce the enrolment amongst patients without the event outcome by 10% compared to the 9% (under the CAT3, FP, RCS) and 8% under the linear method - assuming the numbers of true positives do not increase (see Table 5.5 and Figure 5.8 below for more details).

Table 5.5: Comparison of net benefits and reduction of false positive results per 100 patients according to different statistical prediction models assuming various threshold probabilities.

Threshold probabilities (%)	Treat all	Categorisation (3 groups)			Categorisation (5 groups)			Linearisation		
		Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients
0.05	0.05	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 0)	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 0)	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 0)
0.10	0.00	0.01 (0.00, 0.03)	0.01 (0.00, 0.02)	9 (0, 22)	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)	10 (1, 23)	0.01 (0.00, 0.03)	0.01 (0.00, 0.02)	8 (0, 22)
0.15	-0.06	0.00	0.06 (0.04, 0.08)	33 (22, 46)	0.00	0.06 (0.04, 0.08)	33 (22, 46)	0.00	0.06 (0.04, 0.08)	33 (21, 46)
0.20	-0.13	0.00	0.13 (0.10, 0.15)	50 (41, 60)	0.00	0.13 (0.10, 0.15)	50 (41, 60)	0.00	0.13 (0.10, 0.15)	50 (41, 60)

Fractional Polynomials			Restricted cubic splines		
Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients
0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 5)	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 4)
0.01 (0.00, 0.03)	0.01 (0.00, 0.03)	9 (0, 23)	0.01 (0.00, 0.03)	0.01 (0.00, 0.03)	9 (0, 23)
0.00	0.06 (0.04, 0.08)	33 (21, 46)	0.00	0.06 (0.04, 0.08)	33 (21, 46)
0.00	0.13 (0.10, 0.15)	50 (41, 60)	0.00	0.13 (0.10, 0.15)	50 (41, 60)

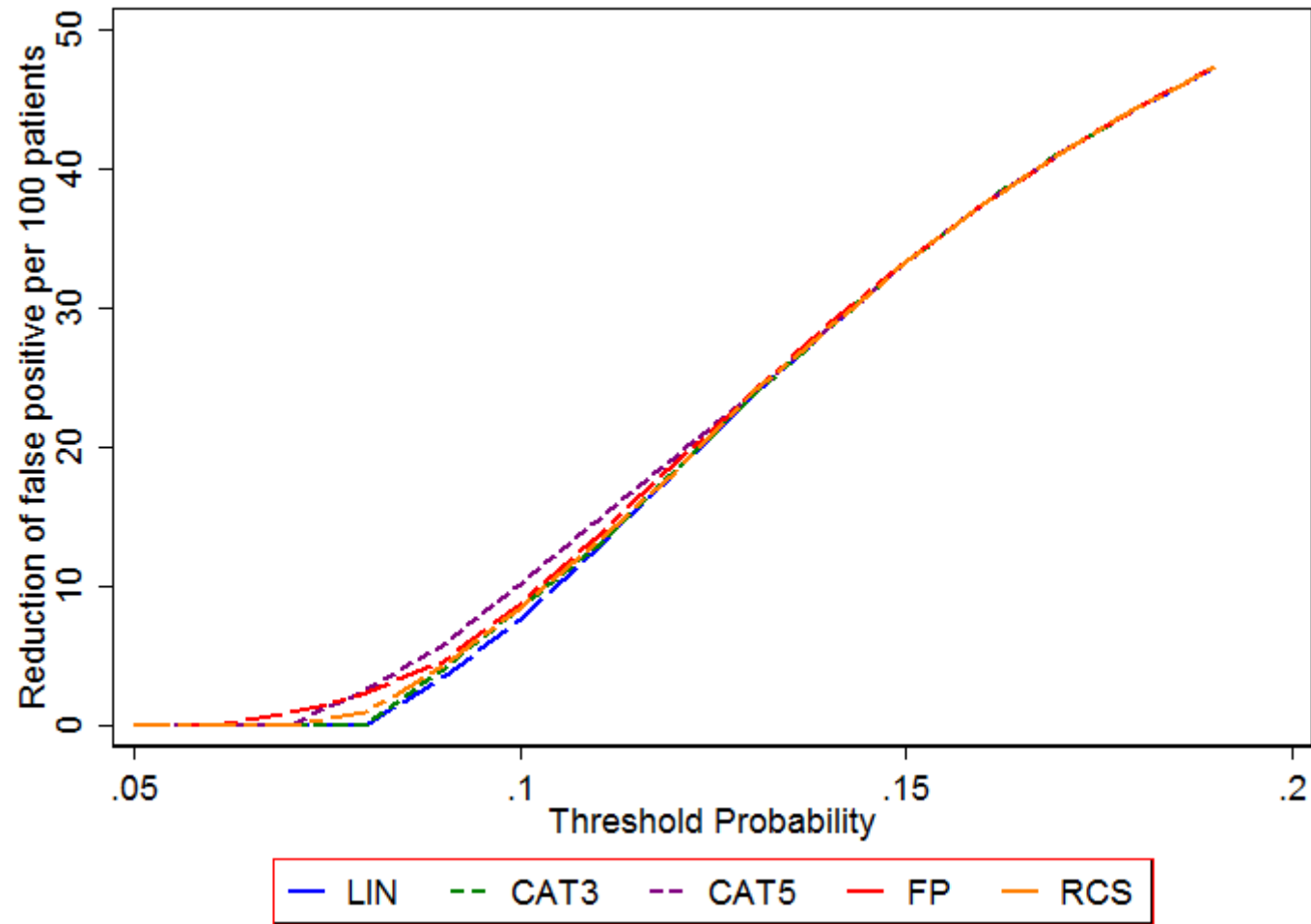


Figure 5.8: Comparison of various statistical prediction models showing the net reduction of false positives per 100 patients in the linear predictor-outcome datasets

The net benefits attained in (a) linear threshold, (b) nonlinear threshold and (c) quadratic or U predictor-outcome relationship datasets using CAT3, CAT5, linear, FP, RCS models are summarized in Figure 5.9. The top row in Figure 5.9 shows the net benefits from these statistical models in (a) linear threshold, (b) nonlinear threshold and (c) quadratic or U predictor-outcome relationship datasets respectively. The bottom row provides the results showing the net reductions for false positives associated with each prediction model in these datasets.

In Figure 5.9, the CAT3 model had the least clinical net benefits compared to the methods of CAT5, linearisation, FP, and RCS when fitted in both linear and nonlinear threshold datasets. In the linear threshold predictor-outcome dataset, the CAT3 method was clinically useful when the threshold probability, p_t was between 5% and 17% and retained the least benefits in this probability range. In the same dataset, the FP and RCS models had similar clinical net benefits when the threshold probability, p_t was between 5% and 22%. The plots comparing the net clinical benefits of these methods are provided in Figure 5.9 (top left (a)). In Figure 5.9 (top left (a)), the FP and RCS models outperformed the CAT3 method when $14\% < p_t < 22\%$. In addition, the FP and RCS approaches outperformed both the CAT5 and linear models when $17\% < p_t < 22\%$. Graphically, the CAT5 method was clinically useful when p_t was between 5% and 19% whilst the linear model achieved its usefulness when $5\% < p_t < 21\%$. Hence, the linear approach was a better strategy than the CAT5 when $19\% < p_t < 21\%$ and $14\% < p_t < 21\%$ against the CAT3 - which is not better than any strategy of treating patients with the event outcome when $p_t > 17\%$ (see Figure 5.9 - top left (a)).

In nonlinear threshold datasets presented in Figure 5.9 (top middle (b)), categorising the predictor into three categories during analysis produced a model with clinical utility when p_t was between 5% and 18%. In this probability range, the CAT3

slightly outperformed the linear model when p_t was between 5% and 11%. In contrast, the CAT5 (with more categories) had better fit (with larger clinical benefits) than both the CAT3 and linear models when $5\% < p_t < 18\%$. However, the linear model outperformed the two methods of categorisation (CAT3 & CAT5) when p_t was large (between 19% and 24%). Fitting the FP and RCS models in similar dataset produced similar clinical net benefits when p_t was between 5% and 33%. In this probability range, the FP and RCS models had larger clinical benefits compared to the CAT5 method when p_t was between 18% and 33%. Otherwise, the three methods of CAT5, FP and RCS had almost similar clinical benefits when $p_t < 18\%$. When compared to the methods of linearisation, the FP and RCS models had larger clinical benefits when $5\% < p_t < 15\%$ and $20\% < p_t < 33\%$. At the threshold probabilities between 15% and 20% the methods of linearisation retained similar large clinical benefits as in the FP and RCS models. The plots comparing the net benefits of five models in nonlinear threshold datasets are provided in Figure 5.9 – top right (c) below.

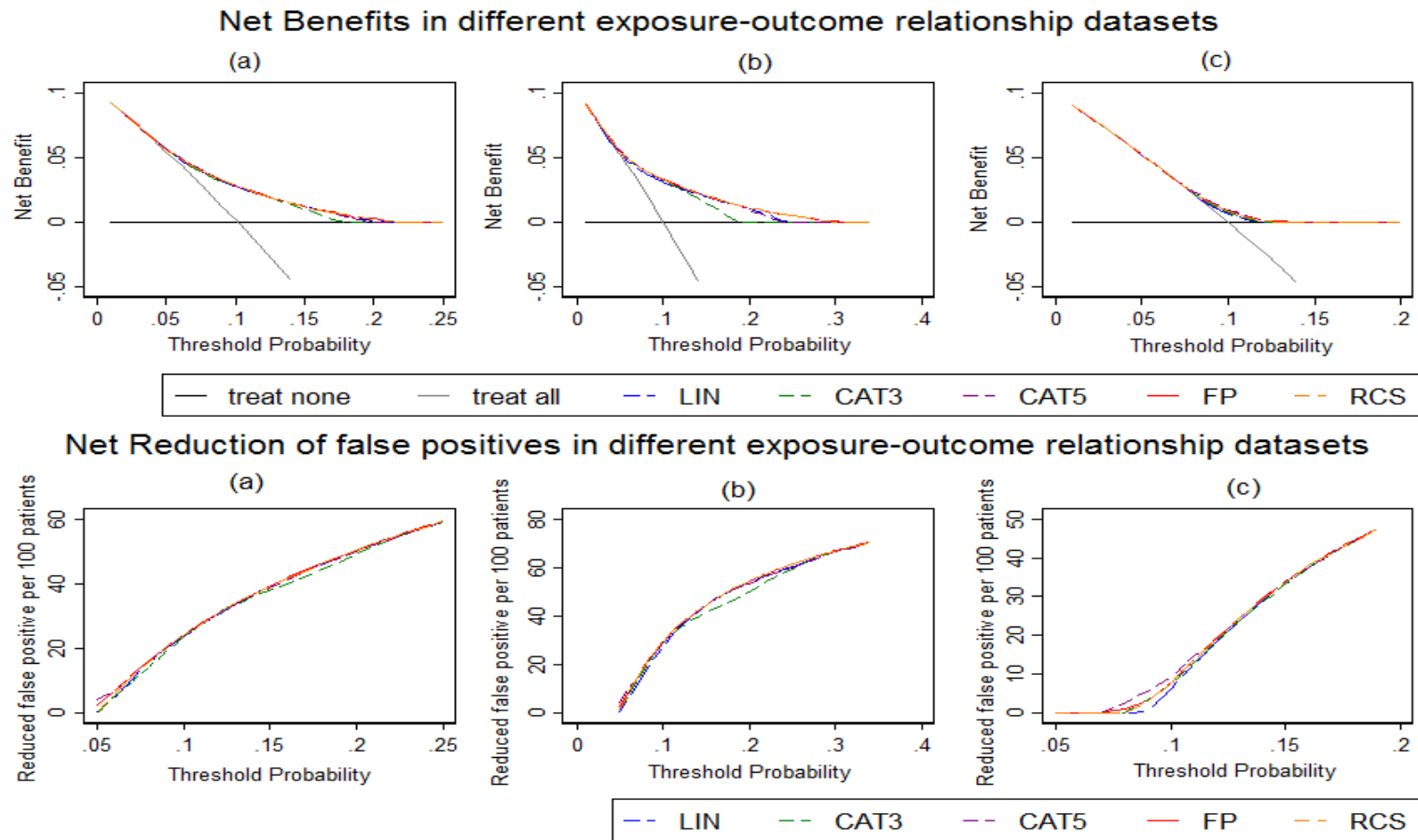


Figure 5.9: The median predicted curves attained from a simulation study (with 1000 replicates) showing the Net Benefits and Reduced false positive per 100 patients using various statistical models (FP, RCS, CAT3, CAT5 and linearisation approaches) in (a) linear threshold, (b) nonlinear threshold and (c) quadratic or U predictor-outcome relationship datasets.

The results showing false-positive reductions per 100 patients in different logistic regression models are provided in Figure 5.9 – bottom. Overall, the CAT5, linear, FP and RCS models performed better than the CAT3 method when fitted in threshold (linear and nonlinear) and quadratic datasets. Evidence of underperformance in CAT3 models was noticeable in threshold datasets. The CAT3 method was struggling at the central distribution of the threshold probabilities, p_t 's in these datasets – producing a smaller reduction in false positives than the CAT5, linear, FP and RCS models (see Figure 5.9- bottom ((a) – (b))). The other additional tables showing detailed analyses output of net benefit and the reduction of false positives per 100 patients using the CAT3, CAT5, linear FP, and RCS models in threshold and quadratic datasets are provided in Appendix D (see Table 5.6 - Table 5.8).

5.5 Discussion

This section starts by discussing the general approach adopted in this chapter. Section 5.5.2 follows next to outline the main findings of this chapter. Limitations, future work emerging from this work are provided in section 5.5.3 and 5.5.4 respectively.

5.5.1 General approach

The practices of categorising or linearising continuous predictor variables in prognostic models have been criticised in the literature (Bennette and Vickers, 2012, Collins et al., 2016). The alternative methods of handling continuous predictor variables include using fractional polynomials (FP) and restricted cubic splines (RCS). FP and RCS are principal competitors for estimating functional forms but there exist few studies comparing the two methods (Hollander and Schumacher, 2006, Govindarajulu et al., 2009, Binder et al., 2013). This chapter investigated through simulations the methods of categorisation and linearisation against the alternative approaches of using

FPs and RCS in logistic regression models with a continuous predictor. In the simulation, two forms of categorisation including CAT3 and CAT5 models were considered for evaluation. The aim was to get an insight on the properties of these approaches in various prognostic models where linear and nonlinear logit functions were observed. Identification of local features such as turning points (or threshold estimation) was also of interest. Hence, the choice and performance of various logistic prognostic models against ‘true’ functions were very critical in this Chapter.

For predictive accuracy, performance measures including discrimination, calibration, and the clinical utility were exhibited to encourage and guide medical researchers with limited statistical background on their usage with CAT3, CAT5, linear, FP, and RCS logistic regression models. The three performance measures of discrimination, calibration, and clinical usefulness have been recommended to improve the reporting of prognostic models in the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis) guidelines - see Item 16 of the recommendations (Moons et al., 2015).

The applications of the FP and RCS models in Stata are readily available. Their implementations in the simulations were performed using the standard (or default) settings available to non-statistician to sensitize and encourage their usage. Furthermore, the simulations were exemplified by using the alcohol-hypertension relationship scenarios in epidemiology. However, the results of the simulation could be extended to similar settings where generalised linear models are applied.

5.5.2 Summary of main results

The practices of handling continuous predictor variables using CAT3 and CAT5 methods were not adequate for prognostic functions assumed in the simulation. The two approaches yielded step functions that did not reflect true shapes in the data - suggesting

risk probabilities that were submerged and varied according to predictor categories. Therefore, the number of predictor categories under the CAT3 and CAT5 methods influenced risk variations in the fitted datasets. For example, the CAT3 (with three categories) allowed the event risk to vary by few (or wider) categories whilst the CAT5 (with five categories) had more (or narrow) categories reflecting risk changes. In contrast, linearisation models performed well only when the relationship between the predictor and outcome was linear. The alternative methods of FP and RCS improved the results of the predictive functions in nonlinear threshold or quadratic datasets. The prognostic models developed using the FP and RCS methods produced fits that were very close to the ‘true’ nonlinear thresholds or quadratic associations in the simulations. In linear threshold datasets, none of the methods investigated in this study were able to identify the actual association shape in the simulations.

In the simulation, the FP functions were characterised by unstable tails at the lower distribution of the predictor. The four logit functions fitted with FPs produced fits with some spike behaviour at zero values of the predictor. This behaviour affected the 95% CIs of the FP models at the lower tails - making them irrationally wide and biologically implausible. In application studies, this problem could be dealt with by ignoring the spike in the fitted models (Lorenz et al., 2017). Alternatively, the FPs could be fitted at high/large predictor values above zero - by shifting the origin of the data, adding a small constant value to predictor values (Royston and Sauerbrei, 2008). The predictive models produced using fractional polynomials are based on natural logarithms thus may not be possible when the predictor assume zero or non-positive values.

The results on optimal thresholds (or turning points) suggested some underestimation when fitting the FP and RCS models in linear and nonlinear threshold

datasets. Both the FP and RCS methods underestimated the ‘true’ optimal probabilities of the event outcome and shifting the positions and the locations of the optimal predictor to the left. Findings on the coverage probabilities of optimal event rates in these threshold datasets suggest under-coverage in the FP models and conservative estimates when fitting the RCS models. For quadratic datasets, there were some inconsistencies in the directions and positions of the estimated optimum predictor when fitting the FP and RCS prognostic models. The FP model underestimated the ‘true’ optimal predictor values – shifting the position to the left. In contrast, the RCS models overestimated the position of the ‘true’ optimal predictor – shifting the position to the right. Apart from that, both the FP and RCS models underestimated the optimal probability of the event outcome in the ‘true’ function. The coverage probabilities of the ‘true’ optimal event rates were less than the 95% nominal level assumed when fitting the quadratic data with the FP and RCS models.

Generally, there was poor discrimination and miscalibration when applying the five methods of analysis in the simulations. In all the simulated datasets, the RCS methods had greater c-index scores whilst the CAT3 retained the least c-indexes. However, the differences between the five methods of analysis were not substantially large or worse. The CAT5 competed fairly with other methods of analysis outperforming the linear and FP models in some scenarios. The results on calibration were characterised by lower predicted probabilities than those observed in ‘true’ functions when applying the two methods of categorisation (CAT3 and CAT5) in the simulations. However, the CAT3 and CAT5 retained better plots than those observed under the linear, FP and RCS models. In contrast, calibration plots attained by methods of linearisation, FP and RCS showed combination of agreements and disagreements (between predicted and observed probabilities) when applied in the same datasets. Reflecting on this, these are important findings, they provide insights on discrimination

and calibration (key measures of prognostic models). There is an ongoing debate about whether excellent discrimination and calibration can both be achieved in prognostic models. In this study, none of the prognostic models achieved both excellent discrimination and calibration in the simulated datasets. These findings are consistent with the argument raised by Diamond (Diamond, 1992) that prognostic models cannot achieve both excellent discrimination and calibration since the models that maximise discrimination does so at the expense of calibration.

The results on clinical usefulness showed better performance when fitting the CAT5, linear, FP and RCS models in linear, thresholds and quadratic datasets than the CAT3 approach. However, the differences between the five methods were not substantially large or worse when applied in linear or quadratic datasets. Thus, the underperformance of CAT3 was predominantly large in linear and nonlinear threshold datasets. The CAT3 struggled at the central distribution of these datasets producing smaller false-positive reductions those observed under the CAT5, linear, FP and RCS models. The findings of poor net-benefits in categorical models (with few categories) were also reported by Collins and colleagues (2016). These authors also found some improvement on net-benefits when the number categories were increased in categorical models. The latter agrees with the results comparing CAT5 against CAT3 models. The CAT5 (with more categories) retained functions with improved net-benefits than the CAT3 fits.

5.5.2.1 The difference between CAT3 and CAT5 approaches

The five categorical (CAT5) models had better predictive ability (in terms of discrimination, calibration and net-benefits) than when three categories (CAT3) were assumed in the simulation. However, the two approaches were similar in character; their application produced step functions that did not reflect actual predictor-outcome relationships. Thus, the improvements on models with large number of categories come

as a trade-off for more complex functions (with more parameters) which would generally be inefficient, e.g. using >10 categories to approximate a non-linear relationship.

5.5.3 Limitations

Firstly, this study has only focused on single predictor models. In reality, there exist other variables that may influence or improve the performances of prognostic models. However, it was not possible to incorporate such variables in the simulations since it was difficult to envision how they would affect the predictive models. The results were clearer when considering one predictor variable. Nevertheless, similar findings may apply in general - when two or more continuous variables are adjusted for in these models.

Secondly, the predictor variable was generated assuming a continuous uniform distribution. This claim is unpopular and ungeneralizable in most settings. Skewed or 'spike' at zero (SAZ) distributions such as the lognormal, normal or a mixture of distribution are popular and more realistic in epidemiology. Thus, the directions and shapes of the predictor-outcome models were likely to be different in similar studies with skewed predictors. For example, if the 'true' predictor-outcome function was steeper at the tails because the predictor effects are concentrated in one end of the distribution, defining the end categories too broadly would obscure the direction and the 'true' predictor-outcome functions in that region. In this situation, the end categories would assume the event outcomes are homogeneous (amongst higher and lower predictor values) which is not correct (Greenland, 1995a, Bennette and Vickers, 2012). Like in this chapter, it is expected that the alternative methods of FPs would produce fits with some spike behaviour at the zero values of the predictor. However, this behaviour would be more striking - skewed or SAZ predictor variables have larger proportions of zeros than the uniformly distributed measures. This behaviour is

influence by FPs inability to deal with zero predictor values (their FP powers of the form $x^r [\ln(x)]^j$ are included in the family of FP curves) hence only require positive values (Greenland, 1995b). In contrast, the RCS models would produce linear fits at the extreme tails of the predictor variables. The RCS are naturally constrained to be linear above the last knot and below the first knot (often placed at the tails of the predictor distributions).

The third issue is about interpretation of certain features of the FP and RCS curves. The RCS fits linearly at the tails whilst the FP tends to fit curves even when it should be linear. Thus, true curves may be missed and misinterpreted in skewed datasets (with a large proportion of zeros and non-negative values). Hence, some statistical methods developed for this problem may be appropriate. Examples of such methods include, two-part models proposed by Duan et al (1983) and the compound Poisson exponential dispersion model proposed by Jørgensen (1987, 1997). However, a two-part model by Duan et al (1987) is suggested as a reasonable approach for many application studies (Min and Agresti, 2002).

5.5.4 Future work

The present simulation work focused on developing prognostic models based on single predictor variable. The motivation was to assess the four modelling approaches in various nonlinear risk functions excluding the influence of other possible variables (a priori) and maximise on generalizability. However, in reality, epidemiological studies are characterised by many covariates that influence and affect the final models. Therefore, an application study is suggested to assess the methods of categorisation, linearisation, FPs and RCS under the multivariable setting.

5.6 Conclusions

Based on the results of this chapter, researchers may be tempted to use large number of categories for predictive analysis. However, the performance improvements on such models come as a trade-off for more complex functions that are inefficient. Thus, this research concludes by recommending that:

- i. Researchers do not categorise continuous predictors for development of prognostic models. Alternative approaches for handling continuous predictor variables such as FPs and RCS are available.
- ii. Flexible regression approaches including fractional polynomials or restricted spline models be used as a minimum check for the presence of nonlinearity when the 'true' predictive function is unknown.
- iii. When nonlinearity is suspected or unknown, researchers should be careful about the reporting of clinical thresholds (or turning points) using the FP and RCS functions. The estimation of clinical thresholds (or turning points) using these functions in unknown relationships could be misleading – producing inconsistent findings.
- iv. When the predictive function is suspected to be linear, the assumption of linearising the predictor or applying RCS produces adequate functions. The FP method is not sufficient for linear predictive models; produce fits that are biologically implausible – characterised by artefacts at the lower tails of the predictor variable and extremely wider CIs.

Finally, if prognostic models are to be used by practitioners or clinicians, it is important to validate the risk prediction for clinical credibility, accuracy, and efficiency. Thus risk performance measures including discrimination, calibration, and clinical utility must always be performed for validation of any prognostic model.

Chapter 6

Examining the alcohol-hypertension association in type 2 diabetes patients using the UK Biobank

6.1 Background

Diabetes (also known as Diabetes Mellitus (MD)) is amongst serious non-communicable diseases targeted by world leaders for action due to steadily increasing cases reported over the past decades (World Health Organization, 2016). The estimated global estimates for the adult population living with diabetes stood at 422 million in the year 2014. The estimates depict a twofold increase in the global prevalence rate of 4.7% in the 1980s to 8.5% in 2014 (World Health Organization, 2016).

The consequences of diabetes include morbidity and mortality and are accelerated by allied complications and sequelae of hypertension such as kidney diseases, cardiovascular diseases, neuropathy, blindness and lower extremity amputations (Bebb et al., 2007, Deshpande et al., 2008). Evidence suggests that hypertension is significantly higher in diabetic population than in non-diabetics particularly common amongst those with type 2 diabetes (Barnett, 1994, World Health Organization, 2016). A recent review suggests that more than 60% of patients with type 2 diabetes have hypertension (Colosia et al., 2013). Therefore, maintaining tighter and lower blood pressure levels amongst patients with type 2 diabetes is essential to control and manage allied complications. However, this is not easy; the number of patients diagnosed with hypertension amongst those with type 2 diabetes is increasing. The reasons for rising prevalence of hypertension in type 2 diabetes population maybe attributed to lifestyle factors such as the consumption of high-calorie diets and

sedentary behaviour in different racial, ethnic, and social groups (Lago et al., 2007, Blomster et al., 2014).

Alcohol consumption is another major lifestyle factor associated with the disease complications amongst patients with type 2 diabetes (Beulens et al., 2005, Blomster et al., 2014). Heavy or excess alcohol intake is known to elevate blood pressure levels (Stamler et al., 2003, Mori et al., 2016) perhaps due to increased levels of low-density lipoprotein (LDL), increased blood clotting and changes in the myocardium and ventricular fibrillation, which are all linked to adverse cardiovascular outcomes (Mckee and Britton, 1998). Thus, some studies recommend moderate alcohol drinking (Razay et al., 1992, Blomster et al., 2014, Gepner et al., 2015). The consumption of alcohol in moderate quantities is associated with reduced incidence of risk amongst cardiovascular diseases and mortality (Razay et al., 1992, Blomster et al., 2014, Gepner et al., 2015), due to increased levels of high-density lipoprotein (HDL) cholesterol and reduced coagulation (Pearson, 1996). In patients with type 2 diabetes, the risk-benefits of moderate alcohol consumption is questionable due to the recommendation of tighter blood pressure levels (Judd et al., 2011, Gepner et al., 2015, Gepner et al., 2016). Some studies discourage patients living with type 2 diabetes of using alcohol in moderation (Bantle et al., 2008). In 2012, a systematic review and meta-analysis reported an increasing trend between moderate alcohol consumption and blood pressure in males and a protective relationship with a decreasing risk of hypertension in females (Briasoulis et al., 2012). These findings and recommendations discouraging the use of alcohol are contradictory and complicates the potential benefits of alcohol consumption in people living with diabetes. The National Institute for Health and Care Excellence guidelines (NICE, 2015) recommends an individual preference for moderate alcohol consumption (3-4 units/day in men and 2-3 units/day in women).

This chapter aims to assess the association between alcohol consumption (exposure) and hypertension (outcome) in patients with type 2 diabetes using the UK Biobank data. A causal relationship between alcohol consumption and hypertension has been established in cross-sectional and prospective studies using the general population datasets (Chang and Park, 1991, Gillman et al., 1995, Moreira et al., 1998, Fuchs et al., 2001). However, this relationship is not well studied in a population with diabetes patients (Saremi et al., 2004). The presence of diabetes has the potential to modify the alcohol-hypertension relationship worsening the risk of hypertension at every unit of alcohol intake. The UK Biobank provides large, generalizable and contemporary data that is sufficient to investigate hypertension in patients with type 2 diabetes (Allen et al., 2012). A recent study by Eastwood and colleagues (2016) reported approximately 5% (23,842/502,619) of participants with type 2 diabetes in the UK Biobank. The proportion is similar to the prevalence rate of 5% (3,500,000/66,000,000) reported in the UK population with type 2 diabetes (Diabetes UK, 2015, National Statistics, 2017).

In this study, the confounding factors for adjustment were identified using a causal diagram known as a Directed Acyclic Graph (DAG). The theory of DAGs and its application in the UK Biobank data is introduced in the next section.

6.1.1 Introducing DAGs

In epidemiology, the issue of establishing causality is a challenging one. Evidence of causation cannot be validated based on non-experimental studies. Non-experimental studies can only offer evidence for association. Strong evidence for causation requires experimental data - which is rare and expensive to produce (Law et al., 2012). To strengthen causal inference in non-experimental studies, causal path diagram have been suggested (Greenland et al., 1999). The theory of causal path diagrams (focusing on the Directed Acyclic Graphs (DAGs)) is provided below.

6.1.1.1 The theory of DAGs

This section describes the theory of DAGs, focusing on the structure, sources of bias and selection of ‘minimal sufficient’ set of covariates. Further details supporting the theory presented here can be attained in the references including Greenland et al (1999), Law et al (2012), Sauer and VanderWeele (2013) and Textor (2013).

6.1.1.1.1 Definition and use of DAGs

Textor (2013) defined a DAG as “*a graphical model that depicts a set of hypotheses about the causal process that generates a set of variables of interest*”. Essentially, DAGs are used to encode investigators’ *a priori* assumptions about the associations between variables in causal structures. Additionally, DAGs help researchers to achieve the following: (1) diagnose the sources of bias and (2) select a set of covariates that explains or allows the estimation of causality from observed data (Greenland et al., 1999, Sauer and VanderWeele, 2013).

6.1.1.1.2 The structure, source of bias and selection of covariates

The structure of DAGs contains directed *arcs* (arrows), linking *nodes* (variables) and their *paths*. A *path* suggests the existence of *known*, *likely* and *assumed* relationships between any two variables, with an arrow representing causality (Law et al., 2012). For instance, ‘A causes B’ would be represented as $A \rightarrow B$, where A and B are nodes and the arrow between them is an arc. In the example, $A \rightarrow B$, A is the *parent* nodes whilst B is the *child*. In a path connecting three nodes such that $A \rightarrow B \rightarrow C$, A is known as the *ancestor* of C, and C is a *descendent* of A; while B is a *child* of A and *parent* of C. The node B lies on the causal pathway between A and C thus is considered as a *mediator* variable. The paths with all arcs following the same direction of causality (as in the examples) are known as *direct* (or *causal*) *paths*. In contrast, a non-causal path is known as a *backdoor path*. For example, when $A \leftarrow B \rightarrow C$, a backdoor path exist

between A and C through B. The node B is a common cause of A and C thus is considered as a *confounder* variable. A *blocked path* occurs if it contains at least one *collider*. A node is a *collider* when both arcs entering and leaving the node have arrows pointing at it. For example, when $A \rightarrow B \leftarrow C$, the path between A and C is blocked by a node B (a collider). The node B is also called an outcome (or common effect) of A and C (Greenland et al., 1999, Law et al., 2012, Textor, 2013). In DAGs, *blocked path* represent independence whilst the *unblocked path* indicate the presence of an association between variables (Sauer and VanderWeele, 2013).

Hence, DAGs can be used to infer on *dependence* and *conditional independence* when their causal structure is correctly specified. The procedure linking the structure of a DAG to statistical independence is known as the *d-separation criterion* (Greenland et al., 1999, Sauer and VanderWeele, 2013). Suppose X and Y nodes are d-separated conditional on Z if all the paths from X to Y are *blocked conditional on Z*; then if a DAG is correctly specified, X and Y are *conditionally independent* given Z. In contrast, *conditional associations* occur when adjusting for *colliders*. *Mutually independent* variables with a common effect become *conditionally dependent* when statistically adjusting the common effect. This process opens up the backdoor paths and introduces confounding errors on the estimates (Greenland et al., 1999, Sauer and VanderWeele, 2013). The removal of confounding errors requires background knowledge to differentiate colliders, mediators, and confounders when creating a causal DAG. If the full causal structure is unknown, colliders, mediators, and confounders may behave the same in the exposure-outcome models. To guide the development of a complete DAG the following consideration are necessary according to Sauer and VanderWeele (2013):

- i. Creating DAGs should not be restricted to measure variables from the study data.

- ii. Capturing any common cause of any other two variables on the DAGs.
This is an important point of demonstrating causality.
- iii. Variables that only causally relate to one other variable maybe captured or ignored. However, common causes must always be captured for causality.
- iv. Identifying a set of covariates that minimises the confounding bias on the DAGs. The minimal set of covariates blocks all the backdoor paths and does not open closed pathways by conditioning on the colliders.

Based on these guidelines, a suitable DAG provided in Figure 6.1 was constructed to identify a set of covariates to adjust for counfounding in the alcohol-hypertension models using the UK Biobank. An online DAGitty software available at (www.dagitty.net/) was used to encode and specify relationships between variables in a DAG presented. The causal relationships assumed were justified using evidence from the literature (see Appendix E).

6.1.1.2 Application of DAGs using the UK Biobank data

Figure 6.1 provides a hypothetical causal relationship between alcohol consumption (as exposure variable) and hypertension (outcome variable) taking into consideration possible mediators, competing exposures, confounding variables and colliders available in the UK Biobank dataset. The schematic diagram was constructed taking into account multiple lifestyles, dietary and behavioural factors affecting the relationship between alcohol consumption and the risk of having hypertension.

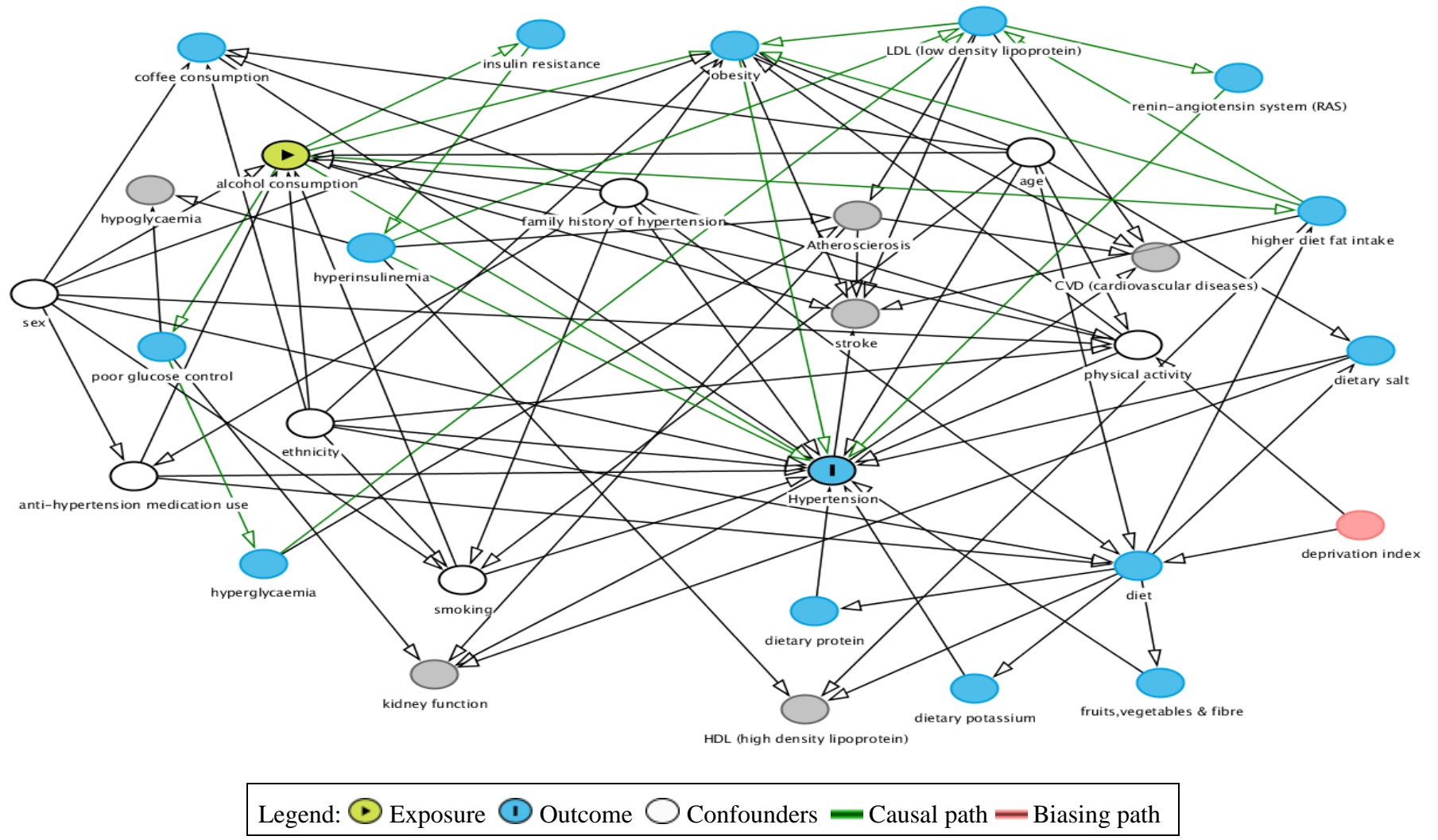


Figure 6.1: DAGitty schematic view of confounding adjustment for an alcohol-hypertension relationship

In Figure 6.1 a set of confounders satisfying the condition of ‘minimal sufficiency’ include age, anti-hypertension medication use, family history of hypertension, physical activity, ethnicity, sex, and smoking. This set of confounding variables closes all non-causal pathways between alcohol consumption and hypertension in the diagram. Adjusting for these variables will be sufficient to remove the confounding bias in the alcohol-hypertension models.

Currently, no analysis has been carried out in diabetes patients to examine risk modification factors of hypertension in the UK Biobank (Eastwood et al., 2016), the present study also aims to investigate how age (continuous variable) and use of antihypertensive medication (binary variable) modify the alcohol-hypertension in the data. Establishing how age and the use of antihypertensive medication modify the alcohol-hypertension relationship may guide health practitioners to develop target specific strategies or interventions to control and manage hypertension.

6.1.1.2.1 Why effect modification of age and use of antihypertensive medication?

When investigating the alcohol-hypertension relationship, it is important to consider the effect modification of age and antihypertension medication use. This is because the benefits or harm associated with alcohol consumption may differ depending on individual’s age or on whether a patient is on antihypertension medication or not. In the literature, little is known about the age-related difference of alcohol consumers on the risk of hypertension (Vanleer et al., 1994). In this chapter, the age-related differences and the chances of having hypertension are investigated assuming *alcohol X age* interactions in the adjusted models.

For patients on antihypertension medication and non-users, it is also not clear whether the association between alcohol consumption and hypertension varies in the two groups (Wakabayashi, 2010). There exist little research in this area (Beevers et al.,

1990). To explore this area, an investigation on whether medication use modifies the relationship between alcohol consumption and hypertension in type 2 diabetes patients is also performed. In the analysis, it is hypothesized that the *alcohol X medication use* interactions are present in the adjusted models.

6.1.2 Specific study objectives

The main aim of this chapter is to investigate the alcohol-hypertension relationship in patients with type 2 diabetes using various logistic modelling approaches. The issues of interest about the alcohol-hypertension association include whether the relationship is linear/nonlinear or involves a threshold dose of alcohol, whether age or antihypertensive medication use modify the benefits/harm associated with alcohol drinking. Using the UK Biobank, the specific objectives of this chapter includes:

- i. To investigate the association between alcohol consumption and the odds of hypertension in patients with type 2 diabetes adjusting for selected confounding variables identified using a DAG.
- ii. To investigate effect modification of age and antihypertensive medication use in the adjusted multivariable alcohol-hypertension models.

6.1.2.1 The rationale of the study

The study is important due to public and clinical interest in the subject. Investigating the relationship between alcohol consumption and hypertension in patients with type 2 diabetes is necessary to assist clinicians in developing target strategies and interventions to control the harm associated with alcohol drinking. In addition, this study forms an example of methods studied in Chapter 4 and Chapter 5. However, the focus is on explanatory analysis since epidemiologists are mostly interested in causal inference than predictive analysis.

6.2 Subjects and Methods

6.2.1 UK Biobank participant's characteristics

This research used self-reported baseline data from the UK Biobank with over 500, 000 participants aged between 40 to 69 years. The UK Biobank was set-up between 2006 and 2010 by recruiting participants from the National Health Service register using healthcare data linkage systems (Eastwood et al., 2016). About 9.2 million people living within 25 miles (40 km) of the 22 UK Biobank assessment centres located throughout England, Wales and Scotland were invited by mail to participate and the response rate of 5.5% was achieved (Allen et al., 2012, Fry et al., 2017). The aim was to provide a resource that will enable researchers to investigate genetic, environmental and lifestyle determinants of a wide range of diseases in middle and older age population in the UK (Allen et al., 2012). Amongst those who agreed to participate in the study, baseline information collected includes the data on lifestyle, environment, medical history, physical measurements, and biological samples. Detailed information on how the data was collected and other assessment procedures may be found elsewhere (<http://www.ukbiobank.ac.uk>). The data is available to other researchers worldwide provided an application for use has been granted by the UK Biobank team. The application and review process is done online through the UK Biobank website and is carried out in four stages including registration, submission of preliminary application, submission of main application and the signing of material and transfer agreement (MTA) contract. Overall, the whole application and review process requires 3-4 months to be completed. A detailed schematic diagram showing the UK Biobank application process is provided in Figure 6.2. The research application protocol approved by the UK Biobank team for this research is provided in Appendix G as a supplementary material.

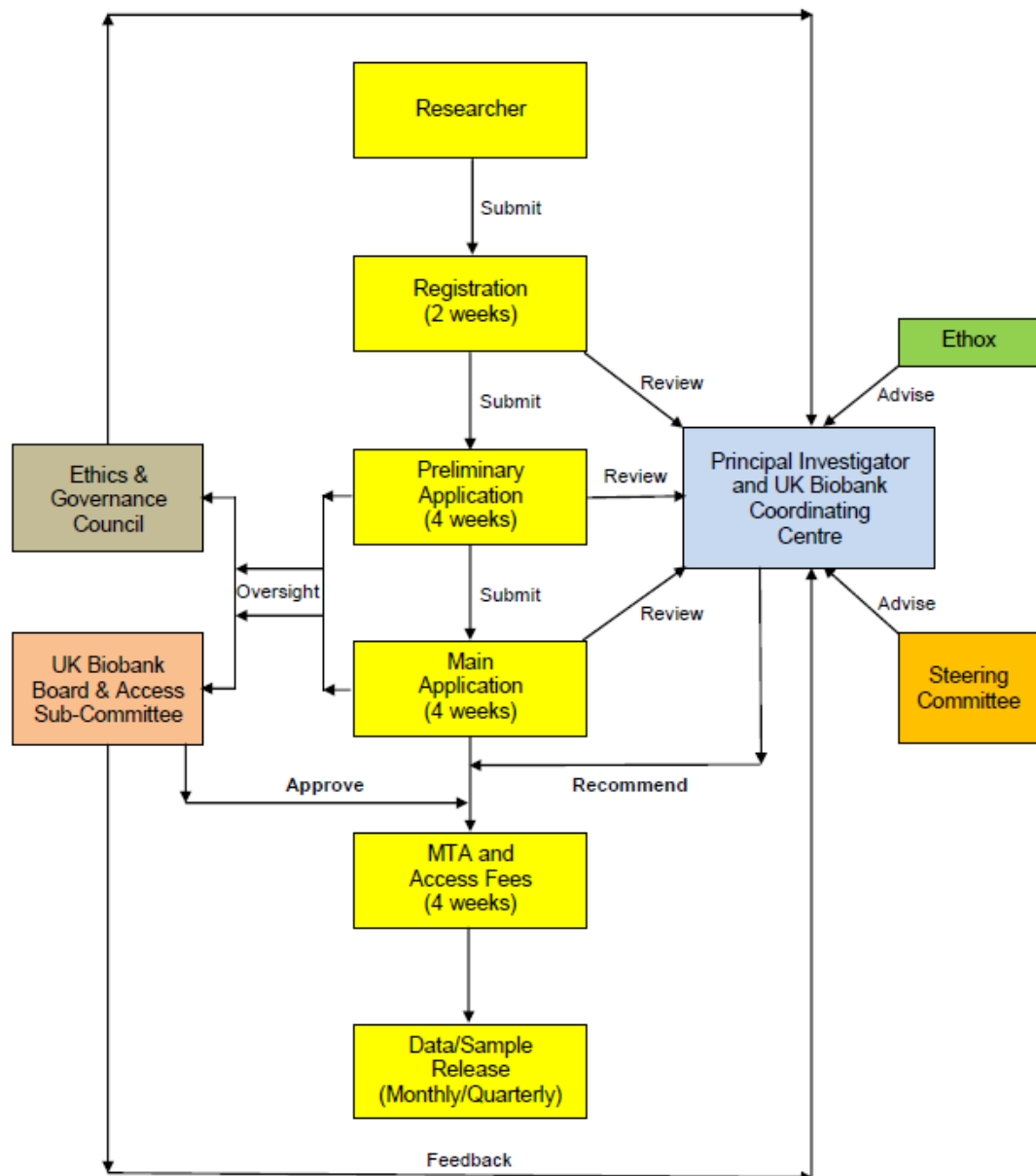


Figure 6.2: The UK Biobank application and review process

Source: adapted from the UK Biobank Access procedure manual (UK Biobank, 2011a)

6.2.1.1 Type 2 diabetes patients

This thesis chapter focuses on a subsample of 23,842 type 2 diabetes patients identified using the algorithm developed by Eastwood and colleagues (Eastwood et al., 2016) when defining patients with diabetes in the UK Biobank. In medical studies using self-reported data, the prevalence of diseases cannot be established with certainty. Bergmann and colleagues (Bergmann et al., 2004) observed some disagreement on the

medical history of participants using self-reported and in-person administered interviews. To bridge this gap, Eastwood et al. (2016) developed and ran an algorithm to assign prevalent diabetes and types using the self-reported information in the UK Biobank. The medical history information collected through the online touchscreens questionnaire and nurse's interviews were used in the algorithm defining patients as 'probable' or 'possible' diabetes cases when they have greater certainty or less certainty of the disease respectively.

Therefore, this study aims to include all type 2 diabetes patients identified in UK Biobank who meet the following criteria:

6.2.1.1.1 Inclusion Criteria

- Self - reported type 2 diabetes (nurse interviews)
- Self-reported type 2 diabetes medications (online touchscreens and nurses interviews)
- Age of diagnosis for diabetes ≥ 36 years (amongst European origin) or ≥ 30 years (amongst South Asian or African-Caribbean origin).

Participants excluded from this study were identified as follows:

6.2.1.1.2 Exclusion criteria

- Non-diabetes participants (i.e. all patients not reporting any diabetes from the nurse interviews, or gestational diabetes in both the nurse's interviews and touchscreens or any diabetes medication in both nurses and touchscreens)
- Self-reported type 1 diabetes patients that also includes self-reported insulin use < 12 months post-diagnosis or self-report current insulin use (touchscreens and nurses interviews)

- self-reported gestational diabetes amongst the females who were pregnant (touchscreens and nurse's interviews)

6.2.1.2 Hypertension and blood pressure measurements

The definition of hypertension in people living with diabetes as used in this chapter refers to any patient with systolic blood pressure (SBP) levels at or above 130 mmHg (World Health Organization, 2006, British Cardiovascular Society et al., 2014).

The SBP measurements were collected using the Omron digital blood pressure monitor which was operated by registered, trained and certified nurses during the assessment visits at the clinics (UK Biobank, 2011b). The blood pressure of each participating patient was collected and reported in millimeter per mercury (mm Hg). The participants were asked to sit on a chair with their feet parallel to each other, toes pointing forward and soles of feet flat on the floor. The right arm was used to take the measurement. Participants were asked to loosen or remove any restrictive clothing that could obstruct circulation of blood. Since the resting BP measurements were required, the nurses took care not to engage the participants in the conversation. This procedure was repeated for the second BP measurements. After completing the first measurement, the participants were allowed at least one minute break for rest; the rubber inflation tubing was disconnected from the Omron monitor with cuff left in place. The participants were asked to gently shake their arm and open and close their hand before the second measurement commenced (UK Biobank, 2011b). In the analysis, the two repeated systolic blood pressure measurements were averaged for use; classifying patients as having high blood pressure or not.

6.2.2 Alcohol intake estimates

The data on alcohol consumption was collected during the assessment visits at clinics using the touch screens and web-based questionnaires. In the questionnaire, alcohol drinkers and non-drinkers were identified by responses relating to the question

“About how often do you drink alcohol?” Non-alcohol drinkers responded as “Never” whilst alcohol drinker’s responses were selected from the options including “daily or almost daily”, “three or four times a week”, “once or twice a week”, “one to three times a month”, and “special occasions only”. The responses for participants who drink more often than once or twice a week were followed by another question requesting them to provide average weekly consumption in numbers of pints, glasses (in various sizes) of beers or ciders, wines, champagnes, spirits, fortified wines and other alcoholic drinks including alcopops. Otherwise, those who consume alcohol occasionally or one to three times a month were asked to provide average monthly consumption. Capturing intake in weekly or monthly quantities is essential since alcohol consumption is an episodic event. Weekly/monthly quantities provide coverage for occasional drinkers who are likely to be missed amongst those reporting daily alcohol intake.

To standardise the units of alcohol consumption provided by respondents, the number of drinks were converted into grams of alcohol consumption per day based on the UK alcohol guidelines (House of Commons Science and Technology Committee, 2012) using the following procedures:

- i. According to the UK alcohol guidelines (House of Commons Science and Technology Committee, 2012), one unit is equivalent to 8g or 10 ml of any standard alcohol drink. In this chapter, a small glass of wine or champagne (125 ml) is equivalent to 1.5 units; a pint of beer or cider was taken as 3 units (a full-strength pint of a beer or cider is 4 units whilst light beers or cider is 2 units). Otherwise, one shot of spirits or fortified wines is 1 unit whilst other alcoholic drinks including alcopops were taken as 1.5 units. These conversion rates were adopted from the Health Survey of England (HSE) report (Fat and Fuller, 2012). Major surveys including the General Household Survey (GHS) estimate the contents of

alcohol based on these conversion units. A detailed conversion table adapted from the HSE 2011 report is provided in Table 6.1.

- ii. Based on the standard conversion provided in (i) above, the amount of alcohol units or contents for each beverage type consumed was calculated then converted in grams by multiplying the number of units by 8.
- iii. For each alcohol drinker, the total consumption per day/week/month in grams was summed considering all beverage type.
- iv. Finally, where weekly alcohol consumption was provided, the sum in (iii) was divided by seven days to obtain the average daily consumption. Amongst participants reporting monthly alcohol consumption, the sum in (iii) was divided by four to attain weekly quantities. The resulting weekly consumption was further divided by seven for the average daily intake.

Table 6.1: The conversion units for estimating the contents of alcohol drinks

Type of drink	Measure	Units of alcohol
Normal strength beer, lager, stout, cider, shandy (less than 6% ABV)	Pint	2
	Can or bottle	Amount in pints multiplied by 2.5
	Small cans (size unknown)	1.5
	Large cans or bottles (size unknown)	2
Strong beer, lager, stout, cider (6% ABV or more)	Pint	4
	Can or bottle	Amount in pints multiplied by 4
	Small cans (size unknown)	2
	Large cans or bottles (size unknown)	3
Spirits and liqueurs	Glass (single measure)	1
Sherry, martini and other fortified wines	Glass	1
Wine	Small glass (125ml)	1.5
	Medium glass (175ml)	2.0
	Large glass (250ml)	3.0
	Bottle	9.0
Alcopops	Small can or bottle	1.5

Source: Adapted from the Health Survey of England report (Fat and Fuller, 2012)

6.2.3 Statistical analysis

To characterise the alcohol-hypertension relationship adjusting for confounders, the methods of categorisation, linearisation, fractional polynomials and restricted cubic splines were applied using logistic regression models. Details of these methods are provided in Chapter 2 of this thesis. The next sub-sections 6.2.3.1-6.2.3.4 highlight the particular assumptions made under each modelling approach as specifically applied in this chapter.

6.2.3.1 Categorisation

In categorical analysis, alcohol drinking patients were divided into three groups based on the amount of alcohol consumption (in g/day). The fourth group constitutes non-alcohol drinkers (0 g/day). Within the alcohol consumers, the boundaries or cut-points of the three drinking categories were established using tertiles of the alcohol

consumed. To compute the odds of hypertension, the three groups were then compared to non-alcohol drinkers (0 g/day) adjusting the model for confounding variables. Continuous confounders were entered in the model as linear variables in the analysis without any transformation and categorical confounders were entered as dummy variables.

6.2.3.2 Linearisation

In linear modelling, the alcohol consumption amounts were kept continuous and were analysed assuming the alcohol-hypertension relationship was linear. A unit increase in the amount of alcohol consumption was assumed to have a constant change on the odds of hypertension across the range of alcohol intake values. That is, an increase in alcohol consumption from 5 to 6 g/day results in the same change on the odds of hypertension as an increase from 59 to 60 g/day. Further, the alcohol-hypertension relationship was adjusted for confounders taking continuous confounders as linear variables and treating categorical confounders as dummy variables.

6.2.3.3 Fractional polynomials

In the analysis involving fractional polynomials, the alcohol-hypertension relationship was assumed to be nonlinear. FPs has the advantage of keeping the alcohol consumption measures continuous and allowing for nonlinearity in the data. During model building, the alcohol intake values were power transformed to allow the alcohol-hypertension relationship to be fitted with first or second order degree FP function (that is, for FP_m , $m \leq 2$ degree model). In practice, FP functions with $m > 2$ are rarely observed thus, FP1 or FP2 models should be sufficient for alcohol-hypertension relationships reported in epidemiological studies. Chapter 2 provide additional details on how the FP models are usually set up based on a set of restricted powers. Moreover, once the FP1 and FP2 models was established, the best fitted function was selected

using both the Akaike's Information Criterion (AIC) score and the likelihood ratio test (LRT).

To account for confounding variables identified through the DAG, the best fitted FP function was then adjusted for confounding by entering continuous covariates as linear variables and categorical covariates as dummy variables in the final multivariable model.

The multivariable fractional polynomial (MFP) algorithm (Royston and Sauerbrei, 2005) developed for the multivariable model building was not applied in this study. When implemented, the MFP algorithm combines the selection of variables using the backward elimination (BE) process and determine their functional forms through the FP function selection procedure (FSP) (Royston and Sauerbrei, 2005, Sauerbrei et al., 2007). Selection of variables through the automated stepwise procedures such as the BE is known to produce biased results with inflated p-values and standard errors (Blanchet et al., 2008).

6.2.3.4 Restricted cubic splines

To estimate the RCS model, the alcohol consumption data was split into a series of connected 'segments' joined with k knots. The number of knots, k was allowed to vary such that $k = 3$ or 4 implying two different RCS models were fitted in the analysis. The purpose of varying the number of knots was to allow and assess flexibility within the fitted models. The proposed knots should be sufficient for any plausible alcohol-hypertension shapes reported in epidemiological studies. The RCS models with $k \geq 5$ are likely to produce over fitted functions.

Within the fitted RCS functions, the knots positions were equally spaced across the percentile distribution of alcohol consumption data using Harrell's method of knots placement described in section 2.2.5.1. This knots selection method is less subjective

and allows reproducibility and comparison of results between studies (Heinzi and Kaider, 1997). At the two extreme knots (below the first knot and above the last knot), the alcohol-hypertension relationship was assumed to be linear. Nonlinearity was assumed to be occurring inside the inner knots. Based on this approach, an adequate RCS fit (with $k = 3$ or 4 knots) was chosen using both the AIC score and the likelihood ratio test (LRT).

The multivariable model was developed by entering continuous confounders as linear variables and treating categorical confounders as dummy variables in the best fitted alcohol-hypertension function.

Overall, the models attained with the procedures above were ranked according to their AIC scores. The model with the least AIC score suggest a better fit. Thus, the AIC scores for these models were reported for both unadjusted and adjusted functions.

6.2.4 Stratification analysis

Beyond adjusting for confounders, analyses were performed to investigate whether the use of anti-hypertensive medication (binary predictor) and age (continuous predictor) modify the relationship between alcohol consumption and hypertension. An additional interest was also to explore if effect modification was preserved in nonlinear functions. To achieve these objectives, effect modification was investigated using margins plots for visual and interpretable graphs (Williams, 2012, Royston, 2013). Typical regression output tables with interaction terms and p-values are hard to interpret and communicate to readers when nonlinearity is present in the data (Lamina et al., 2012).

In the analyses, the two interaction terms (*alcohol X age*) and (*alcohol X medication use*) were entered and assessed separately in different multivariable models obtained through the categorisation, linearisation, FP and RCS approaches. For each

modelling approach, two multivariable models were tested; one with *alcohol X age* interaction term and another one with the *alcohol X medication use* interaction term.

The procedures for testing these interactive terms include:

- i. Plotting the difference in probabilities of hypertension between patients who are not on medication and those using anti-hypertensives at different values (or categories) of alcohol consumption. This was applicable in models inspecting categorical by categorical or continuous by categorical interaction terms.
- ii. Assessing the average marginal effect (AME) of age on the probability of hypertension assuming values of alcohol consumption are held constant (vice-versa).
- iii. Plotting the predicted probability of hypertension for all combination of age (in years) and some specified range of alcohol consumption (in g/day) using contour or functional diagrams. Note: (ii) and (iii) were applied in adjusted models where two continuous variables are being assessed for interaction.

6.2.5 Sensitivity analysis

Sensitivity analysis was performed to evaluate the influence of type 2 diabetes status on key findings. This was done by excluding ‘possible’ type 2 diabetes patients in the analysis. Using the same procedures in sections 6.2.3.1 - 6.2.3.4, the assessment of the alcohol-hypertension relationships was further replicated in ‘probable’ type 2 diabetes patients to validate the final conclusions.

6.3 Results

6.3.1 General characteristics

Table 6.2 shows the characteristics of 23,842 patients with type 2 diabetes included in the study for analysis. Counts (percentages) were used for describing the occurrence of categorical variables, means (standard deviations) to summarise the distribution of continuous variables which are normally distributed and median (interquartile range) for variables which deviate from normality. Of 23,842 patients, 20,569 (86%) and 3,273 (14%) individuals were classified as ‘probable’ and ‘possible’ type 2 diabetes cases respectively. A t-test performed to assess the mean systolic blood pressure difference between ‘probable’ and ‘possible’ type 2 diabetes cases was not statistically significant ($t=0.33$, $p\text{-value} = 0.74$). A similar mean systolic blood pressure of 141 mmHg was observed amongst ‘probable’ and ‘possible’ type 2 diabetes patients with standard deviations of 17.2 and 18.2 respectively.

The assessment of data quality revealed 5% ($n=1,338$) of type 2 diabetes patients with missing values on systolic blood pressure (SBP). This implied that 5% of the total sample ($n=23,842$) would be omitted in the analysis because of missing SBP readings. The alcohol drinking status was not revealed in nearly 0.5% ($n=114$) of the respondents. Amongst those identified as current alcohol drinkers ($n=19,773$), the information on the amount of alcohol consumption was missing on nearly 22% ($n=4425$) of the participants. A summary of missing data provided in Table 6.2, suggest incompleteness on family history of hypertension - 4% ($n=1017$), smoking status - 1% ($n=256$), ethnicity - 0.8% ($n=202$), and physical activity - 10% ($n=2285$). Apart from this, the data on age, sex, and anti-hypertension medication use were complete.

The general characteristics of type 2 diabetes patients suggested that participants were adults in middle and older ages with a median age of 62 years (IQR=56-66). Male participants were more dominant with 63% ($n=15,009$) compared to 37% ($n=8,833$) of

females. The majority of participants, 86% (n=20,531) were from the White ethnic group. The other ethnic groups including the Blacks, Asians, and Mixed constituted about 13% (n=3,109) of total patients classified as diabetic (see Table 6.2).

Based on the definition of hypertension used in this study (any patient with SBP levels at or above 130 mmHg), 70% (n=16,659) of patients with diabetes were also hypertensive with the mean SBP of 148 mmHg (SD=13.8). Apart from the latter, 25% (n=5,845) of patients with diabetes were non-hypertensive with the mean SBP of 121 mmHg (SD=7.3). Amongst the 16,659 patients with hypertension and diabetes, 49% (n=8,240) individuals reported a family history of high blood pressure and 67% (n=11,163) were using antihypertensive medications.

The majority of individuals, n=13,019 (54%) were either previous or current smokers whilst n=10,567 (44%) have never smoked. On average patients moderately exercised 3 days (SD=2.4) per week. The data on alcohol consumption was rightly skewed showing daily median intake of 9.7 g with interquartile range values between 0.2 and 25.1 g (see Table 6.2 and Figure 6.3 below).

Table 6.2: The general characteristics of n=23,842 diabetes patients included in the study

Data characteristics	Overall (n=23,842)
Age, years (median [IQR])	62 [56-66]
Sex (n [%])	
Female	8,833 (37%)
Male	15,009 (63%)
Hypertension (n[%])	
Yes	16,659 (70%)
No	5,845 (25%)
Antihypertensive medication use (n[%])	
Yes	15,421 (65%)
No	8,421 (35%)
The family history of hypertension (n [%])	
Yes	11,664 (49%)
No	11,161 (47%)
Smoking status (n [%])	
Never	10,567 (44%)
Previous	10,406 (44%)
Current	2,613 (11%)
Ethnicity (%)	
Whites	20,531 (86%)
Asians	1,683 (7%)
Blacks	852 (4%)
Mixed/Other ethnic groups	574 (2%)
Physical Activity (PA) (mean [SD])	
No. of days per week of moderate PA 10+ minutes	3.3 [2.4]
Alcohol intake (median [IQR])	
Alcohol intake, g/day	9.7 [0.2-25.1]

A graphical display of alcohol consumption (measured in g/day) amongst this study sample is given in Figure 6.3 below:

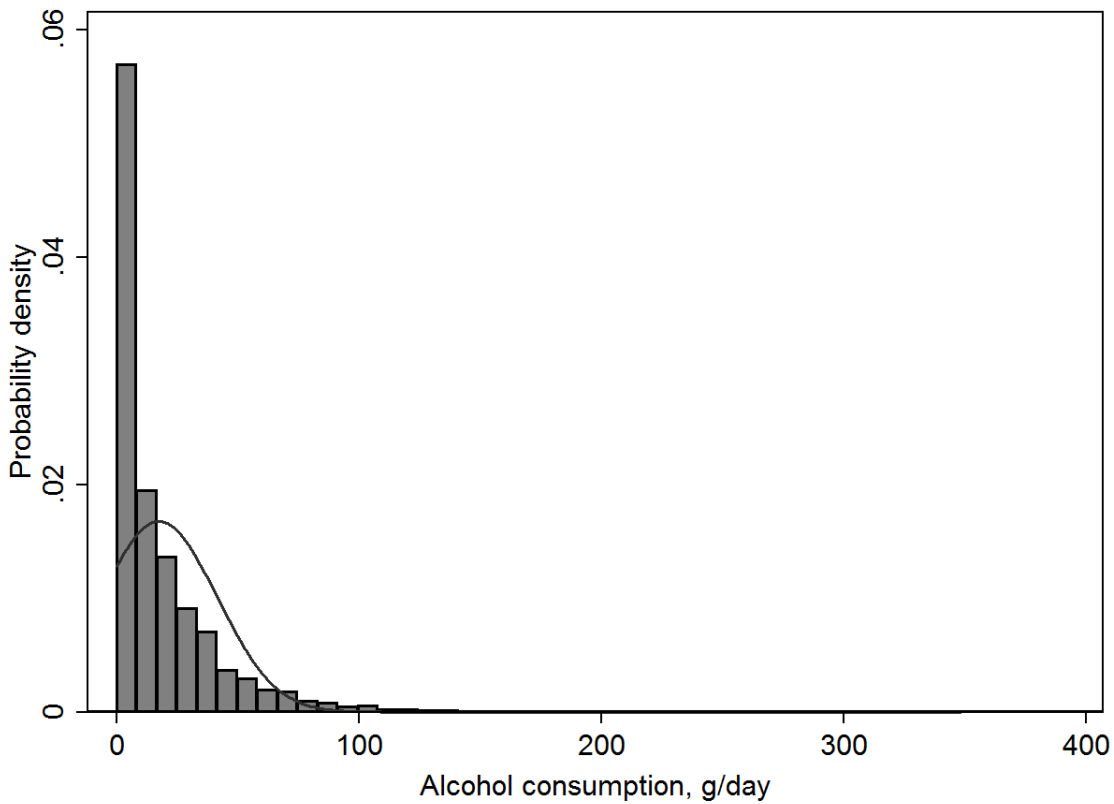


Figure 6.3: Histogram of alcohol consumption (in g/day) with normal curve

6.3.2 Model fits

A summary Table 6.3 showing the AIC scores of different unadjusted logistic regression models applied in the analysis of alcohol-hypertension relationship is presented below. The logistic regression models were obtained based on the four modelling approaches outlined in section 6.2.3.1 - 6.2.3.4.

Table 6.3: Summary statistics obtained after fitting various unadjusted logistic regression models

Modelling approaches	Degrees of freedom (excluding the intercept)	Akaike's Information Criterion (AIC)
Categorisation	3	20724
Linearisation	1	20739
Fractional polynomials (FPs):		
First order degree – FP1 (0.5)	2	20712
Second order degree – FP2 (1, 2)	4	20707
Restricted cubic splines (RCS):		
With 3 knots (RCS3)	2	20713
With 4 knots (RCS4)	3	20713

In Table 6.3, the largest AIC score of 20739 occurred when the relationship between alcohol consumption and the odds of having hypertension was linearised. Categorising the alcohol consumption data produced the model with the next largest AIC value of 20724.

For nonlinearity, the best fitting first-order degree fractional polynomial (FP1) had alcohol consumption term transformed with power 0.5 and the AIC score of 20712. The best fitting second-order degree fractional polynomial (FP2) had powers (1, 2) and the AIC value of 20707 (see Table 6.3). A likelihood ratio test performed between the FP2 (1, 2) and FP1 (0.5) functions showed an insignificant difference between the two models ($LRT = 2.12$, $p = 0.14$), hence the FP1 (0.5) function was favoured to characterise the adjusted alcohol-hypertension relationship in the UK Biobank.

For restricted cubic spline models (RCS), similar AIC scores were obtained when fitting three or four knots functions (see Table 6.3). A likelihood ratio test performed between the two RCS functions suggests an insignificant models difference ($LRT = 2.87$, $p = 0.10$). Therefore, a three knot function (RCS3) was chosen to describe the adjusted alcohol-hypertension relationship in the data.

Based on ‘minimally sufficient’ set of confounding variables identified from the DAG, the categorical, linear, FP1 (0.5) and RCS3 models were then adjusted for confounding.

Table 6.4 shows the unadjusted and adjusted odds ratio estimates together with their 95% confidence intervals (CIs) from the method of categorisation with four categories.

Table 6.4: The unadjusted and adjusted Odds ratios (ORs) of hypertension and their 95% confidence intervals obtained using the method of categorisation (CAT).

Alcohol consumption, g/day	No. of observations	No. of hypertension cases	Unadjusted OR (CAT Model)		P-trend	Adjusted OR (CAT Model)		P-trend
			Estimate	95% CI		Estimate	95% CI	
0	4,579	3,173	1.00	-	0.025	1.00	-	0.001
0-9.7	4,752	3,425	1.14	1.05 - 1.25		1.05	0.95 - 1.16	
9.7-25.1	4,570	3,455	1.37	1.25 - 1.51		1.22	1.10 - 1.36	
25.1+	4,484	3,632	1.89	1.71 - 2.08		1.71	1.52 - 1.93	

Covariates in the adjusted models include; age, anti-hypertension medication use, family history of hypertension, physical activity, ethnicity, sex, and smoking.

To compare the unadjusted and adjusted odds ratios from the linear, FP1 and RCS3 models, the category based estimates in Table 6.4 were replaced using logistic regression from these methods. In the logistic regression models fitted using the linear, FP1 and RCS3 approaches, the odds ratios were estimated at different values of alcohol consumption. As an illustration, Table 6.5 shows the odds of hypertension in the three models estimated at 0, 5, 17.5, 37.5, 62.5 and 87.5 g of alcohol consumption per day. In the data, less than 5% of the respondents reported alcohol consumption above 90 g/day.

Treating non-drinkers (as the reference), Table 6.4 and Table 6.5 suggested that the odds of hypertension were increasing for every unit of alcohol consumption. This was observed across the four methods of analyses. Furthermore, adjusting for confounders reduced the odds of hypertension; the unadjusted models had larger odds compared to adjusted estimates.

A display of the alcohol-hypertension association curves based on the four modelling approaches is shown in Figure 6.4. Additional curves trimmed ≤ 45 g/day of alcohol consumption were also made available in Figure 6.13 in the appendices. The choice to trim the functions ≤ 45 g/day of alcohol consumption was mainly for illustration and also to compare the alcohol-hypertension curves with functions in the range between 0-90 g/day. Large alcohol intakes (>90 g/day), may influence the functions to behave wildly at the upper tail.

Table 6.5: The unadjusted and adjusted odds ratios (ORs) of hypertension & their 95% confidence intervals obtained from the best fitting linearisation (LIN), fractional polynomials - first order degree (FP1) and the restricted cubic spline with 3 knots (RCS3) models. The odds of hypertension was modelled as a function of alcohol consumption, g/day.

Alcohol consumption, g/day	No. of observations	No. of hypertension cases	Ref. points	OR Estimates (LIN Based Model)		OR Estimates (FP1 Based Model)		OR Estimates (RCS3 Based Model)	
				Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
				OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
0	4,579	3,173	0 (ref)	1.00	1.00	1.00	1.00	1.00	1.00
0-9.7	4,752	3,425	5.0	1.05 (1.04-1.06)	1.05 (1.04-1.06)	1.18 (1.09-1.28)	1.08 (0.98-1.18)	1.13 (1.10-1.16)	1.10 (1.06-1.13)
9.7-25.1	4,570	3,455	17.5	1.20 (1.17-1.24)	1.17 (1.13-1.21)	1.42 (1.32-1.53)	1.30 (1.19-1.42)	1.44 (1.34-1.56)	1.33 (1.21-1.45)
25.1-49.9	2,885	2,313	37.5	1.48 (1.39-1.58)	1.41 (1.31-1.52)	1.72 (1.58-1.87)	1.56 (1.41-1.73)	1.74 (1.59-1.89)	1.58 (1.42-1.76)
49.9-74.9	1,021	830	62.5	1.92 (1.73-2.14)	1.77 (1.57-1.99)	2.05 (1.85-2.27)	1.85 (1.63-2.10)	1.95 (1.76-2.15)	1.83 (1.62-2.06)
74.9+	578	489	87.5	2.50 (2.15-2.89)	2.21 (1.87-2.62)	2.35 (2.08-2.66)	2.12 (1.83-2.46)	2.18 (1.89-2.52)	2.11 (1.79-2.49)

6.3.3 Alcohol-hypertension association curves

In Figure 6.4, the left panel represents the estimated functions from the unadjusted models trimmed ≤ 90 g/day of alcohol consumption. The adjusted models are presented in the right panel.

Compared to non-alcohol drinkers (0 g/day), the odds of hypertension increased in steps based on the three alcohol drinking groups assumed in the categorical models. For instance, in the adjusted model, 0-9.7 (lower), 9.7-25.1 (middle) and 25.1+ (upper) alcohol drinking categories was associated with OR=1.05 (CI=0.95-1.16), OR=1.22 (CI=1.10-1.36) and OR=1.71 (CI=1.52-1.93) respectively (see Figure 6.4).

In Figure 6.4, linearising alcohol consumption measures was associated with increasing odds of hypertension in both unadjusted and adjusted linear models. A constant change in the odds of hypertension was observed for every additional g of alcohol intake.

In Figure 6.4, the curves attained using FP1 models were visibly different. A large 'spike' at zero units of alcohol consumption was observed when fitting the adjusted model compared to the unadjusted model. However, there was no biological interpretation associated with this wild behaviour. The adjusted FP1 function was more meaningful when greater than zero units of alcohol was consumed. The alcohol-hypertension association curve in the adjusted FP1 model showed steeper and increasing slope on the odds of hypertension when small and moderate quantities of alcohol were consumed. For larger amounts of alcohol consumption, the slope in the FP function was shallow - showing a monotonically increasing OR trend (see Figure 6.4 for details). Furthermore, in the adjusted FP function, lower odds of hypertension (i.e. OR < 1) were observed when less than 2.2 g/day of alcohol was consumed. For example,

when 0.1 g of alcohol was consumed, the predicted odds of hypertension in the adjusted model was 0.89 (CI=0.80-1.00).

The RCS model with three knots (placed at the 10th, 50th, and 90th percentile) depicts a positive association with the odds of hypertension. Unlike in the FP1 model, the odds of hypertension obtained with RCS3 fit was never below one. The ORs in the RCS3 function was always positive with steep slopes observed when small and moderate units of alcohol were consumed. Just like in FP models, when large amounts of alcohol were consumed, the slope in the alcohol-hypertension function attained using the RCS was shallow but with a monotonically increasing OR trend. Overall, near similar nonlinear curves were observed when fitting the adjusted FP1 and RCS3 models for alcohol consumption exceeding 2.2 g/day. The graphs comparing the unadjusted and adjusted functions are provided in Figure 6.4 below.

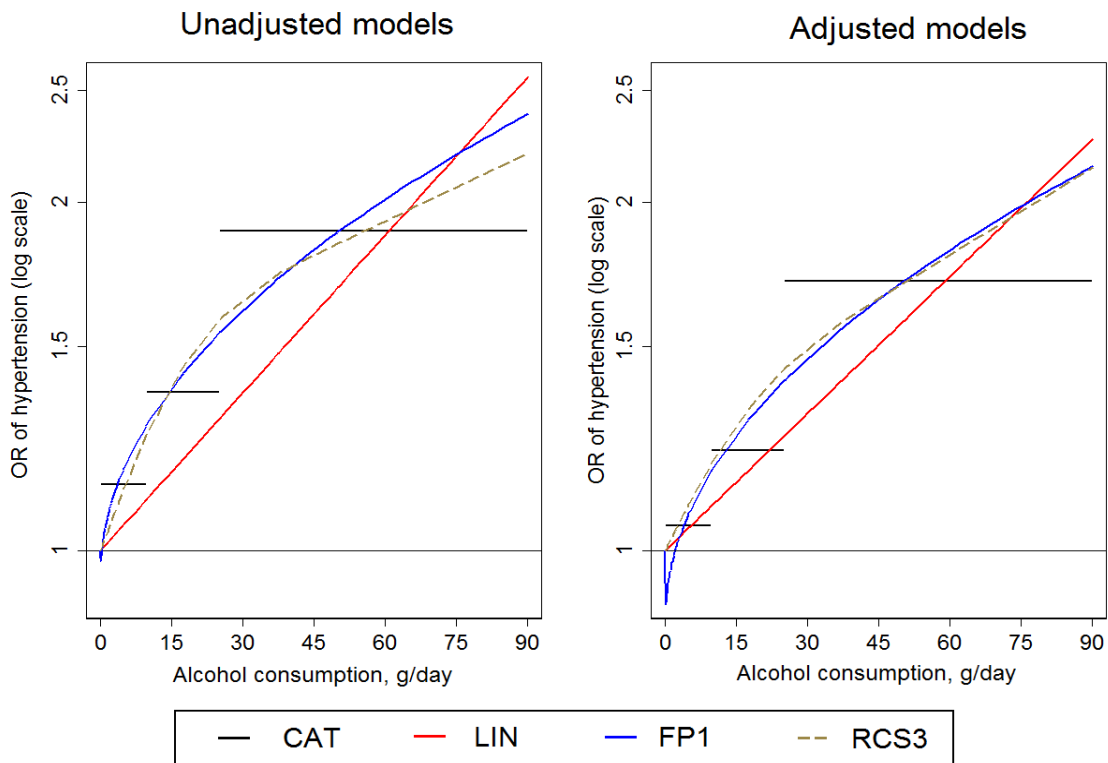


Figure 6.4: The unadjusted and adjusted odds of hypertension (on log scales) estimated using categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption (in g/day).

The graphs comparing the predicted odds of hypertension together with their 95% CIs using the adjusted models were presented in Figure 6.5. The presentation shows the predicted functions assuming 0-90 g/day of alcohol consumption. Similar graphs trimmed ≤ 45 g/day of alcohol consumption were also provided in Appendix F for comparison (see Figure 6.14).

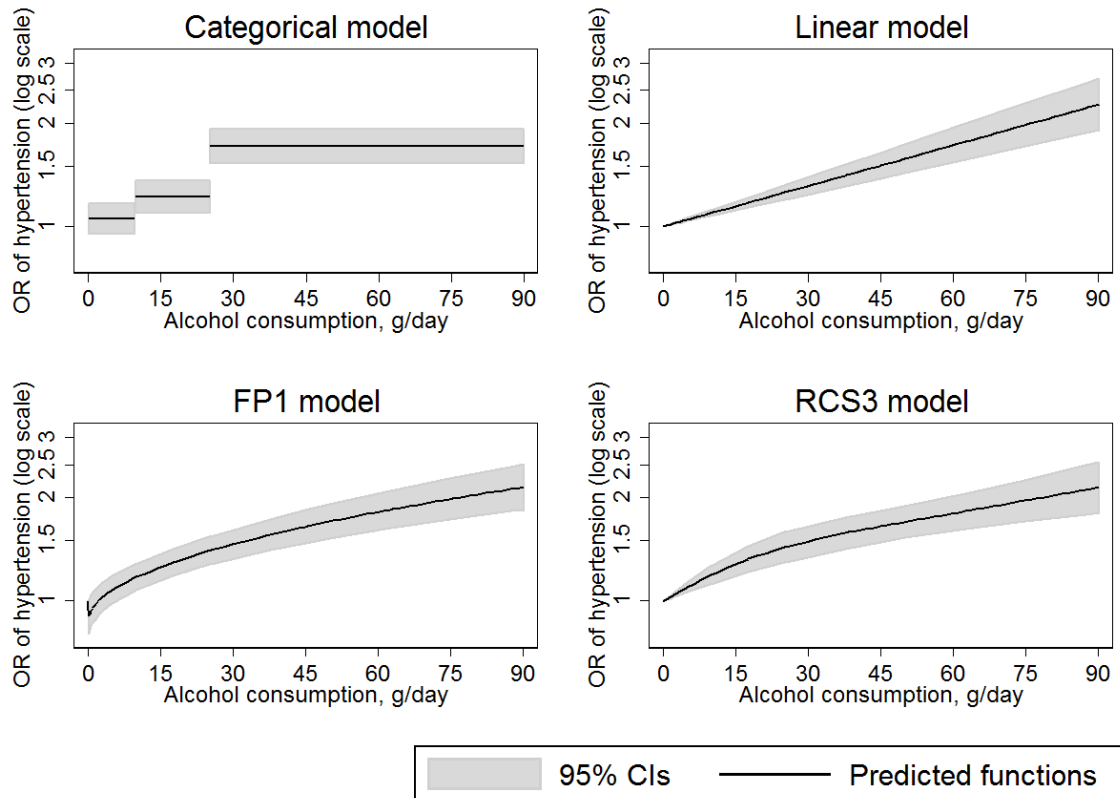


Figure 6.5: The adjusted odds of hypertension (on log scales) together with their 95% CIs estimated using the categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption, g/day.

In Figure 6.5, narrow CIs were observed at lower alcohol intakes when fitting the linear and RCS models. In contrast, the FP model had wider confidence intervals when smaller quantities of alcohol were consumed. However, in the three models, the width of predicted CIs increased with an additional unit of alcohol intake. Thus, the three models had wider CIs when larger amounts of alcohol were consumed. The CIs produced using categorical models increased in steps and the CIs width was not affected by variation in the data. Similar CI width was observed across the three groups assumed in the analysis.

6.3.3.1 Final adjusted models

Formally, the influence of confounders in the final models cannot be interpreted due to the application of a DAG. Nonetheless, Figure 6.11 & Figure 6.12 shows the odds ratios and the 95% CIs of several covariates adjusted for the alcohol-hypertension relationships (see Appendix F).

Overall, the FP approach produced a better fit compared to the other adjusted models. The adjusted FP function had the lowest AIC score of 17440 followed by the RCS function with the AIC score of 17443. The adjusted models that linearised and categorised the alcohol measures (exposure) had the AIC scores of 17449 and 17450 respectively.

6.3.4 Stratification analysis

The results showing whether antihypertensive medication use (binary variable) or age (continuous variable) modifies the alcohol-hypertension relationship are provided in the next sub-sections 6.3.4.1-6.3.4.4. The results obtained in the analysis using the four methods were different thus the outcomes were presented separately for comparison.

6.3.4.1 Effect modification of medication use and age using the categorical model

The differences in probabilities of hypertension between patients on anti-hypertension medication against non-users across the four categories of alcohol consumption (non-drinkers, lower, moderate and heavy drinkers) are shown in Figure 6.6 (a). Shown in the figure are also the 95% CIs of the predicted probabilities. In the two medication groups, the chances of having hypertension increased positively with alcohol drinking categories. When comparing the probabilities within the four drinking categories, the chance of having hypertension was high amongst patients on treatment

than non-medication users. However, the probability difference was narrow amongst heavy drinkers compared to moderate, light and non-alcohol drinkers (see Figure 6.6).

This suggests that antihypertensive medication use weakens the harm associated with alcohol consumption in heavy drinkers compared to patients who consume alcohol moderately, lightly or non-alcohol drinkers.

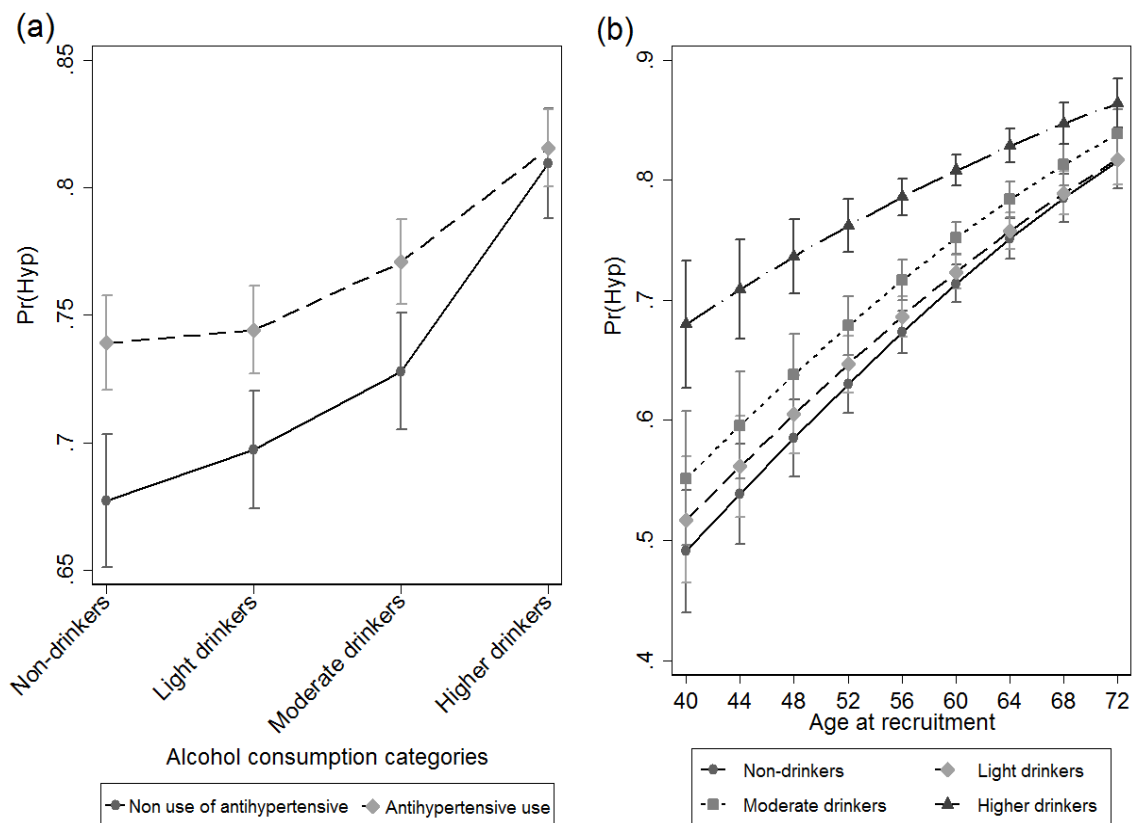


Figure 6.6 (a): The predicted probabilities of hypertension between patients on medication against non-medication users. The probability differences were computed based on the four categories of alcohol consumption. Figure 6.6 (b) shows the probabilities of hypertension between different categories of alcohol consumption against patient's age. The predicted probabilities were attained through the adjusted multivariable categorical model.

Figure 6.6 (b) shows the difference in probabilities of hypertension in the four groups of alcohol consumption against age predicted using the categorisation model. In all the four categories of alcohol consumption, the chances of having hypertension increased with age. Heavy alcohol drinkers had greater probabilities of being hypertensive compared to other drinking categories across the age range. In comparison, non-alcohol drinkers retained the lowest probabilities across the age range. Light alcohol drinkers followed with probabilities slightly higher than non-alcoholic patients but not greater than moderate drinkers (see Figure 6.6 (b)). In Figure 6.6 (b), when patients were young, the predicted probabilities of hypertension were small with large differences observed between the four drinking categories. In contrast, older patients had greater chances of hypertension with narrow differences between the alcohol drinking categories. For example, the predicted probabilities of hypertension in 40-year-old patients were recorded as 0.49 (CI=0.44-0.54), 0.52 (CI=0.46-0.57), 0.55 (CI=0.50-0.61) and 0.68 (CI=0.63-0.73) amongst non-alcohol drinkers, light, moderate and heavy alcohol drinkers respectively. Additionally, the predicted probabilities of hypertension in 72-year-old patients were large with narrow differences amongst the four alcohol drinking groups compared to the 40-year-olds. The probabilities of hypertension in 72-year-old patients were predicted as 0.82 (CI=0.80-0.84), 0.82 (CI=0.80-0.84), 0.84 (CI=0.82-0.86) and 0.86 (CI=0.84-0.88) amongst non-alcohol drinkers, light, moderate and heavy alcohol drinkers are respectively.

The narrow probability differences observed amongst older patients suggest the harm associated with alcohol drinking was worse in young patients.

6.3.4.2 Effect modification of medication use and age using the linear model

The predicted probabilities of hypertension based on the linear models for patients on anti-hypertension medication and non-users are presented in Figure 6.7 (a). Based on the various amount of alcohol consumption (g/day), the difference in

probabilities between the two groups (medication user's vs non-users) was also captured and presented for evaluation using Figure 6.7 (b). Overall, the graph showed a significant alcohol-hypertension association with the presence of an *alcohol X medication use* interaction term. The slopes of predicted probabilities of hypertension between patients on medication and non-users were different. However, similar predicted probability and zero probability difference of hypertension between patients on medication and non-medication occurred when 73 g of alcohol was consumed (see Figure 6.7 (a)-(b)). Below 73 g of alcohol consumption, the predicted probability of hypertension amongst patients on antihypertension medication was large than in non-medication users. When patients consumed more than 73 g of alcohol, the predicted probabilities of hypertension was slightly high in non-medication users compared to those using anti-hypertensives (see Figure 6.7 (a)).

From the results, antihypertensive drug's efficacy was notable at every additional unit of alcohol intake. The slopes functions amongst patients on medication and non-medication users narrows with every unit of alcohol consumption. At 73 g of alcohol drinking, the difference of probabilities between patients on medication and non-medication users was zero. Above 73 g, non-medication users had greater chances of hypertension compared to antihypertensive medication users.

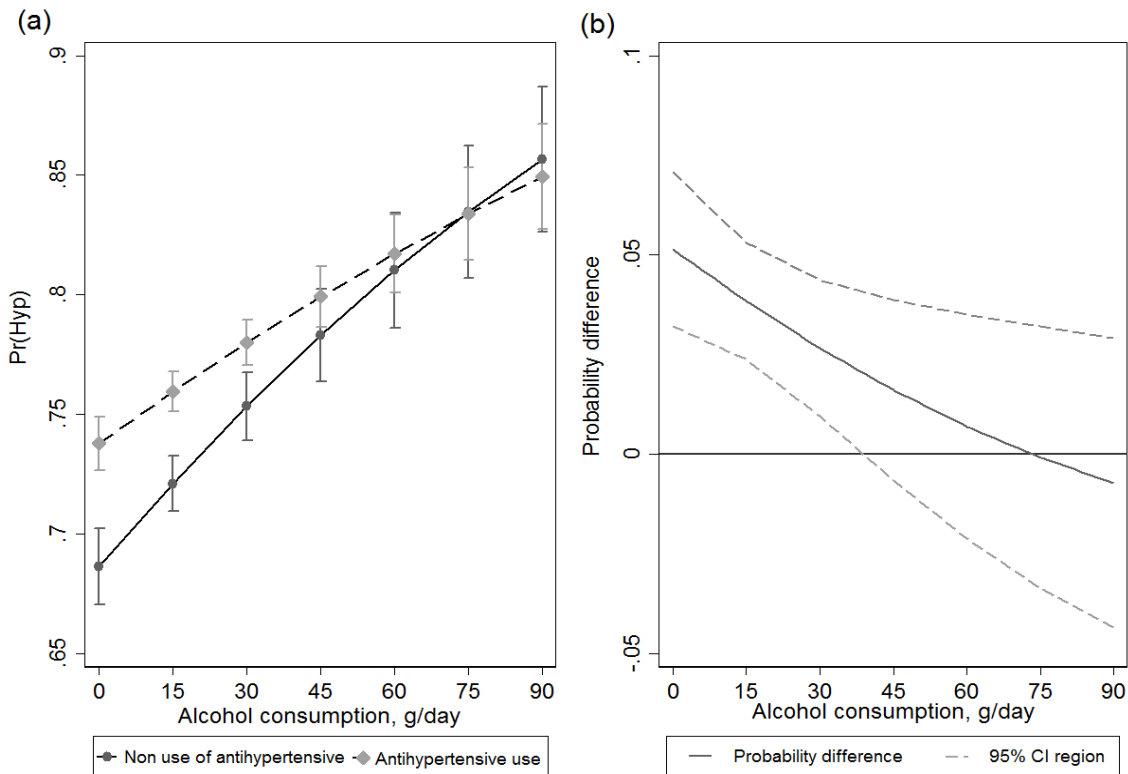


Figure 6.7 (a): The predicted probabilities of hypertension between patients on anti-hypertension against non-users across the different amount of alcohol consumption (g/day). Figure 6.6 (b) shows the difference in probabilities between anti-hypertension medication users and non-users at different quantities of alcohol consumption (g/day). The predicted probabilities were attained through the adjusted multivariable linear model.

Figure 6.8 presents the effects of alcohol consumption (g/day) on the predicted probability of hypertension according to different patient's age. It was also evident from the graph that the association between age and the probabilities of hypertension varied with the different amount of alcohol consumption (g/day). Moreover, there was evidence of *alcohol X age* interaction term in the linear model. This was suggested by the curvature of the contour lines in Figure 6.8. The omission of the interaction term produces straight contour lines.

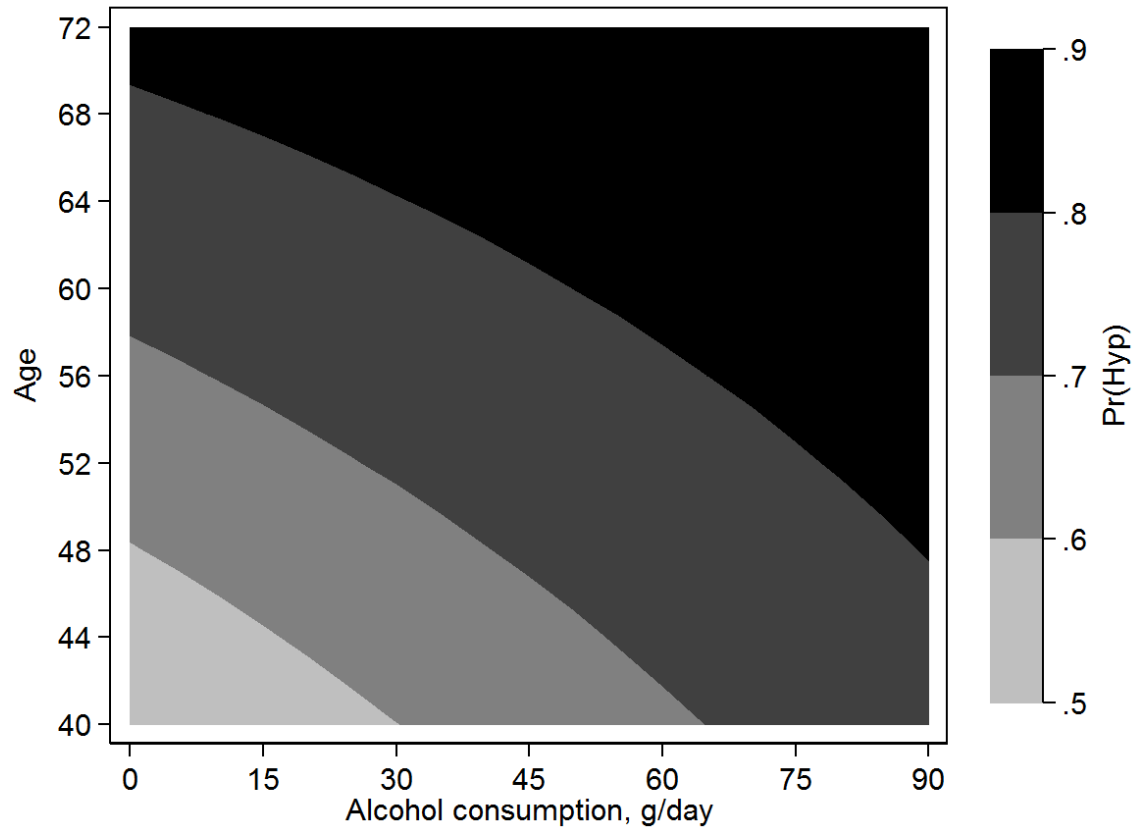


Figure 6.8: The predicted probabilities of hypertension across different levels of alcohol consumption (g/day) and age from the adjusted multivariable linear model

In Figure 6.8, a 44 year old person drinking 15 g of alcohol per day had 60% chance of hypertension. When the patient's age was held constant, the probability of hypertension increased with alcohol consumption. For example, when an individual's drinking capacity was increased to 60 g/day amongst the 44-year-olds, the predicted probability of hypertension also increased to 80%.

Also, when the amount of alcohol consumption (g/day) was held constant, the predicted probability of having hypertension increased with age. For instance, 48-year-old people consuming 30 g of alcohol per day had 70% chance of hypertension. For the same units of alcohol intake – 30 g/day, an older person aged between 68 or 72 years had 90% chance of hypertension.

6.3.4.3 Effect modification of medication use based on fractional polynomial and restricted cubic spline models.

The probabilities of hypertension between patients on antihypertension medication and non-users estimated from the FPI and RCS models are presented in Figure 6.9. The probabilities of hypertension were large amongst patients on antihypertensive medication compared to non-medication users and increased with every unit intake of alcohol. The slopes of predicted probability functions were the same when fitting the FP functions (see Figure 6.9 (column (a))). Thus, the antihypertensive drug's efficacy was undetectable across the range of alcohol consumption levels.

The RCS fit produced wider probability differences amongst patients consuming < 30 g of alcohol per day. For alcohol consumption >30 g/day, the slopes of predicted probability functions between antihypertensive medication users and non-users were identical and the probability difference was narrow (see Figure 6.9 (column (b))). The narrow probability difference amongst patients consuming > 30 g of alcohol per day suggests the efficacy of antihypertensive drugs in this region

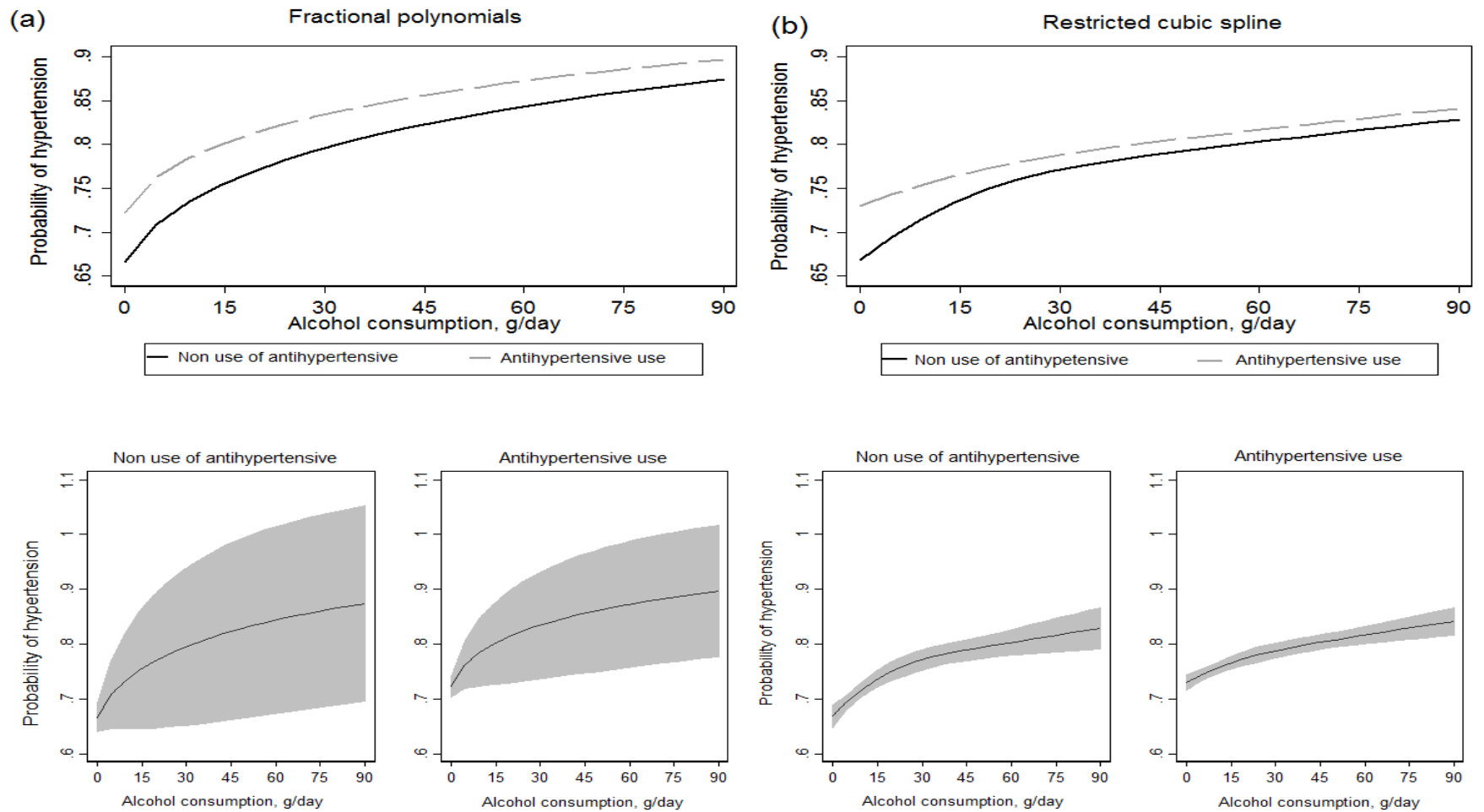


Figure 6.9: shows the difference in probabilities of hypertension between patients on antihypertensive medication and non-users at different levels of alcohol consumption (g/day) from the FP1 model with a power transformation of 0.5 and RCS with 3 knots.

6.3.4.4 Effect modification of age based on fractional polynomial and restricted cubic spline models

The results showing whether age modifies the alcohol-hypertension relationship from the FP and RCS models are provided in Figure 6.10. In the two models, the chance of hypertension increased with age (holding alcohol consumption units constant). Similarly, when age was held constant, the chance of having hypertension increased with each unit of alcohol consumption. The differences in predicted probabilities of hypertension across different ages were wide at the lower tail of alcohol consumption scale and narrow at the upper tail when fitting the FP model. Although a similar pattern was observed when fitting the RCS model, the latter had slightly wider probability differences at the upper tail than when fitting the FP model. This suggested the presence of a significant *alcohol X age* interaction term in the FP model which was not strong in the RCS model. For any two age groups, the difference in mean probabilities of hypertension obtained using the RCS function reduced firmly with the consumption of more alcohol (see Figure 6.10).

Moreover, for any age held constant, the FP model predicted large probabilities of hypertension compared to fitting the RCS function. At the tails of alcohol consumption (x-scale), the FP model produced functions with steep and increasing slope compared to narrow slopes obtained in the RCS fit. The differences in the two model fits are shown in Figure 6.10.

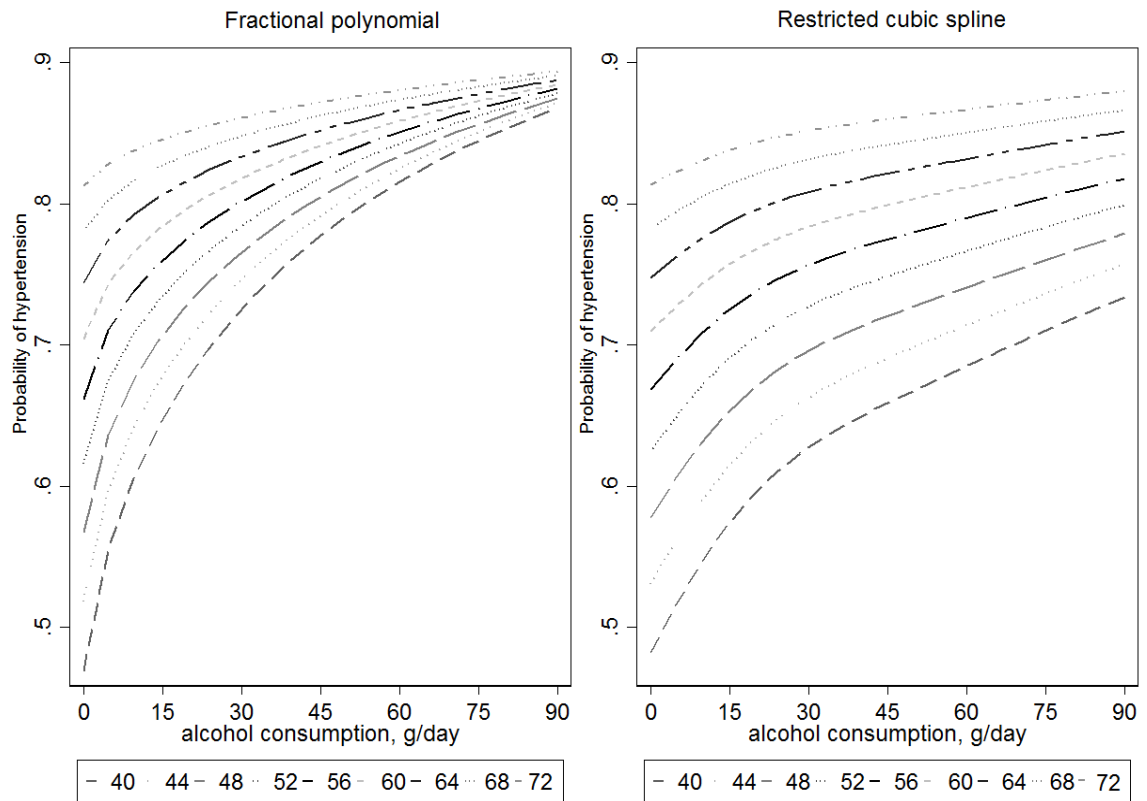


Figure 6.10: The predicted probabilities of hypertension across different levels of alcohol consumption (g/day) and age from the adjusted FP and RCS models

The results in Figure 6.10 suggest that the harm associated with alcohol consumption was worse in young adult population than in older patients. The latter is supported by steep slopes observed in young aged functions at the lower tails of alcohol consumption and narrow probability differences at the upper tails obtained when comparing the older and the younger adult population groups. A similar pattern was observed when fitting the FP and RCS models; however, the *alcohol X age* interaction in the RCS model was weaker.

6.3.5 Sensitivity analysis

Excluding ‘possible’ type 2 diabetes patients in the analysis did not substantially change the odds of hypertension in the fitted models. For any unit of alcohol consumption, the categorical, linear, FP and RCS models slightly overestimate the odds of hypertension in ‘probable’ type 2 diabetes patients. However, similar alcohol-

hypertension shapes were obtained in patients with or without ‘possible’ type 2 diabetes. As an illustration, the estimated odds of hypertension across different units of alcohol consumption in patients excluding ‘possible’ cases are given in Appendix F (see Table 6.6-Table 6.7). Figure 6.15 shows alcohol-hypertension associations obtained across the four methods of analysis in the dataset excluding ‘possible’ diabetes cases. The sensitivity analysis involving interactions were not performed because of insignificant changes observed between the main effects models.

6.4 Discussion

This section discusses key findings, challenges and limitations, strength and opportunities, novelty and future studies from this chapter. Summary of findings and how they compare with results from other studies are provided in Section 6.4.1. The discussion of challenges and limitations focussing on the data and issues of this chapter are provided in section 6.4.2. Section 6.4.3 focuses on the strength and opportunities whilst section 6.4.4 - 6.4.5 highlight the novelty and possible future studies.

6.4.1 Summary of key findings

The overall objective of this study was to investigate the association between alcohol consumption and hypertension in type 2 diabetes patients using the UK Biobank. The prevalence of hypertension amongst patients with type 2 diabetes was observed at 70% in the UK Biobank. This proportion was higher than the figure reported in the general UK population. The Health Survey for England (HSE), estimate hypertension at 28% of the general UK adult population (National Statistics, 2016). In the present study, tighter systolic BP cut-off point (≥ 130 mmHg) was used in defining patients with hypertension amongst type 2 diabetes patients. In the HSE, the systolic BP cut-off point defining hypertension in the general population was less tight (≥ 140 mmHg). Hence, the high proportion of hypertension subjects observed in the present

study. In other words, the proportions of hypertension in diabetes cohort could vary depending on the definition of tighter BP control in the study. For example, the prevalence of hypertension (assuming SBP ≥ 160 mmHg) in diabetic adults was estimated at 51% using a population based analysis from the HSE conducted in 1991 to 1994 (Colhoun et al., 1999). In the latter, the prevalence of hypertension could be higher when applying the recent guidelines suggesting a stringent reduction of SBP in diabetes patients. The recent Joint British Societies' consensus recommendations for the prevention of cardiovascular diseases (JBS3) and the WHO guidelines suggest SBP <130 mm Hg amongst patients with diabetes (World Health Organization, 2006, British Cardiovascular Society et al., 2014). Compared to patients with both hypertension and type 2 diabetes in the UK Biobank, the proportion of non-hypertensive individuals with type 2 diabetes was low (~25%). Non-hypertensive patients had an average SBP of 121 mmHg compared to 148 mmHg observed amongst those with hypertension and type 2 diabetes. These findings characterising the prevalence of hypertension (or mean SBP) in patients with type 2 diabetes should form the basis for upcoming studies. The UK Biobank has large, generalizable and contemporary data sufficient to investigate hypertension in patients with type 2 diabetes (Allen et al., 2012).

The relationship between alcohol consumption and the odds of hypertension in type 2 diabetes patients suggested the presence of nonlinearity when fitting the FP and RCS models. The adjusted ORs in the two models were different for alcohol consumption below 2.2 g/day and comparable when the consumption exceeds 2.2 g/day. For example, a large 'spike' at zero units of alcohol consumption was observed when fitting the FP function. In addition, the FP function had ORs below 1 for alcohol consumption between 0.1 and 2.2 g/day. In this range, the ORs of hypertension were increasing at every unit of alcohol intake. In contrast, the 'spike' at zero was not observed when applying RCS. Furthermore, the ORs in the RCS model was always

positive across the alcohol consumption range. According to Royston and Sauerbrei (Royston and Sauerbrei, 2008), the ‘spike’ at zero observed under the FP models has no biological interpretation. Royston and Sauerbrei (2008) recommend researchers to ignore the ‘spike’ behaviour in their models and begin interpreting the FP curves where alcohol consumption exceed zero (i.e. $X > 0$). Alternatively, the ‘spike’ at zero in FP models can be removed by shifting the origin of alcohol intake units by adding a small constant δ to each observation. However, this transformation method has been criticised because the resulting FP models could be influenced by the choice of the constant δ (Ambler and Royston, 2001). The other approaches of dealing with ‘spike’ at zero covariates such as alcohol consumption include a two stage FP procedure suggested by Jenkner and colleagues (Jenkner et al., 2016). The two stage FP procedure recommend including a binary indicator variable (non-alcohol drinkers vs alcohol drinkers) in the final FP models to eliminate the spike at zero. The disadvantage of the two stage FP procedure is that it makes a strong assumption on the relationship between alcohol consumption and the odds of hypertension. The inclusion of a binary indicator in the final FP model will suggest a relationship separating the exposure at zero and the continuous part.

At the alcohol consumption range where the FP and RCS functions are comparable, the OR curves increased positively with steeper slopes observed between 2.2 and 17.5 g/day of alcohol intake. For alcohol consumption exceeding 17.5 g/day, the slopes in both the FP and RCS models were shallow but with monotonically increasing OR trends. The functions attained with the traditional methods showed non-negative monotonic relationships that increased in step (when categorising alcohol consumption measures) and linearly (when treating alcohol consumption measures as a linear variable). In comparison, the four methods of analysis used in this study were in agreement suggesting the presence of monotonically increasing relationships between

alcohol consumption and the odds of hypertension. However, the methods of categorisation and linearisation prohibit the detection of nonlinearity on the data. In the four methods, no optimal amount of alcohol consumption was linked with reduced odds of hypertension amongst patients with type 2 diabetes. These results contradict findings in the general population studies reporting the J or U association shapes suggesting the optimal risk of hypertension amongst the light or moderate alcohol drinkers (Gillman et al., 1995, Moreira et al., 1998). For example, an optimal odd of hypertension was reported amongst light drinkers in the general study performed by Moreira and colleagues (1998). After adjustment for sex, age, education, BMI and use of antihypertensive drugs; light drinkers (0-30 g/day) had OR=0.82 (CI=0.51-1.30), moderate drinkers (30-60 g/day) had OR=2.39 (CI=1.11-5.16) and heavy drinkers (≥ 60 g/day) had OR=2.02 (CI=0.88-4.64) of hypertension compared to non-alcohol drinkers (reference category). Thus, the ever rising or positive odds of hypertension amongst the light alcohol drinkers in the present study could be a suggestion that alcohol worsens the incidence of hypertension in patients with diabetes than in the general (or non-diabetics) population. Light-moderate drinking and occasional heavy drinking were both associated with increased risk of hypertension in patients with type 2 diabetes (Saremi et al., 2004).

In the analysis stratified by antihypertensive medication use (medication user's vs non-users); the chances of having hypertension were high amongst patients using antihypertensive medication compared to non-medication users. The latter was observed across the four methods of analysis. These results were consistent with findings in the Japanese male population (without diabetes) showing higher percentages of alcohol drinkers in the group receiving antihypertensive therapy for hypertension than in the group not receiving antihypertensive(s) (Wakabayashi, 2010). Under the method of categorisation and linearisation, the slopes showing the predicted probabilities of

hypertension between patients on medication and non-users were different suggesting significant alcohol consumption influence and the existence of *alcohol X medication use* interaction terms. The difference in probabilities was narrow towards heavy alcohol drinkers compared to light, moderate and non-alcohol drinkers. These results suggest antihypertensive efficacy amongst heavy drinkers. The harm associated with alcohol consumption was visibly weak amongst heavy drinkers compared to light and moderate drinkers. The evidence of antihypertensive medication efficacy amongst hypertensive patients has also been reported in a randomised placebo-controlled trial (Maheswaran et al., 1990) and a cross-sectional study (Wakabayashi, 2010) performed to examine the influence of alcohol consumption. In the trial, the systolic BP amongst hypertensive patients participants given metoprolol drugs dropped significantly compared to those who were given placebo (Maheswaran et al., 1990). In a cross-sectional performed by Wakabayashi (2010), the systolic BP was significantly high in heavy (22-44 g/day) and very heavy (≥ 44 g/day) drinkers compared to non-drinkers when subjects were not receiving therapy for hypertension. In subjects receiving therapy, no significant difference was obtained when comparing systolic BP for light (<22 g/day), heavy and very heavy drinkers to non-drinkers. The FP and RCS analyses revealed the absence of the *alcohol X medication use* interaction terms. The comparison between the slopes of probability functions of hypertension amongst patients using antihypertensive medication versus non-medication users was the same when fitting the FP model. This means that the efficacy of antihypertensive medication was not detectable in the FP model. In contrast, the RCS model suggested the efficacy of antihypertensive amongst patients consuming more than 30 g/day of alcohol. When comparing the RCS functions predicting the occurrence of hypertension amongst antihypertensive medication users and non-users, identical slopes with narrow probability differences were observed at alcohol consumption exceeding 30 g/day. When alcohol consumption was below 30

g/day, the comparison between the RCS functions of medication users versus non-users produced wider probability differences suggesting inefficacy of antihypertensive drugs. To my knowledge, this is the first study investigating the application of the four methods of analysis and the efficacy of antihypertensive use in alcohol-hypertension relationships focusing on diabetes patients.

In the analysis stratified by age; the probabilities of hypertension were high amongst aging patients. This was observed across the four methods of analysis. Heavy older alcohol drinkers had a greater chance of having hypertension compared to young non-alcohol drinkers or young individuals drinking lightly or moderately. The differences in predicted probabilities of hypertension across different ages were wider at the lower tails of alcohol consumption than at the upper tails. This implies that younger adults were severely affected by the harm associated with alcohol consumption compared to older patients. The *alcohol X age* interaction was also observed in the four models however, it was not strong when fitting the RCS function. Existing studies investigated the alcohol-hypertension relationship in general population (Okubo et al., 2014). Okubo and colleagues (2014) found consistent linear associations between alcohol intake and the risk of hypertension in middle-aged (40-59 years) and older (60-79 years) individuals suggesting the absence of the *alcohol X age* interaction term in the data. To my knowledge, this is the first largest study showing that the relationship between alcohol consumption and hypertension in a diabetic population vary by age.

6.4.2 Challenges and limitations

This section discusses the potential challenges and limitations in this chapter.

6.4.2.1 Data

There are issues of inconsistent reporting in alcohol studies linked with possible response biases. Potential sources of the response biases in the UK Biobank includes (1) under-reporting particularly amongst heavy drinkers, (2) inability to remember the past

alcohol intake activities, and (3) inability to comprehend and estimate alcohol content accurately according to the assumed standard drinks and sizes. These sources of bias are discussed in details below.

Under-reporting amongst heavy drinkers: Heavy drinkers are likely to under-report the amount of alcohol they drink, perhaps for reasons of social desirability. If this occurs, the heavy drinkers would be misclassified and analysed as moderate, light or non-drinkers. This would then affect the predicted functions by overestimating probabilities and changing their slopes at the lower tails of the alcohol consumption scale. Hence, it is possible that the curvature observed at the lower tails of the alcohol distribution when applying the FP and RCS models could be artefacts caused by under-reporting amongst heavy drinkers (see Figure 6.4). Unfortunately, it was not possible to check this since the actual misclassification error was not known.

Inability to remember the past alcohol intake activities: Forgetting is another potential source of response bias in self-reported alcohol consumption data. Evidence suggest higher estimates in studies that inquire about recent alcohol intake activities (e.g. in the last 24-hours) than those estimating the consumption over a longer period of time (in weeks or months) (Lemmens et al., 1992, Stockwell et al., 2004). Based on the latter, occasional drinkers were likely to underestimate their alcohol intake in the present study. This could affect the predicted probabilities of hypertension in the four methods - lowering the estimates across the alcohol consumption range. The actual errors associated with forgetting and its influence in the present study was also difficult to quantify.

Inability to comprehend and estimate alcohol content accurately: The UK Biobank touchscreen questionnaire provided pictures of different drinks and serving sizes to standardise alcohol intake response's and increase the reliability of answers

provided by participants. This approach assumes the serving sizes and contents of drinks in licenced premises and home measures are the same. These assumptions have the potential to underestimate the actual contents and amounts of alcohol consumed by participants. The large variation of alcohol contents and drinking sizes could affect the slopes and estimates in the predicted models. However, the standardized alcohol measurements in the UK Biobank were in agreement with those in the UK alcohol guidelines (House of Commons Science and Technology Committee, 2012). The UK alcohol guidelines are revised and updated frequently to capture changing patterns of beverage preferences and availability. Hence, the predicted estimates in the present study should have minimum bias from the responses given by participants.

6.4.2.2 Methods and application

The discussions on methodological challenges and limitations in this chapter are provided below.

The clinical definition of hypertension amongst patients with diabetes includes individuals on blood pressure lowering medication and/or systolic BP \geq 130 mm Hg (Judd et al., 2011). However, in this chapter, the operative definition of hypertension excluded patients on antihypertensive medication, focusing only on those with systolic BP \geq 130 mm Hg. This has the potential to affect the results in other similar studies since the prevalence rates of hypertension are likely to be underestimated in such investigations. Nonetheless, it was not statistically feasible to consider antihypertensive medication use as part of the definition for patients with hypertension in the present study. This is because antihypertensive medication use (yes/no) was treated as a confounding variable in the analysis.

In the analyses, continuous confounders were untransformed to establish the relationship between alcohol consumption and hypertension. Untransformed

confounders simplify the analysis and preserve the scales of continuous confounders during statistical modelling minimising the bias on the estimates (Brenner and Blettner, 1997, Groenwold et al., 2013). A simulation study assessing various covariate-risk associations, Brenner and Blettner (1997) found suboptimal confounding when covariates were continuous (either linear or nonlinear) and biased OR estimate or residual confounding when continuous covariates were analysed as categorical variables. Chen and colleagues (Chen et al., 2007) also obtained satisfactory odds ratios (ORs) when age (continuous confounder) was treated as a linear or nonlinear variable in the analysis and biased ORs when age was dichotomised. Based on the findings of these studies, biased OR estimates was less likely in functions treating both the exposure and confounders as continuous variables in the analyses. Models dichotomising the exposure and analysing confounders as continuous variables had a larger bias on their estimates. This has implications for the present study; the estimated ORs of hypertension attained with the method of categorisation (see Table 6.4) could be more biased than the estimates in Table 6.5 (obtained when both the exposures and confounders were continuous variables).

Knots selection procedure was another aspect considered when assessing the alcohol-hypertension relationships using the RCS functions. The RCS functions are known to be sensitive to the number of knots and their placement (Durrleman and Simon, 1989, Desquilbet and Mariotti, 2010). In this study, the RCS function suitable for the alcohol-hypertension relationship amongst patients with type 2 diabetes had 3 knots. The RCS model with 4 knots was less efficient with more parameter estimates and unstable function. These findings are in agreement with the suggestion made by Durrleman and Simon (1989) that fewer knot RCS models were likely to provide adequate fits for most phenomena observed in medical studies.

Finally, the incomplete data in the predictors were omitted during statistical modelling. The adjusted functions were attained using $n=15,967$ complete records, which represent about 67% of the whole dataset ($n=23,842$). The remaining $n=7,875$ observations representing 33% of the dataset, had one or more missing values and were completely ignored in the analysis. This implies that the predicted alcohol-hypertension functions could be less precise due to reduced sample size or loss of power on the data. Also, the confidence intervals of the predicted functions could be incorrect due to the potential bias in standard errors (White, 2015). An attractive approach for handling missing values of predictors in regression modelling is the multiple imputation (MI) method (Rubin, 2004). The MI method is common in epidemiology and clinical studies. In its application, M complete datasets are generated such that they correctly reflect the distribution of missing data given the observed values. The benefits of the MI methods is that it restores the natural variability of the missing values and incorporates uncertainty in the data for valid statistical inference (Kang, 2013). The main challenge with this imputation method is that it is computationally expensive and based on assumptions that may lead to errors when generating the data (Humphries, 2013). In addition, more complications could occur in studies investigating nonlinearity and interaction terms. Available standard software's implementing the MI procedures are limited - imputation datasets maybe based on assumptions that are incomparable with the 'true' models (Bartlett et al., 2015). Thus, the recent studies (Bartlett et al., 2015, White, 2015) recommend not choosing the MI method when investigating nonlinear associations and interactions in the data. Hence, the MI procedure was not implemented in this study.

6.4.3 Strengths and opportunities

Apart from the weaknesses noted above, there are particular strengths worth noting in this work. Firstly, it is important to recognise that UK Biobank is a large

cohort study with over 500, 000 participants. Therefore, a large sample was used in establishing and analysing the alcohol-hypertension relationships. Secondly, to control for confounding bias, the DAG technique was applied in the modelling process. The DAG approach reduce the potential confounding bias in alcohol-hypertension models through the identification of minimal sufficient sets of variables to adjust for in the analysis. In contrast to the DAG based approach, stepwise selection procedures are common for identifying confounders however they have been discouraged on many grounds particularly because variable selection depends on p-values alone (Greenland and Neutra, 1980, Núñez et al., 2011). Thirdly, the research covers a variety of methods of analysis to study the alcohol-hypertension relationships. The two popular methods of categorisation and linearisation were compared with nonlinear approaches including fractional polynomials and restricted cubic splines. The practice of categorisation and linearising the alcohol intake measures restrict the shapes of the alcohol-hypertension associations to step and linear functions ignoring nonlinearity in the data. Fourth, the research employs simpler graphical methods of analysis for visualization and interpretation of interactions (or modification) in the data. The graphical methods were illustrated using the categorical, linear and nonlinear alcohol-hypertension relationships attained in this chapter. Finally, a sensitivity analysis was performed to verify the consistency of the reported alcohol-hypertension relationships.

6.4.4 Novelty

A literature search on Google Scholar, PubMed and Web of Science using the following keywords; UK Biobank, type 2 diabetes, hypertension, alcohol, associations or relationships suggest this is the first study investigating the alcohol-hypertension relationships and interactions amongst type 2 diabetes patients in the UK Biobank. Therefore, the findings reported in this chapter are novel and have potential implications for public health and medical practice.

6.4.5 Future studies

To clarify the causal relationship between alcohol consumption and hypertension in patients with type 2 diabetes, further research is needed using longitudinal data. The present study investigated the relationship using self-reported baseline data. The cause and effect pathways will be more reliable in a longitudinal study than the data obtained at a single point in time.

Consideration for another research is also suggested to develop a multiple imputation model for missing alcohol-hypertension data in the UK Biobank and to provide guidelines for its implementation. The proposed imputation needs to demonstrate the missing data mechanism, correct specification of the data imputation model and the implementation process.

Finally, since the alcohol data is prone to measurement/misclassification errors, further investigations are also required to (1) establish the effects of these errors on the alcohol-hypertension relationships and (2) assess the methods suitable for correcting or reversing the bias in the data. This could be achieved through a simulation study demonstrating the effects of different random errors in the data and correction procedures.

6.5 Conclusions

This study identified a greater proportion of type 2 diabetes patients with poor blood pressure control. Furthermore, alcohol consumption was associated with the increasing odds of hypertension in the data. The four methods of analysis applied in the study suggested monotonically increasing functions in the data. These results suggest the need for aggressive strategies to manage and control hypertension in type 2 diabetes patients. Additional findings also revealed that antihypertensive medication use and patient's age modify the odds of hypertension in alcohol drinkers across the four

methods of analysis. In the four models, the harm associated with alcohol drinking was worse in young patients than in older patients. Furthermore, the efficacy of antihypertensive use was observed amongst heavy alcohol drinkers when comparing medication users against non-users. These findings are in support of clinical guidelines or strategies that explain the risk of hypertension in patients with type 2 diabetes taking into consideration (1) the efficacy of antihypertension medication use and (2) the severity of alcohol drinking across different ages.

This study also demonstrated the existence of nonlinearity in real data. The functions characterising the alcohol-hypertension showed the presence of nonlinearity when fitting the FP and RCS models. This is an important finding that should encourage researchers working in the same area to consider nonlinearity in their studies. Characterising nonlinear exposure–outcome relationships accurately is very important in epidemiology because such relationship studies inform policies that influence individuals' health outcomes.

Chapter 7

Discussion

This discussion chapter synthesises the work carried out in this PhD.

The chapter is divided into sub-headings listed below:

- 1) A brief introduction
- 2) Synthesis and interpretations of key findings
- 3) Research implications and contributions
- 4) Challenges and limitations
- 5) Strengths and opportunities
- 6) Recommendations for future research
- 7) A list of publications arising from this PhD thesis
- 8) A conclusion

7.1 Introduction

In medical studies, it is important to be able to characterise the shapes of the predictor-outcome relationships accurately. This is because such relationship studies have the potential to inform health policies that in turn impact on individuals' health outcomes. Evidence in published surveys suggests the common practice of categorisation of continuous variables when reporting predictor-outcome relationships in medicine (Pocock et al., 2004, Turner et al., 2010). The practice of categorisation does not make use of within category information. Thus, the final models may be inappropriate due to loss of information. The alternative methods such as fractional

polynomials (FPs) and restricted cubic spline (RCS) approaches are available for handling continuous data during statistical modelling but they not widely used.

To encourage and promote the use of FPs and RCS models amongst medical researchers with little background in statistics for application of these methods; the first objective of this research was to conduct a new survey demonstrating the current extent of categorised continuous variables in observational studies. Taking lessons from the latter, novel simulation studies (based on causal and predictive models) and an application study based on real dataset then followed. The purpose of these studies was to compare the performances and properties of the linear, FPs and RCS regression approaches (as the alternatives) against the practice of categorising continuous data. A simple conceptual framework linking these chapters is provided in Figure 7.1 below.

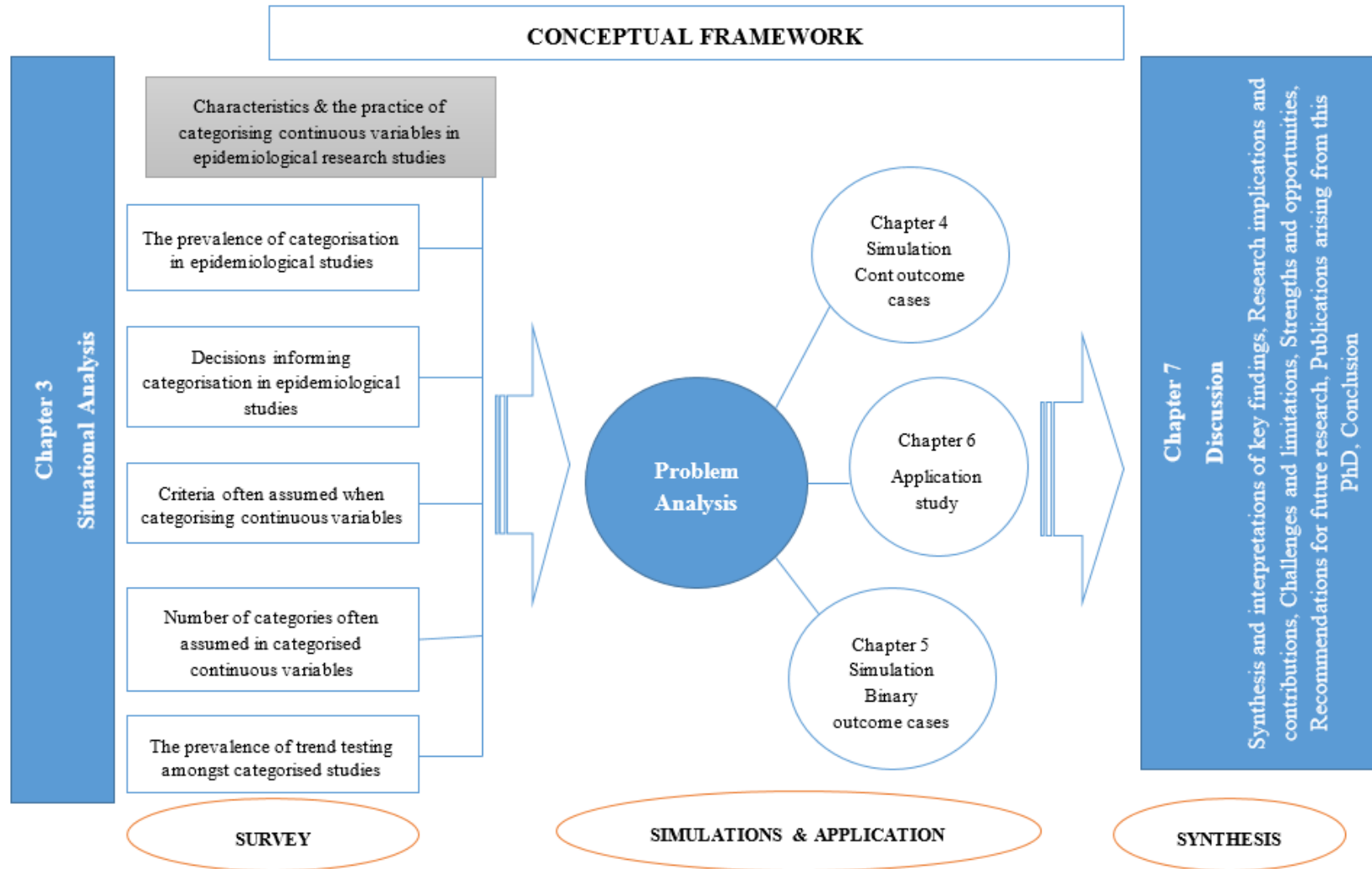


Figure 7.1: Conceptual framework and linkages of the thesis results chapters

7.2 Synthesis and interpretation of key findings

This section discusses key findings emerging from this PhD, bridging chapters together.

7.2.1 The current practice of reporting and analysing continuous variables in observational studies

The findings of the survey study conducted in Chapter 3 revealed the presence of categorisation of continuous variables in epidemiological studies. Amongst the articles investigating the associations between the continuous exposures and disease outcomes, 61% (CI = 39%, 80%) of them transformed the exposure variables into categorical measures for analysis. Amongst these articles, the justifications informing the choice of categorisation was only explained in 7% (CI = 0%, 34%) of the studies. These findings show that things have not improved since the release of the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines (Von Elm et al., 2007). Thus, researchers could be unaware of the existing STROBE guidelines or simply ignoring them underestimating the consequences of categorisation during statistical modelling. The implications of categorisation were investigated in simulation studies provided in Chapter 4 and Chapter 5. The findings in these chapters revealed the inability of this method to discover the ‘true’ associations in the simulations and estimate turning points in the data. Furthermore, categorical models were characterised by poor performance measures in the simulations. This is because categorisation does not use the within category information. The alternative methods of FPs and RCS that preserve continuous data are available. However, the survey results in Chapter 3 suggest that such methods are rarely used for reporting exposure-outcome relationships in epidemiology.

7.2.2 Implications of categorisation and comparison of alternative methods

Given the high incidence of categorisation in Chapter 3, the researcher wanted to establish how bad categorisation was amongst studies investigating exposure-outcome relationships. To achieve this, two methods of categorisation - assuming three (CAT3) and five (CAT5) categories were compared against the alternative approaches of FPs and RCS through the simulation studies presented in Chapter 4 and Chapter 5. The simulations also incorporated linear regression models – which offer simplicity amongst studies that keep continuous variables in the analysis. The reasons informing the choice of FPs and RCS were as follows; (1) the two approaches are considered powerful and flexible in fitting both complex and linear functions, (2) they are readily available for implementation in most statistical programs, (3) the comparisons between them are lacking and little is known about their results and (4) they possess special unique features - RCS are constrained to be linear at their tails whilst FPs offers flexibility and are constrained through a set of powers, $p_j \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Besides the FPs and RCS methods, the other approaches such as LOESS and GAMs are available but they were not considered in this research. The application of LOESS in multivariable setting is limited (i.e. only useful in single-predictor models) whilst the GAMs are computational expensive (Beck and Jackman, 1998, May and Bigelow, 2005).

To aid the evaluation of methods used in this research, novel simulations were performed - assuming plausible relationship scenarios in epidemiology. The plausible scenarios were exemplified by using the alcohol-blood pressure relationships found in the literature. However, it was difficult to envision the influence of other covariates (or confounding) in the simulations thus, single predictor-outcome models were considered for evaluation. The predictor-outcome shapes in the simulations comprised the linear function, two different thresholds functions (linear and nonlinear - tick shaped) and U

shaped function – referred as ‘true’ shapes for estimation using the CAT3, CAT5, linearization, FPs and RCS regression approaches.

7.2.2.1 Ability to discover ‘true’ predictor-outcome relationships

The research in Chapter 4 and 5 showed that categorising or linearising continuous predictor or exposure (used interchangeably in this chapter) produces functions that fail to lend themselves on the ‘true’ nonlinear shapes. The CAT3 and CAT5 approaches produced step functions – concealing detailed information about the actual shapes in the simulations. For instance, the relations between the predictor and outcome variables were constant within the assumed categories. This implies that ‘true’ biological changes occurring in linear, thresholds and U shaped associations were lost due to constant variation within the assumed categories. In contrast, the linearisation approach produced fits that lied entirely on the ‘true’ function only when the relationship was linear. Besides that, the linear models produced inadequate fits when applied in thresholds (both linear and nonlinear) and U shaped datasets - underestimating the ‘true’ outcomes at the lower and upper predictor values. For moderate exposure values, the outcomes were overestimated when fitting the linear functions in thresholds and U shaped datasets. Overall, the results in Chapter 4 (for Gaussian outcome models) were similar to those in Chapter 5 (for binary outcomes). The CAT3, CAT5, and linearisation methods were inadequate for characterising thresholds and U-shaped relationships. Thus, applying these methods in similar scenarios will produce inaccurate functions that mislead individuals’ health outcomes. The alternative methods of FPs and RCS improved the results in thresholds (both the linear and nonlinear) datasets – producing fits that were very close to the ‘true’ functions. However, none the predicted fits lied entirely on ‘true’ threshold functions. This was true in both the continuous and binary outcome models. The RCS models also produced near approximate fits in U-shaped datasets under the continuous and binary

models. In contrast, the FP method produced precise fits that lied entirely on the 'true' U-shaped function when both the predictor and outcome were analysed as continuous variables (i.e. continuous models). However, near approximate fits that do not lie entirely on the 'true' curve were obtained when applying FPs under the binary outcome scenario. This latter implies that any deviation from the assumption of normality to binary outcomes will reduce the precision of the FP function.

One of the limitations in this PhD is studying a few nonlinear shapes in the simulation chapters. Some of the functions found in epidemiology but not considered in the simulations include the asymptotic, J-shaped and sigmoidal relationships. For example, an asymptotic relationship was reported between blood carboxyhaemoglobin (or plasma thiocyanate concentrates) and daily smoking of cigarettes, reaching an asymptote after consumption rate of 25 cigarettes per day (Vesey et al., 1982). Evidence of J-shaped relations includes the associations between BMI and coronary heart disease or mortality (Jee et al., 2006, de Gonzalez et al., 2010). A sigmoid relationship exists between Vitamin C ingestion (or dosage) and its concentration (Paoletti et al., 1998). Supposing such relationships existed in real application studies, the categorical and linear models will fail to identify (or detect) the 'true' shapes in the data. The findings in Chapter 4 and Chapter 5 suggest the methods of categorisation would hide the actual relations in such scenarios producing step functions. In contrast, the linear models would restrict these nonlinear functions to be linear (which is misleading and inaccurate). For alternative methods, the RCS model would likely struggle to lend itself on the J-shaped function at the lower exposure since the RCS model predict linear fits at the tails. However, the RCS should not struggle to accurately identify the asymptotic fits since such functions tend to be linear at their tails. In contrast, the FP functions would likely to lend itself on the J-shapes than the asymptotic fits. The J curves belong to the quadratic functions thus, it is expected that the FP fit identify them as it does with

the U-shaped functions in Chapter 4 and Chapter 5. On the other hand, the FP would likely struggle to lend itself on asymptotic functions since its fits tend to be too flexible at the tails. For sigmoid functions, the RCS (with 3 knots) as assumed in this thesis would be insufficient. Sigmoid functions have more than one turning point thus, additional internal knots maybe required for the actual fit. Similarly, FPs may be inadequate; requiring high order degree fit (>2) - depending on the complexity of the sigmoid shapes.

7.2.2.2 Turning point estimation

The simulations in Chapter 4 & Chapter 5 showed that the FP and RCS functions were not adequate for predicting the location of turning points (or thresholds) in linear and nonlinear threshold datasets (tick shaped). The FP and RCS models underestimated the ‘true’ outcomes in the two threshold functions – shifting the ‘true’ location of the exposure/predictor variables to the left. Moreover, the two methods produced varying estimates in similar datasets (i.e. under both the linear and nonlinear threshold functions). This has huge clinical implications, different optimal exposure/outcome values often reported in application studies could be due to these methodological variations. The alternative change point regression methods (Muggeo, 2003, Breitling, 2015) that were not explored in this PhD maybe suited for estimating the positions of turning points (together with their CIs) in the simulations. Although the FP and RCS models were not adequate for predicting the location of turning points for threshold scenarios in the simulations; the FP models produced precise estimates in U-shaped functions (under the continuous outcome scenarios). In contrast, the RCS models misspecified the position of the ‘true’ turning points – shifting it to the right. The RCS results were similar across the continuous and binary outcome scenarios in Chapter 4 and Chapter 5.

The CAT3 and CAT5 approaches were not suitable for estimating exact location of turning points in the simulations. The two methods of categorisation forced the predicted outcomes to remain invariant within each category of the predictor - producing interval estimates. Considering that the estimated turning points may be useful for establishing treatment dosages or therapeutic interventions, interval estimates are not explicit thus individual's health outcomes could be compromised. Therefore, I suggest that researchers use such estimates for quantitative hypothesis – testing whether the precise location of the turning point is within the predicted intervals.

7.2.2.3 Goodness of fit

The RMSE (also a predictive measure) was used in Chapter 4 to measure the variation between the predicted curves and the 'true' functions in the simulations. In contrast, discrimination and calibration were used in Chapter 5 for prognostic models with binary outcomes. Discrimination (measured using c-index scores) was defined as the ability of the predicted model to separate patients with different event outcomes whilst calibration (assessed by graphical plots) was the extent of agreement between observed and predicted outcomes (Harrell et al., 1996). In Chapter 4, the two methods of categorisation (CAT3 and CAT5) had larger RMSE scores compared to the FP and RCS methods when fitted in linear, thresholds and U – shaped datasets. The FP functions retained the smallest RMSE when fitted in nonlinear datasets (thresholds and U-shaped associations) followed by the RCS fit. When the association was linear, the linearisation approach produced the best fit - with minimum RMSE followed by the RCS and FP fits respectively. In addition, the FP method was likely to reject linearity more often than the RCS approach - even when the true relationship was linear (resulting in high type I errors). Applying the RCS model in linear datasets produced

type I errors close to 5%, whilst the type I errors in FP models varied 14% and 25%. The latter suggests the FP approach was likely to produce over fitted models when applied in linear datasets than the RCS functions – an implication that the RCS is more conservative than the FP approach. The suggestion that the RCS is conservative than the FP approach was also observed in binary outcome models. The RCS models retained slightly larger c-index scores than the FP method in binary outcome functions (occurring after transforming continuous outcomes into binary variables) – an indication that the RCS approach was better in such scenarios. Categorising the continuous predictor variables into few (CAT3) categories (under the binary outcome models) displayed functions with the least discrimination than the linear, FP, RCS approaches. However, the CAT5 (with more categories) improved the discrimination results. In contrast, the CAT5 competed fairly with the linear, FP and RCS functions - slightly outperforming the linear and FP methods in some scenarios. Moreover, the CAT5 retained functions with better calibration plots than the linear, FP and RCS approaches (whose functions were characterised by combinations of agreements and disagreements in the simulations). Based on these results, researchers might be tempted use models with more categories in predictive analysis. However, the improvements on models with large number of categories come as a trade-off for more complex functions (with many parameters or degrees of freedom) than the categorical models (with few categories) or other methods that preserve and make full use of predictor information in the analysis.

7.2.3 The differences between continuous and binary outcome models

The simulation in Chapter 5 investigated the performances of categorisation, linearisation, FP, and RCS methods in binary logistic models. The survey research in Chapter 3 showed that binary outcome models were very popular in medical studies. Often binary logistic models are investigated after categorisation of continuous outcomes - using prevailing definitions (or clinical cut points) of disease conditions for

the purpose of prediction and explanatory analysis. However, there exist few studies investigating predictive models (Shmueli, 2010). The simulations in Chapter 5 focused on the performance of different methods used for handling continuous predictors when developing predictive models under the logit framework. These simulations were different from the normal error regression models in Chapter 4 where both the predictor and outcome variables were continuous. Furthermore, the prognostic models in Chapter 5 formed an example of categorisation for both the predictor and outcome variables. Considering the different simulation framework in chapter 4 and 5, this was an opportunity to study several performance measures for continuous and binary outcome models. Notable differences observed between the results comparing the methods of categorisation, linearisation, FP and RCS in chapter 4 and 5 are discussed in sections 7.2.3.1 and 7.2.3.2.

7.2.3.1 Confidence Intervals of the estimated functions

The most outstanding difference observed between binary and continuous models in the simulations was their 95% CIs. The binary outcome scenarios retained functions with wider CIs compared to the settings where the predictor and outcome variables were continuous. Based on these findings, any deviation from the assumption of normality (or Gaussian distribution) will reduce the precision or accuracy of functions being estimated. Putting this in context, this implies that transforming continuous outcomes into binary variables would reduce the accuracy of the estimated functions under these methods. Generally, the information is lost when continuous outcomes are transformed into binary variables.

Overall, the fitted functions had the widest CIs at the tails of the predictor variables than at the centre. This pattern was the same across the binary and continuous outcome models. However, the comparison between the fitted functions showed some striking difference at the lower tails of the predictor variables. Across the four

association shapes in the simulations, the FP curves had wider CIs at the lower tails of the predictor than the categorical, linear and RCS fits. This was mainly due to zero values occurring at the lower tail of the predictor variable. FPs are based on natural logarithms thus they not able to handle zero values. This consequently causes unstable fit at the lower tails - resulting in wider CIs. Reflecting on this problem, this could be one of the reasons why FPs are not widely used in medical studies (as seen in the survey in Chapter 3). Researchers may not be comfortable with using FPs because their resulting trends functions may be biologically implausible.

7.2.3.2 Coverage probabilities of the turning points

The average coverage probabilities of the ‘true’ turning points exceeded the 95% nominal levels when applying the FP and RCS functions under the continuous outcome scenarios. The latter was observed in both the thresholds and U-shaped functions considered in Chapter 4 (under all simulation conditions – varying sample sizes and noise in the data). In contrast, the average coverage probabilities of 72% and 76% were observed when fitting the FP functions in linear and nonlinear threshold datasets respectively, under the binary scenarios. In the same datasets, the RCS had conservative estimates above the 95% nominal level. For U-shaped functions, both the FP and RCS had the average coverage probabilities below the 95% nominal level. However, the RCS was closer to the nominal level with the average rate of 87% than 66% under the FP. These results suggest underperformance of FP in binary outcome scenarios compared to the continuous outcome cases. In contrast, the RCS was not adversely affected by deviations from the assumption of normality – the results remained the same except in U-shaped functions. Reflecting on these results, the coverage rate below and far from the 95% levels implies under-coverage and lack of fit. Thus, under-coverage and lack of fits were more likely under the binary outcome scenarios.

7.2.4 Application study - Examining the alcohol-hypertension association in type 2 diabetes patients using the UK Biobank

For the application of methods studied in Chapter 4 and Chapter 5, an investigation was carried out in Chapter 6 to assess the association between alcohol consumption and the odds of hypertension in patients with type 2 diabetes using the UK Biobank. In addition, Chapter 6 was also aimed at investigating whether patients' age and antihypertensive medication use modify the alcohol-hypertension relationships. Moreover, Chapter 6 was also an example of real world data and use of categorisation, linearisation, FP and RCS approaches.

This chapter showed that the proportion of type 2 diabetes patients with poor blood pressure control (defined as SBP \geq 130 mmHg in diabetics) in the UK Biobank (70%) is substantially high. Taking into account this finding, aggressive strategies are needed to manage and control hypertension in type 2 diabetes patients in the UK Biobank.

Amongst type 2 diabetes patients, the odds of hypertension increased monotonically with the consumption of alcohol in the UK Biobank. This relationship was observed when fitting the categorical, linear, FP and RCS models. The application of the categorisation method reported as the popular practice in Chapter 3 yielded a non-negative monotonic function that increased in steps. As observed in Chapter 4 and Chapter 5, the step functions produced through the categorical approach obscure 'true' relationships in the data due to loss of information. The alternative methods that keep the continuous variables in the analysis resulted in a non-negative linear relationship when fitting linear models. In contrast, the FP and RCS revealed the presence of nonlinearity in the UK Biobank. The adjusted odds of hypertension in the FP and RCS models were different for alcohol consumption below 2.2 g/day and comparable when the consumption exceeded 2.2 g/day. Notable differences at the

alcohol consumption range between 0 and 2.2 g/day include (1) a large ‘spike’ at zero units of alcohol consumption and (2) the ORs <1 for alcohol intake between 0.1 and 2.2 g/day in the FP function. In contrast, the ‘spike’ at zero consumption of alcohol was not observed when applying the RCS model. Furthermore, the RCS function was always positive and increasing with ORs >1 . Between 2.2 and 17.5 g/day of alcohol consumption, the two functions were comparable and had steeper and increasing slopes on their ORs. When alcohol consumption exceeded 17.5 g/day, the slopes in the FP and RCS models were shallow with OR trends that increase monotonically. No optimal amount of alcohol consumption was linked with the reduction of odds of hypertension amongst patients with type 2 diabetes in the UK Biobank. This contradicts findings in general population where the optimal risk of hypertension has been reported amongst the light or moderate alcohol drinkers (Gillman et al., 1995, Moreira et al., 1998). The present study found increasing or positive odds of hypertension amongst the light and moderate alcohol drinkers.

Additional findings also suggest that antihypertensive medication use and patient’s age modify the odds of hypertension. In the analysis stratified by antihypertension medication use (medication user’s vs non-users); the probabilities of having hypertension were high amongst patients on antihypertensive medication compared to non-medication users and the harm associated with alcohol drinking was visibly weak amongst heavy drinkers compared to light and moderate drinkers. The latter was observed in the categorical, linear and RCS models and suggest the efficacy of antihypertensive use amongst heavy alcohol drinkers. In contrast, the efficacy of antihypertensive medication was undetectable when fitting the FP model. The findings suggesting the efficacy of antihypertensive use has also been reported in other clinical studies examining the influence of alcohol consumption on hypertension (Maheswaran et al., 1990, Wakabayashi, 2010). In the analysis stratified by age, the probabilities of

having hypertension were high amongst aging patients. However, young adult patients were severely affected by the harm associated with alcohol drinking than older patients. The four models applied in the analysis were in agreement with these findings and suggested the presence of *alcohol X age* interaction (which was not strong when fitting the RCS function) in the data. In contrast, Okubo and colleagues (2014) found consistent linear associations between alcohol intake and the risk of hypertension in middle-aged (40-59 years) and older (60-79 years) individuals suggesting the absence of *alcohol X age* interaction term in general population. Thus, further research is needed to confirm the presence of *alcohol X age* interaction in diabetes patients. The suggested *alcohol X age* interaction may only be in this study.

7.3 Research implications and contribution to knowledge

The research implications and contributions of each chapter in this thesis are discussed below:

For the first time after several years in existence, a new piece of research was carried out in Chapter 3 to assess the current practice of reporting and analysing continuous variables in observational studies according to the STROBE guidelines. The findings showed a higher incidence of categorisation - raising concerns about the adequacies of analysis and quality of reporting continuous exposure or risk factors in epidemiology. This result was an indication that researchers may be unaware of the existing STROBE guidelines or they simply ignore them. This result also suggests researchers might not be attracted to the alternative methods of FP and RCS suitable (shown in Chapter 4 and Chapter 5 of this thesis) for improving the reporting relationships in epidemiology. Taking note of these results, it is so important to encourage researchers to adopt the alternative approaches for improved reporting. The common practice of categorisation does not make use of within category information -

yield functions that obscure underlying relationships (shown in Chapter 4 and Chapter 5). Therefore, the costs of categorisation may be huge in public health if the exposure-outcome relationships are not well established. For instance, patients' health outcomes may be wrongly classified resulting in misguided interventions.

The research in Chapter 4 suggests the FP and RCS functions were not adequate for predicting both the 'true' fits and turning points in a linear or nonlinear threshold (tick shaped) datasets. The actual positions of turning points were underestimated (shifted to the left of the exposure scale) when fitting FP and RCS models in the two threshold functions. However, the two methods produced varying estimates when applied to a similar threshold dataset. Under the U-shaped datasets, the FP model was able to accurately identify both the 'true' curve and the exact location of turning points. In contrast, the RCS overestimated the positions of the 'true' turning points in the same datasets (shifted to the right of the exposure scale). In comparison, applying the methods of categorisation in the simulations produced step functions, which were inadequate for the 'true' relationships and estimation of turning points. Essentially, these findings are novel and can also be extended to binary outcome scenarios to guide medical researchers reporting similar shapes in their studies. The researcher was not aware of any simulation study evaluating the estimation of turning points in plausible exposure-outcome relationships using the categorisation, FP, and RCS methods. Even though researchers often visualise and approximate the location of the turning points based on the shapes of predicted association functions (Pastor and Guallar, 1998).

The research in Chapter 5 was based on novel simulations examining the predictive ability of prognostic models developed using the CAT3, CAT5, linearisation,

FP, and RCS methods under three key measures of discrimination, calibration, and clinical utility. Overall, there was poor discrimination and miscalibration when applying the five methods in the simulations. However, the RCS methods had greater c-index scores whilst the CAT3 retained the least c-indexes (although the difference between the five methods was not substantially large or worse). The calibration results were characterised by lower predicted probabilities than those observed in 'true' functions when applying CAT3 and CAT5). However, the two methods of categorisation were better calibrated than the linear, FP and RCS models that showed combinations of agreements and disagreements between predicted and observed probabilities in the simulations. For clinical usefulness, the CAT5, linear, FP and RCS methods showed better clinical net-benefits when applied in linear, thresholds and quadratic datasets than the CAT3 approach. The findings of poor net-benefits on categorical models (with few categories) and improvements on models with large categories were also reported by Collins and colleagues (2016). Overall, these are novel findings that provide insights about categorisation, linearisation, FP and RCS methods in prognostic modelling. The results imply that categorising continuous predictors into few categories during model development would produce models with the least predictive accuracy (discrimination and calibrations) and poor clinical net benefits than the prognostic models developed using the CAT5, linear, FP and RCS approaches. The poor performance of the categorical model with few categories was attributed to information loss occurring when continuous predictors are transformed into categories during model development. In contrast, the CAT5 achieves better performances by trading off its many categories for more complex step functions (which is not efficient), whilst the other methods fully utilise the continuous predictor information. Therefore, with the emerging field of machine learning; researchers working in this area might find this research useful to validate their models. Predictive analytics go hand-in-hand with machine learning

where big data – large volumes of raw structured, semi structured and unstructured data are used to estimate or predict future outcomes (Obermeyer and Emanuel, 2016).

A literature search on Google Scholar, PubMed and Web of Science using the following keywords; UK Biobank, type 2 diabetes, hypertension, alcohol, associations or relationships suggest Chapter 6 was the first study investigating the alcohol-hypertension relationship amongst type 2 diabetes patients in the UK Biobank using the methods of categorisation, linearisation, FPs, and RCS. Findings in this chapter showed a substantially large proportion of type 2 diabetes patients with poor blood pressure control (defined as SBP \geq 130 mmHg in diabetics) in the UK Biobank (70%). The odds of hypertension increased monotonically with the consumption of alcohol in the UK Biobank amongst the type 2 diabetes patients suggesting nonlinear relationships when fitting the FP and RCS models. In contrast, the odds of hypertension were increasing in steps when fitting the categorical model whilst the linear model produced a positively linear relationship. The additional findings suggested antihypertensive medication use and patients' age modifies the odds of hypertension. In the analysis stratified by antihypertension medication use (medication user's vs non-users); the probabilities of having hypertension were high amongst patients on antihypertensive medication than non-medication users and the harm associated with alcohol drinking was visibly weak amongst heavy drinkers compared to light and moderate drinkers. The latter was observed in the categorical, linear and RCS models and suggested the efficacy of antihypertensive use amongst heavy alcohol drinkers. In contrast, the efficacy of antihypertensive medication was undetectable when fitting the FP model. In the analysis stratified by age, the probabilities of having hypertension were high amongst aging patients. However, young adult patients were severely affected by the harm associated

with alcohol drinking than older patients. The four models applied in the analysis were in agreement with these findings and suggested the presence of *alcohol X age* interaction (which was not strong when fitting the RCS function). Overall, these are novel results that suggest the need for aggressive strategies to manage and control alcohol induced hypertension in type 2 diabetes patients and clinical guidelines explaining the risk of hypertension in patients with type 2 diabetes taking into account (a) the efficacy of antihypertension medication use and (b) the severity of alcohol drinking according to patient's age. On the methodological side, researchers are encouraged to consider nonlinearity when investigating the associations between alcohol consumption and hypertension in patients with type 2 diabetes. Considering nonlinearity and characterising the exposure-outcome relationships accurately is important because such studies inform policies that influence individuals health outcomes. Taking this application study as an example, ignoring nonlinearity will not recognise steep OR slopes (increasing at a decreasing rate in FP and RCS models) requiring health policy makers to pay attention at lower values of the exposure. Thus, this study is also an example of errors that can result in any epidemiological study not addressing the issues of nonlinearity and categorising continuous variables – a common practice observed in the survey provided in Chapter 3.

7.4 Challenges and limitations

It is important to point out that there were some challenges and limitations encountered in this PhD. These challenges and limitations are discussed in section 7.4.1- 7.4.3 below.

7.4.1 PhD scope/coverage

It was beyond the scope of this PhD thesis to compare all the methods available for analysing continuous exposure-outcome relationships. This PhD covered the

common methods of categorisation and the alternative approaches of linearisation, FP, and RCS for characterising relationships in epidemiology. Despite these methods, there exist other approaches for analysing exposure-outcome relations that were not proposed but may have benefited from consideration in this PhD.

Assuming knowledge of ‘true’ exposure-outcome relationships in epidemiology, the performances of categorisation, linearisation, FP and RCS approaches were investigated in simulation studies under the framework of continuous and binary outcomes (see Chapter 4 & Chapter 5). The simulations covered few exposure-outcome relationships exemplified by alcohol-blood pressure association scenarios found in the literature. These example scenarios were only used to illustrate the applications of the methods, therefore could not be interpreted as estimates for the causal effects of alcohol on blood pressure (confounding also not considered). Apart from the latter, the properties and applications of these methods could be extended to any similar exposure-outcome relations studied in Chapter 4 & Chapter 5 provided the exposures are continuous variables.

7.4.2 Methods applications

The mathematical expressions of the best-fitted FP and RCS functions attained when estimating the alcohol-hypertension relationships in Chapter 6 were complex and difficult to interpret. This was not surprising because the FP and RCS methods are known for not providing interpretable parameters when estimating exposure-outcome relationships (Royston and Altman, 1994, Heinzl and Kaider, 1997, Steenland and Deddens, 2004). Because of this challenge, parameter estimations and their interpretations were avoided. Functional interpretations and comparisons of results were achieved by using graphs and tabulations. This is the best practice recommended for reporting results when applying these methods (Royston and Altman, 1994, Royston et al., 1999).

In spline regression modelling, the procedures for deciding the number of knots and their placement in exposure-outcome studies are unclear (Smith, 1979, Durrleman and Simon, 1989). This PhD applied a reasonable approach recommended by Harrell (2001) for knots placement over the quantiles distribution of the exposure. Harrell's method of knots selection is useful when a prior knowledge about the exposure-outcome relationships is unknown. The procedure for knots placements in this method is less subjective and allows reproducibility and comparison of results between studies (Heinzl and Kaider, 1997). Other strategies include adaptive procedures based on standard algorithms for "optimal" knots selection (Morton, 1988, Friedman and Silverman, 1989, Luo and Wahba, 1997, Zhou and Shen, 2001). The disadvantage with adaptive procedures is that they are subjective - there exists no standard algorithm which produces the best possible number and position of knots from the data alone (Morton, 1988, Zhou and Shen, 2001). Additionally, these methods exhibit computational burden in large samples because sets of candidate knots have to be examined to establish the 'optimal' number of knots and their positions (Zhou and Shen, 2001).

7.4.3 Data

The simulations in Chapter 4 & Chapter 5 were setup assuming uniformly distributed predictors. This is another limitation of these simulation studies - uniformly distributed predictors may not be too realistic. In epidemiology, skewed or 'spike' at zero (SAZ) situations such as those under the lognormal, normal distributions (or mixture with uniform) are common. Thus, the simulation results would likely be different under the skewed or SAZ situations. For example, if the 'true' predictor-outcome function was steeper at the tails because the predictor effects are concentrated in one end of the distribution, defining the end categories too broadly would obscure the direction and the 'true' predictor-outcome functions in that region. In that situation, the

end categories assume the event outcomes are homogeneous (Greenland, 1995a, Bennette and Vickers, 2012) which is not correct. The alternative methods of FPs would produce fits with some spike behaviour at the zero values of the predictor. However, this behaviour will be more striking due to larger proportions of zeros in skewed distributions. FPs cannot deal with zero predictor values because they based on natural logarithms thus, the spike behaviour around the zero region. In contrast, the RCS models would produce linear fits at the extreme tails of the predictor variables.

The issues of measurement or misclassification errors were not dealt with in this PhD thesis. This means the alcohol-hypertension relationships reported in Chapter 6 (application study) could be exposed to some form of measurement or misclassification bias due to the error-prone alcohol intake measures (exposure). In epidemiological studies, measurement bias is known to occur in models that analyse error-prone continuous exposures without categorising them. In contrast, categorising error-prone continuous exposures is known to induce misclassification bias (Brenner and Loomis, 1994, Dalen et al., 2009). This suggests the linear, FP and RCS models would likely suffer from the measurement bias whilst the method of categorisation would be susceptible to the misclassification bias. Taking note of this problem, the issues of measurement or misclassification errors need to be addressed separately by (1) assessing the direction and magnitude of the potential bias in exposure-outcome relationships and (2) offering corrective measures for the potential bias in the exposure-outcome data. To achieve the latter, a simulation study needs to be performed by exemplifying the alcohol-hypertension relationships in patients with type 2 diabetes as reported in the UK Biobank (see Chapter 6).

In the application study, the exposure-outcome relationships were reported based on complete case analysis – omitting a third of incomplete data in the predictors. This could affect the predicted exposure-outcome functions making them less precise due to reduced sample size and potential loss of power on the data. Also, the confidence intervals of the predicted functions could be incorrect due to the potential bias in standard errors (White, 2015). An attractive approach for handling missing values of predictors in epidemiological studies is the multiple imputation (MI) method (Rubin, 2004). However, the MI procedure was not performed in Chapter 7 of this thesis. The two recent studies recommend not to choose the MI method when investigating nonlinear associations and interaction terms (Bartlett et al., 2015, White, 2015).

7.5 Strengths and opportunities

The strengths and opportunities of this PhD are as follows:

- i. Provide an update on the current practices of categorisation in leading medical publications - raising concerns about the adequacies of analysis and quality of reporting continuous exposure or risk factors in epidemiology.
- ii. Demonstrate and compare the performances and properties of categorical, linear FP and RCS methods based on rational and plausible simulation scenarios in epidemiological studies. The simulations cover the continuous and binary outcome models – focusing on the ability of these methods to characterise the ‘true’ relationships assumed in the data. The assessment and evaluation of model performances were carried out using measures suitable for the simulated datasets.
- iii. Demonstrate and compare the application of categorical, linear FP and RCS methods in real application data - dealing with perceived challenges of confounding, interactions and interpretations of the results.

7.6 Recommendations for future research

Evidence from this PhD thesis suggests the categorical, linear, FP and RCS regression models cannot accurately characterise the threshold association datasets. This calls for further research exploring suitable methods for estimating the ‘actual’ threshold points and their functional parameters.

It would also be interesting to evaluate how the distribution of the exposure affects the performance of categorical, linear, FP and RCS models in more explicit ‘spike’ at zero distributions or under lognormal, normal distributions. To my knowledge, there exists no simulation study comparing the performance of these methods and considering the ‘spike’ at zero situations in different exposure-outcome relationships. In a recent study, the methods of categorisation, linearisation, and fractional polynomials were compared using three case-control datasets with ‘spike’ at zero situation (Lorenz et al., 2017).

Further research is also proposed in the UK Biobank to develop a multiple imputation model for missing alcohol-hypertension data - providing guidelines for its implementation.

Since the exposure-outcome relationships are prone to measurement and misclassification errors, further research is required to (1) assess the direction and magnitude of these potential biases in exposure-outcome relationships and (2) offer their corrective measure in the data. To achieve this, a simulation study may be performed - exemplifying it with the alcohol-hypertension relationships amongst patients with type 2 diabetes as reported in the UK Biobank.

Finally, an application study is proposed using a two-part regression model by Duan et al (1987) or the compound Poisson exponential model by Jørgensen (1987,

1997). Certain features of the curve may not be adequately interpreted when fitting FPs and RCS models in skewed datasets (especially at the tails).

7.7 Publications arising from this PhD thesis

The publications associated with this thesis are given below.

Conference presentations

1. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *A simulation study investigating the performance of traditional and alternative approaches for fitting nonlinear exposure-outcome relationships in epidemiology*. Oral Presentation at the LICAMM Early Career Group Science Day. 27th May 2017, University of Leeds, United Kingdom.
2. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *A survey based study evaluating the incidence and categorisation of quantitative variables in medical research*. Oral Presentation at 39th Research Students' Conference in Probability and Statistics. 14th – 17th June 2016, Dublin, Ireland.
3. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *Fractional polynomial and restricted cubic spline models as alternatives to categorising continuous data: applications in medicine*. Poster Presentation at Faculty of Medicine and Health Postgraduate Symposium. 29th June 2015, University of Leeds, United Kingdom.

Published Protocol

4. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. (2017).
Modelling the alcohol-blood pressure associations in type 2 diabetes patients:
UK Biobank. UK Biobank, United Kingdom.
[\(http://www.ukbiobank.ac.uk/2017/07/mr-onkabetse-mabikwa-modelling-the-alcohol-blood-pressure-associations-in-type-2-diabetes-patients-uk-biobank/\)](http://www.ukbiobank.ac.uk/2017/07/mr-onkabetse-mabikwa-modelling-the-alcohol-blood-pressure-associations-in-type-2-diabetes-patients-uk-biobank/).

Published Research

5. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. (2017).
Assessing the reporting of categorised quantitative variables in observational
epidemiological studies. *BMC Health Services Research*, 17, 201.
 [\(http://doi.org/10.1186/s12913-017-2137-z\)](http://doi.org/10.1186/s12913-017-2137-z).

In preparations

6. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *Comparison of different approaches for modelling associations between an exposure and a continuous outcome – a simulation study*. *International Journal of Statistics and Probability*. In preparation
7. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *Comparison of different approaches for developing prognostic models with binary outcomes – a simulation study*. *Statistics in Medicine*. In preparation
8. Mabikwa, O. V., Greenwood, D. C., Baxter, P. D. & Fleming, S. J. *Examining the alcohol-hypertension association in type 2 diabetes patients using the UK Biobank*. *Diabetes care*. In preparation

7.8 Conclusions

As argued in Chapter 1, this PhD thesis was motivated by the common practice of categorisation and the limited application of alternative methods for analysing unknown predictor-outcome relationships in medicine. The performances and properties of the alternative approaches of fractional polynomials and restricted cubic splines were shown and compared against the popular method of categorisation and the linear approach using different exposure-outcome relationships in simulation studies (assuming continuous & binary outcome models). In addition, the applications of these methods were exemplified using real-dataset from the UK Biobank to encourage their use - demonstrating communication or interpretation of the results from these approaches. Therefore, it is my hope that this PhD will go some way to highlight to practitioners the application of these alternative methods when reporting continuous predictor-outcome relationships in their studies.

Appendix A

A.1 Data collection form

Section 1: General information

D.O.I (Digital Object Identifier)

Reviewers

1: O.V.M

2: S.L.F

3: P.D.B

4: D.C.G

(4) Does this article satisfy the eligibility criterion?

1: Yes

2: No

(If Yes, skip to Q6 otherwise stop entry at Q5)

(5) If No, which eligibility criterion is not satisfied?

(5a) If other, specify

1: Sample size too small

2: Analysis based on pooled data/estimates (e.g. systematic reviews & meta-analysis)

3: Main investigation based on effect modification or interaction effects

4: Based on experimental or randomised control trial designs

5: Non-relevant article (e.g. comments, reviews, correspondences, tutorials, seminars etc.)

6: Non-full text abstract (e.g. poster/conference/meeting abstracts)

7: Cohort or profile update studies

8: Article not relevant

9: Other

(6) Journal

1: International Journal of Epidemiology

2: Epidemiology

3: Journal of Clinical Epidemiology

4: New England Journal of Medicine

5: Lancet

(7) Types of study design

(7a) If other, specify

1: Cohort

2: Case-control

3: Cross-sectional

4: Other

(8) Sample size (no.of participants)

Section 2: Outcomes

(9) The principal disease outcome or condition studied

(10) Type of outcome

(10a) If other, specify

1: Binary: absent/present event (yes/no)

2: Time-to-event or events/person-year

- 3: Ordered categorical
- 4: Unordered categories
- 5: Continuous or quantitative measure
- 6: Other

(10b) For continuous outcomes, how were they analysed?

- 1: Quantitatively/Continuously
- 2: Qualitatively/Categorical/Grouping
- 3: Both
- 4: N/A

Section 3: Exposures

(11) Name of the main exposure or risk factor studied

(12) Nature of the main exposure or risk factor studied

(12a) If other, specify

- 1: Lifestyle
- 2: Environmental
- 3: Pre-existing condition
- 4: Diet
- 5: Biochemical
- 6: Physiological
- 7: Socioeconomic
- 8: Genetic

9: Other

(13) Type of main exposure or risk factor studied

(13a) If other, specify

1: Ordered categorical

2: Unordered categories

3: Continuous or quantitative

4: Other

(14) If the main exposure or risk factor is continuous and was categorised, how many groups were used?

Skip to Q22, if exposures are categorical - (for both unordered & ordered categories)

Details of continuous or quantitative exposure/risk factors

(15) Is the main exposure/risk factor truly quantitative?

1: Yes

2: No

(16) Was it measured or collected quantitatively?

1: Yes

2: No

9: Not specified/Unknown

(17) How was the main exposure/risk factor analysed?

(17a) If other, specify

1: Quantitatively or Continuously

2: Categorically or Ordered categories

3: Both

4: Other

If analysed categorically;

(18) What informed this approach or decision?

(18a) If other, specify

1: Data driven or evidence based categories

2: Hypothesis driven categories

3: N/A

4: Other

9: Not specified

(19) How are the categories analysed, are they ordered categories?

(19a) If No, specify

1: Yes

2: No

3: N/A

(19b) Have they looked at the trend?

1: Yes

2: No

3: N/A

9: Not specified/Unknown

(20) Specify the types of grouping that were used or assumed?

(20a) If other, specify

1: Statistical criteria (e.g. equal sized group)

2. External basis (well established e.g. WHO criteria)

3. Logical grouping/Equally spaced intervals (e.g. 10 year age groups)

4. Random/Arbitrary grouping

5. Quantiles

6. Other

If analysed quantitatively;

(21) Was linearity or nonlinearity between the continuous exposure and the outcome considered?

1: Yes

2: No

3:N/A

9: Not specified

Section 4: Details of the analysis

(22) Specify if any, the model used for producing the estimates between the exposure and the outcome

(23) What were the principal types of statistical results used for reporting the main exposure/risk factor?

1: OR (binary or per unit change in risk factor)

2: RR

3: Difference in % or rates

4: Hazard ratios or other time-to-event measures or rate ratios (e.g. from Poisson)

5. Regression: difference in means or regression coefficients

6: Other

(23a) If Other, specify

Adjustment for other variables/confounders

(24) Are there any other variables considered in the study?

1. Yes

2. No

(25) If Yes, how many are they?

(26) Are there any continuous risk factors analysed as categorical variables?

1. Yes

2. No

(27) If Yes, how many continuous risk factors are analysed as categorical variables?

Presentation of results

(28) How are the results using the categorised principal risk factor or exposure presented?

1: Tables

2: Figures

3: Both

4: Other

(28a) If other, specify

Types of estimates

(a)

(b).

(c).Reference group

1: Yes

1: Yes

1: Lowest

2: No

2: No

2: Middle

3: SE's

3: Highest

4: Other

9: N/A

- (29) Continuous e.g. regression coefficients
- (30) By group for all groups
- (31) By group relative to reference group
- (32) Other, specify

P-values and other statistical significance tests

Types of tests

1: Yes

2: No

3: With CI

- (33a) None
- (33b) Continuous analysis p-value
- (33c) Trend test (i.e. scores for groups)
- (33d) Pairwise p-values (i.e. every group relative to reference group)
- (33e) Global p-values (e.g. LRT model with all dummy vs w/t all dummy)
- (33f) Other, specify

(34) Any other comments?

Appendix B

B.1 Example of the simulation codes used to generate data in chapter 4

```

clear all
capture log close
global nmc = 1000
set seed 231014
set more off
scalar constant = 120
scalar slope = .38
scalar a = 0
scalar b = 60
gen x = .
gen Ey = .
gen y = .
gen u = .

tempname sim
local iter = 1
quietly {
postfile `sim' nobs sigma check using check1, replace
foreach j of numlist 0 300 800 4800 9800 {
    local nobs = 200 + `j'
    set obs `nobs'
    replace x = a+(b-a)*runiform()
    replace x = round(x)
    replace Ey = constant + slope*x
    foreach sigma of numlist 2.5 5.0 7.5 {
        local i = 1
        while `i' <= $nmc {
            local k = `i'+1
            replace u = rnormal(0,`sigma')
            replace y = Ey + u
            fp<x>, scale replace: reg y <x>
            preserve
            levelsof x, local(levels)
            xblc x_1 x_2, covname(x) at (`r(levels)') gen (xlabfp`k' predfp`k'
lbfk`k' ubfp`k')
            keep xlabfp`k' predfp`k' lbfk`k' ubfp`k'
            rename xlabfp`k' xlabfp
            rename predfp`k' predfp
            rename lbfk`k' lbfk
            rename ubfp`k' ubfp
            save FPdta`iter++'
            restore
            post `sim' (`nobs') (`i') (`sigma')
            local i = `i' + 1
        }
    }
}
}

postclose `sim'
use "check1.dta", clear
quietly list ,clean
quietly dir *.dta
quietly {

use FPdta1, clear
    forvalues i = 2/15000 {
        append using FPdta`i'
    }
foreach var of varlist xlabfp predfp lbfk ubfp {
    drop if `var'==.
}
}
save FPxblc

```


B.2 Example of the simulation codes used to generate data in chapter 5

```

clear all
capture log close
global nmc = 1000
set seed 231014
scalar a = 0
scalar b = 60
gen x =.
gen y =.
gen w =.
gen p =.
gen s1 =.
gen s2 =.

tempname sim
local iter = 1
quietly {
postfile `sim' nobs sigma check using check1, replace
foreach j of numlist 800 {

    local nobs = 200 + `j'
    set obs `nobs'
    replace x = a+(b-a)*runiform()
    replace x = round(x)
    replace w = -2.2-0.0128*x+0.00032*x^2
    replace p = exp(w)/(1+exp(w))
    mkspline xs = x, nknots(3) cubic
    replace s1 = xs1
    replace s2 = xs2
    drop xs*

    foreach sigma of numlist 7.5 {
        local i = 1
        while `i' <= $nmc {
            local k = `i'+1
            replace y = rbinomial(1, p)

            logit y s*, iter(5)
            preserve
            levelsof x, local(levels)
            xblc s*, covname(x) at (`r(levels)') gen (xlabrcs`k' predrcs`k'
lbrcs`k' ubrcs`k')
            keep xlabrcs`k' predrcs`k' lbrcs`k' ubrcs`k'
            rename xlabrcs`k' xlabrcs
            rename predrcs`k' predrcs
            rename lbrcs`k' lbrcs
            rename ubrcs`k' ubrcs
            save RCdta`iter++'
            restore

            post `sim' (`nobs') (`i') (`sigma')
            local i = `i' + 1
        }
    }
}

postclose `sim'
use "check1.dta", clear
list clean
quietly dir *.dta
quietly {
    use RCdta1, clear
    forvalues i = 2/1000 {
        append using RCdta`i'
    }
}

foreach var of varlist xlabrcs predrcs lbrcs ubrcs {
    drop if `var'==.
}

}
save RCSxblc

```

Appendix C

C.1 Additional tables in Chapter 4

Table 4.3: Estimated median Root Mean Square Errors (RMSE) and their 95% CI regions obtained when fitting different nonlinear exposure-outcome associations using different regression models

			Shapes of the exposure-outcome relationships			
			Linear association curve	Linear piecewise threshold association curve	Nonlinear piecewise threshold association curve	Quadratic or U-shaped association curve
n	σ	Method				
200	2.5	LIN	0.21 (0.20, 0.22)	2.53 (2.53, 2.53)	7.45 (7.45, 7.45)	8.60 (8.60, 8.60)
		CAT3	2.03 (2.03, 2.03)	3.19 (3.19, 3.19)	6.60 (6.60, 6.60)	6.60 (6.60, 6.60)
		CAT5	1.36 (1.36, 1.36)	2.12 (2.12, 2.12)	4.34 (4.34, 4.34)	4.37 (4.37, 4.37)
		RCS	0.27 (0.26, 0.28)	1.16 (1.15, 1.16)	1.00 (1.00, 1.00)	1.03 (1.03, 1.03)

		FP	0.36 (0.35, 0.36)	0.79 (0.79, 0.79)	0.54 (0.54, 0.55)	0.28 (0.26, 0.28)
	5.0	LIN	0.42 (0.41, 0.44)	2.56 (2.55, 2.56)	7.46 (7.45, 7.46)	8.61 (8.61, 8.61)
		CAT3	2.09 (2.08, 2.09)	3.22 (3.22, 3.22)	6.61 (6.61, 6.61)	6.61 (6.61, 6.61)
		CAT5	1.50 (1.49, 1.51)	2.21 (2.21, 2.22)	4.38 (4.38, 4.39)	4.42 (4.42, 4.42)
		RCS	0.54 (0.52, 0.56)	1.25 (1.24, 1.26)	1.11 (1.10, 1.11)	1.13 (1.12, 1.14)
		FP	0.70 (0.68, 0.71)	0.97 (0.96, 0.99)	0.76 (0.74, 0.77)	0.63 (0.60, 0.65)
	7.5	LIN	0.64 (0.61, 0.68)	2.60 (2.59, 2.61)	7.47 (7.47, 7.47)	8.62 (8.62, 8.62)
		CAT3	2.19 (2.18, 2.20)	3.29 (3.28, 3.29)	6.64 (6.64, 6.65)	6.65 (6.64, 6.65)
		CAT5	1.73 (1.71, 1.75)	2.38 (2.36, 2.39)	4.47 (4.46, 4.47)	4.50 (4.50, 4.51)
		RCS	0.84 (0.81, 0.87)	1.40 (1.38, 1.42)	1.28 (1.26, 1.30)	1.30 (1.28, 1.32)
		FP	1.09 (1.06, 1.12)	1.25 (1.23, 1.28)	1.07 (1.04, 1.11)	1.11 (1.08, 1.14)
500	2.5	LIN	0.13 (0.13, 0.14)	2.81 (2.81, 2.81)	8.32 (8.32, 8.32)	9.73 (9.73, 9.73)

		CAT3	2.18 (2.18, 2.18)	3.67 (3.67, 3.67)	8.16 (8.16, 8.16)	8.07 (8.07, 8.07)
		CAT5	1.37 (1.36, 1.37)	2.20 (2.20, 2.20)	5.04 (5.04, 5.04)	4.98 (4.98, 4.98)
		RCS	0.17 (0.16, 0.17)	1.16 (1.16, 1.16)	1.34 (1.34, 1.34)	1.01 (1.01, 1.01)
		FP	0.21 (0.21, 0.22)	0.81 (0.81, 0.81)	0.51 (0.51, 0.52)	1.17 (1.16, 1.18)
	5.0	LIN	0.26 (0.25, 0.27)	2.82 (2.82, 2.82)	8.33 (8.33, 8.33)	9.73 (9.73, 9.73)
		CAT3	2.20 (2.20, 2.20)	3.68 (3.68, 3.68)	8.16 (8.16, 8.16)	8.08 (8.08, 8.08)
		CAT5	1.43 (1.42, 1.43)	2.23 (2.23, 2.24)	5.06 (5.06, 5.06)	5.00 (5.00, 5.00)
		RCS	0.34 (0.33, 0.35)	1.20 (1.19, 1.20)	1.37 (1.37, 1.38)	1.05 (1.05, 1.05)
		FP	0.44 (0.43, 0.46)	0.87 (0.87, 0.88)	0.61 (0.61, 0.62)	0.34 (0.33, 0.35)
	7.5	LIN	0.41 (0.39, 0.42)	2.83 (2.83, 2.84)	8.33 (8.33, 8.33)	9.73 (9.73, 9.73)
		CAT3	2.24 (2.23, 2.24)	3.70 (3.70, 3.70)	8.17 (8.17, 8.17)	8.09 (8.09, 8.09)
		CAT5	1.53 (1.52, 1.53)	2.30 (2.30, 2.30)	5.09 (5.09, 5.09)	5.03 (5.02, 5.03)
		RCS	0.52 (0.51, 0.54)	1.26 (1.26, 1.27)	1.43 (1.43, 1.44)	1.12 (1.12, 1.13)

		FP	0.69 (0.67, 0.71)	1.00 (0.99, 1.01)	0.76 (0.74, 0.77)	0.59 (0.56, 0.61)
1000	2.5	LIN	0.09 (0.09, 0.10)	2.72 (2.72, 2.72)	8.13 (8.13, 8.13)	9.46 (9.46, 9.46)
		CAT3	2.16 (2.16, 2.16)	3.58 (3.58, 3.58)	7.87 (7.87, 7.87)	7.81 (7.81, 7.81)
		CAT5	1.33 (1.33, 1.33)	2.09 (2.09, 2.09)	4.83 (4.83, 4.83)	4.85 (4.85, 4.85)
		RCS	0.12 (0.12, 0.12)	1.21 (1.21, 1.21)	1.18 (1.18, 1.18)	1.03 (1.03, 1.03)
		FP	0.15 (0.14, 0.15)	0.78 (0.77, 0.78)	0.51 (0.50, 0.51)	0.12 (0.12, 0.12)
	5.0	LIN	0.19 (0.18, 0.19)	2.72 (2.72, 2.72)	8.14 (8.14, 8.14)	9.46 (9.46, 9.46)
		CAT3	2.17 (2.16, 2.17)	3.59 (3.58, 3.59)	7.87 (7.87, 7.87)	7.82 (7.81, 7.82)
		CAT5	1.36 (1.36, 1.36)	2.11 (2.11, 2.11)	4.84 (4.84, 4.84)	4.86 (4.86, 4.86)
		RCS	0.25 (0.24, 0.26)	1.23 (1.23, 1.23)	1.20 (1.19, 1.20)	1.06 (1.05, 1.06)
		FP	0.32 (0.32, 0.33)	0.82 (0.82, 0.82)	0.57 (0.56, 0.57)	0.25 (0.24, 0.26)
	7.5	LIN	0.28 (0.27, 0.29)	2.73 (2.73, 2.73)	8.14 (8.14, 8.14)	9.47 (9.47, 9.47)

		CAT3	2.18 (2.18, 2.18)	3.60 (3.59, 3.60)	7.87 (7.87, 7.87)	7.82 (7.82, 7.82)
		CAT5	1.41 (1.40, 1.41)	2.15 (2.14, 2.15)	4.86 (4.86, 4.86)	4.87 (4.87, 4.87)
		RCS	0.36 (0.35, 0.37)	1.26 (1.25, 1.26)	1.22 (1.22, 1.23)	1.09 (1.08, 1.09)
		FP	0.48 (0.47, 0.49)	0.87 (0.86, 0.88)	0.65 (0.65, 0.66)	0.37 (0.36, 0.39)
5000	2.5	LIN	0.04 (0.04, 0.04)	2.86 (2.86, 2.86)	8.26 (8.26, 8.26)	9.78 (9.78, 9.78)
		CAT3	2.21 (2.21, 2.21)	3.73 (3.73, 3.73)	8.03 (8.03, 8.03)	8.01 (8.01, 8.01)
		CAT5	1.34 (1.34, 1.35)	2.20 (2.20, 2.20)	4.97 (4.97, 4.97)	4.98 (4.98, 4.98)
		RCS	0.06 (0.05, 0.06)	1.18 (1.18, 1.18)	1.39 (1.39, 1.39)	1.02 (1.02, 1.02)
		FP	0.07 (0.07, 0.07)	0.79 (0.79, 0.79)	0.51 (0.51, 0.51)	0.06 (0.05, 0.06)
	5.0	LIN	0.08 (0.08, 0.09)	2.86 (2.86, 2.86)	8.26 (8.26, 8.26)	9.78 (9.78, 9.78)
		CAT3	2.21 (2.21, 2.21)	3.73 (3.73, 3.73)	8.03 (8.03, 8.03)	8.01 (8.01, 8.01)
		CAT5	1.35 (1.35, 1.35)	2.20 (2.20, 2.20)	4.97 (4.97, 4.97)	4.98 (4.98, 4.98)

		RCS	0.11 (0.10, 0.11)	1.18 (1.18, 1.18)	1.39 (1.39, 1.39)	1.03 (1.03, 1.03)
		FP	0.13 (0.13, 0.13)	0.80 (0.80, 0.80)	0.52 (0.52, 0.53)	0.11 (0.10, 0.11)
	7.5	LIN	0.13 (0.12, 0.13)	2.86 (2.86, 2.86)	8.26 (8.26, 8.26)	9.78 (9.78, 9.78)
		CAT3	2.22 (2.22, 2.22)	3.73 (3.73, 3.73)	8.03 (8.03, 8.03)	8.01 (8.01, 8.01)
		CAT5	1.36 (1.36, 1.36)	2.21 (2.21, 2.21)	4.98 (4.98, 4.98)	4.99 (4.99, 4.99)
		RCS	0.16 (0.16, 0.17)	1.19 (1.18, 1.19)	1.39 (1.39, 1.39)	1.04 (1.03, 1.04)
		FP	0.21 (0.20, 0.21)	0.81 (0.81, 0.82)	0.54 (0.54, 0.54)	0.16 (0.16, 0.17)
10000	2.5	LIN	0.03 (0.03, 0.03)	2.84 (2.84, 2.84)	8.23 (8.23, 8.23)	9.72 (9.72, 9.72)
		CAT3	2.20 (2.20, 2.20)	3.71 (3.71, 3.71)	8.22 (8.22, 8.22)	8.17 (8.17, 8.17)
		CAT5	1.32 (1.32, 1.32)	2.13 (2.13, 2.13)	4.96 (4.96, 4.96)	4.96 (4.96, 4.96)
		RCS	0.04 (0.04, 0.04)	1.15 (1.15, 1.15)	1.41 (1.41, 1.41)	1.02 (1.02, 1.02)
		FP	0.05 (0.04, 0.05)	0.79 (0.79, 0.79)	0.50 (0.50, 0.50)	0.04 (0.04, 0.04)

	5.0	LIN	0.06 (0.06, 0.06)	2.84 (2.84, 2.84)	8.23 (8.23, 8.24)	9.72 (9.72, 9.72)
		CAT3	2.20 (2.20, 2.20)	3.71 (3.71, 3.71)	8.22 (8.22, 8.22)	8.17 (8.17, 8.17)
		CAT5	1.32 (1.32, 1.32)	2.14 (2.14, 2.14)	4.96 (4.96, 4.96)	4.96 (4.96, 4.96)
		RCS	0.08 (0.07, 0.08)	1.16 (1.16, 1.16)	1.41 (1.41, 1.41)	1.03 (1.03, 1.03)
		FP	0.09 (0.09, 0.10)	0.79 (0.79, 0.79)	0.51 (0.51, 0.51)	0.08 (0.07, 0.08)
	7.5	LIN	0.09 (0.08, 0.09)	2.84 (2.84, 2.84)	8.24 (8.24, 8.24)	9.72 (9.72, 9.72)
		CAT3	2.20 (2.20, 2.20)	3.71 (3.71, 3.71)	8.22 (8.22, 8.22)	8.17 (8.17, 8.17)
		CAT5	1.32 (1.32, 1.32)	2.14 (2.14, 2.14)	4.96 (4.96, 4.96)	4.96 (4.96, 4.96)
		RCS	0.11 (0.11, 0.12)	1.16 (1.16, 1.16)	1.41 (1.41, 1.41)	1.03 (1.03, 1.03)
		FP	0.14 (0.14, 0.14)	0.80 (0.80, 0.80)	0.52 (0.52, 0.52)	0.11 (0.11, 0.12)

Table 4.4: The proportion of times the test of linearity was rejected in 1000 simulations under nonlinear association datasets fitted using FPs and RCS models (assuming different numbers of observations and noise's (σ)).

True association functions	Methods of Analysis	Sigma (σ)	Number of observations				
			200	500	1000	5 000	10 000
Linear piecewise threshold	FP	2.5	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000
	FP	5.0	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000
	FP	7.5	0.997	1.000	1.000	1.000	1.000
	RCS		0.986	1.000	1.000	1.000	1.000
Nonlinear piecewise	FP	2.5	1.000	1.000	1.000	1.000	1.000

threshold	RCS		1.000	1.000	1.000	1.000	1.000
	FP	5.0	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000
	FP	7.5	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000
Quadratic	FP	2.5	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000
	FP	5.0	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000
	FP	7.5	1.000	1.000	1.000	1.000	1.000
	RCS		1.000	1.000	1.000	1.000	1.000

Table 4.5: Comparison of the average optimal alcohol intake and BP estimates obtained across 1000 simulations (replications) after fitting linear threshold association datasets using different regression models. The 95% CI regions for the estimates are given in brackets. Datasets of various sample sizes ($n=200, 500, 1000, 5000, 10\ 000$) and standard deviations ($\sigma = 2.5, 5.0, 7.5$) are considered.

Linear threshold association datasets			Estimates		
			Optimal alcohol intake	BP at optimum alcohol intake	BP at intake = 20 grams
Sample sizes	Std. dev	Methods			
200	2.5	CAT3	0-23	121.13 (120.52, 121.67)	121.13 (120.52, 121.67)
		CAT5	0-13	121.02 (120.24, 121.73)	122.73 (121.98, 123.46)
		RCS	0 (0, 0)	118.38 (117.26, 119.35)	123.82 (123.38, 124.27)
		FP	8 (0, 9)	119.82 (118.94, 120.66)	123.15 (122.61, 123.68)
	5.0	CAT3	0-23	121.13 (120.01, 122.26)	121.13 (120.01, 122.26)
		CAT5	0-23	121.00 (119.52, 122.39)	122.73 (121.14, 124.11)
		RCS	0 (0, 0)	118.36 (116.27, 120.39)	123.81 (122.88, 124.80)

		FP	6 (0, 11)	119.73 (117.95, 121.28)	123.16 (122.09, 124.11)
	7.5	CAT3	0-23	121.18 (119.36, 122.98)	121.18 (119.36, 122.98)
		CAT5	0-25	120.91 (118.62, 123.04)	122.83 (120.32, 125.07)
		RCS	0 (0, 0)	118.43 (115.11, 121.81)	123.87 (122.43, 125.27)
		FP	0 (6, 12)	119.65 (108.80, 121.77)	123.21 (121.46, 124.69)
500	2.5	CAT3	1-20	120.99 (120.63, 121.37)	120.99 (120.63, 121.37)
		CAT5	1-11	120.98 (120.54, 121.49)	121.98 (121.53, 122.47)
		RCS	1 (1, 1)	119.28 (118.77, 119.86)	123.87 (123.55, 124.16)
		FP	1 (1, 9)	119.94 (119.55, 120.37)	123.30 (122.90, 123.61)
	5.0	CAT3	1-20	121.05 (120.21, 121.77)	121.05 (120.21, 121.77)
		CAT5	1-22	121.03 (120.01, 121.93)	122.02 (121.08, 122.92)
		RCS	1 (1, 1)	119.34 (118.07, 120.53)	123.88 (123.28, 124.45)

		FP	1 (1, 9)	119.97 (119.06, 120.75)	123.25 (122.51, 123.83)
	7.5	CAT3	1-20	120.99 (119.89, 122.18)	120.99 (119.89, 122.18)
		CAT5	1-24	120.92 (119.60, 122.24)	121.99 (120.57, 123.49)
		RCS	1 (1, 1)	119.33 (117.68, 121.16)	123.82 (122.98, 124.76)
		FP	1 (1, 10)	119.86 (118.58, 121.05)	123.18 (122.17, 124.11)
1000	2.5	CAT3	0-21	121.03 (120.79, 121.31)	121.03 (120.79, 121.31)
		CAT5	0-13	121.00 (120.71, 121.33)	121.90 (121.57, 125.26)
		RCS	0 (0, 0)	118.68 (118.30, 119.17)	123.85 (123.66, 124.05)
		FP	8 (0, 9)	119.76 (119.41, 120.32)	123.21 (122.91, 123.49)
	5.0	CAT3	0-21	121.01 (120.49, 121.53)	121.01 (120.49, 121.53)
		CAT5	0-16	120.98 (120.29, 121.60)	121.90 (121.20, 122.58)
		RCS	0 (0, 0)	118.70 (117.69, 119.62)	123.84 (123.40, 124.24)

		FP	8 (0, 9)	119.71 (119.11, 120.50)	123.15 (122.63, 123.63)
	7.5	CAT3	0-21	121.05 (120.25, 121.87)	121.05 (120.25, 121.87)
		CAT5	0-23	120.99 (120.00, 121.89)	121.92 (120.85, 123.03)
		RCS	0 (0, 0)	118.76 (117.39, 120.11)	123.84 (123.23, 124.51)
		FP	8 (0, 10)	119.79 (118.84, 120.77)	123.17 (122.43, 123.83)
5000	2.5	CAT3	0-20	121.00 (120.87, 121.12)	121.00 (120.87, 121.12)
		CAT5	0-12	121.00 (120.84, 121.14)	121.68 (121.52, 121.84)
		RCS	0 (0, 0)	119.14 (118.94, 119.34)	123.72 (123.62, 123.82)
		FP	8 (0, 8)	119.91 (119.59, 120.06)	123.17 (123.04, 123.43)
	5.0	CAT3	0-20	121.00 (120.77, 121.23)	121.00 (120.77, 121.23)
		CAT5	0-12	121.00 (120.70, 121.24)	121.68 (121.37, 121.99)
		RCS	0 (0, 0)	119.15 (118.79, 119.50)	123.73 (123.52, 123.92)

		FP	7 (0, 8)	119.88 (119.50, 120.17)	123.20 (122.94, 123.51)
	7.5	CAT3	0-20	121.01 (120.67, 121.34)	121.01 (120.67, 121.34)
		CAT5	0-12	121.01 (120.56, 121.42)	121.67 (121.20, 122.17)
		RCS	0 (0, 0)	119.16 (118.61, 119.70)	123.72 (123.44, 123.99)
		FP	7 (0, 8)	119.84 (119.37, 120.28)	123.22 (122.85, 123.58)
10000	2.5	CAT3	0-20	121.00 (120.92, 121.09)	121.00 (120.92, 121.09)
		CAT5	0-12	121.00 (120.90, 121.12)	121.66 (121.55, 121.76)
		RCS	0 (0, 0)	119.12 (118.98, 119.27)	123.75 (123.68, 123.82)
		FP	8 (0, 8)	119.97 (119.64, 120.06)	123.16 (123.08, 123.40)
	5.0	CAT3	0-20	121.00 (120.82, 121.17)	121.00 (120.82, 121.17)
		CAT5	0-12	120.99 (120.79, 121.22)	121.65 (121.44, 121.87)
		RCS	0 (0, 0)	119.11 (118.83, 119.40)	123.75 (123.62, 123.89)

		FP	8 (0, 8)	119.92 (119.55, 120.13)	123.19 (123.01, 123.47)
	7.5	CAT3	0-20	121.01 (120.75, 121.24)	121.01 (120.75, 121.24)
		CAT5	0-12	121.01 (120.69, 121.33)	121.65 (121.31, 121.97)
		RCS	0 (0, 0)	119.13 (118.71, 119.53)	123.75 (123.55, 123.95)
		FP	8 (0, 8)	119.90 (119.46, 120.21)	123.21 (122.93, 123.51)

Table 4.6: Comparison of the average optimal alcohol intake and BP estimates obtained across 1000 simulations (replications) after fitting nonlinear threshold association datasets using different regression models. The 95% CI regions for the estimates are given in brackets. Datasets of various sample sizes ($n=200, 500, 1000, 5000, 10\ 000$) and standard deviations ($\sigma = 2.5, 5.0, 7.5$) are considered.

Nonlinear threshold association datasets			Estimates		
			Optimal alcohol intake	BP at optimum alcohol intake	BP at intake = 20 grams
Sample sizes	Std. dev	Methods			
200	2.5	CAT3	0-23	121.02 (120.41, 121.56)	121.02 (121.41, 121.56)
		CAT5	0-26	120.97 (120.22, 121.58)	121.46 (120.71, 122.19)
		RCS	18 (16, 19)	120.01 (119.58, 120.45)	120.12 (119.68, 120.57)
		FP	15 (11, 17)	120.21 (119.67, 120.87)	120.68 (120.08, 121.51)
	5.0	CAT3	0-23	121.02 (119.90, 122.15)	121.02 (119.90, 122.15)
		CAT5	0-27	120.76 (119.43, 122.01)	121.45 (119.87, 122.83)

		RCS	18 (15, 20)	119.98 (119.06, 120.88)	120.11 (119.18, 121.10)
		FP	15 (6, 18)	120.20 (119.00, 121.43)	120.74 (119.63, 122.05)
	7.5	CAT3	0-23	121.07 (119.25, 122.87)	121.07 (119.25, 122.87)
		CAT5	0-27	120.62 (118.51, 122.59)	121.55 (119.05, 123.80)
		RCS	18 (12, 21)	120.02 (118.56, 121.34)	120.17 (118.73, 121.57)
		FP	14 (0, 18)	120.20 (117.90, 121.96)	120.83 (118.99, 122.46)
500	2.5	CAT3	1-20	120.99 (120.63, 121.37)	120.99 (120.63, 121.37)
		CAT5	1-24	120.94 (120.53, 121.30)	121.17 (120.73, 121.67)
		RCS	18 (17, 18)	119.46 (119.15, 119.75)	119.60 (119.28, 119.89)
		FP	15 (12, 16)	120.07 (119.74, 120.55)	120.53 (120.20, 121.25)
	5.0	CAT3	1-20	121.05 (120.21, 121.77)	121.05 (120.21, 121.77)
		CAT5	1-25	120.85 (119.98, 121.60)	121.22 (120.27, 122.11)

		RCS	18 (16, 19)	119.46 (118.88, 119.99)	119.62 (119.01, 120.18)
		FP	15 (10, 17)	120.14 (119.43, 120.87)	120.67 (119.95, 121.62)
	7.5	CAT3	1-20	120.99 (119.89, 122.18)	120.99 (119.89, 122.18)
		CAT5	1-25	120.71 (119.49, 121.77)	121.18 (119.77, 122.68)
		RCS	18 (16, 19)	119.41 (118.59, 120.29)	119.55 (118.71, 120.49)
		FP	15 (7, 17)	120.10 (119.02, 121.22)	120.67 (119.67, 121.93)
1000	2.5	CAT3	0-21	121.00 (120.76, 121.28)	121.00 (120.76, 121.28)
		CAT5	0-24	120.96 (120.69, 121.24)	121.16 (120.82, 121.51)
		RCS	18 (17, 18)	119.72 (119.54, 119.91)	119.84 (119.65, 120.05)
		FP	15 (13, 16)	120.12 (119.86, 120.48)	120.58 (120.32, 121.13)
	5.0	CAT3	0-21	120.98 (120.46, 121.50)	120.98 (120.46, 121.50)
		CAT5	0-25	120.87 (120.23, 121.38)	121.15 (120.45, 121.83)

		RCS	18 (17, 19)	119.71 (119.28, 120.08)	119.83 (119.39, 120.23)
		FP	15 (11, 16)	120.14 (119.66, 120.75)	120.66 (120.10, 121.71)
	7.5	CAT3	0-21	121.03 (120.22, 121.84)	121.03 (120.22, 121.84)
		CAT5	0-25	120.82 (119.89, 121.65)	121.17 (121.10, 122.28)
		RCS	18 (16, 19)	119.72 (119.08, 120.36)	119.84 (119.22, 120.50)
		FP	15 (10, 17)	120.23 (119.46, 121.07)	120.80 (119.99, 121.71)
5000	2.5	CAT3	0-20	121.00 (120.87, 121.12)	121.00 (120.87, 121.12)
		CAT5	0-23	120.99 (120.84, 121.12)	121.10 (120.94, 121.26)
		RCS	18 (17, 18)	119.45 (119.35, 119.54)	119.59 (119.49, 119.69)
		FP	14 (13, 15)	120.19 (119.90, 120.33)	120.88 (120.38, 121.01)
	5.0	CAT3	0-20	121.00 (120.77, 121.23)	121.00 (120.77, 121.23)
		CAT5	0-24	120.96 (120.68, 121.19)	121.10 (120.79, 121.41)

		RCS	18 (17, 18)	119.45 (119.25, 119.64)	119.60 (119.39, 119.79)
		FP	14 (13, 16)	120.13 (119.84, 120.42)	120.81 (120.30, 121.11)
	7.5	CAT3	0-20	121.00 (120.67, 121.34)	121.00 (120.67, 121.34)
		CAT5	0-24	120.91 (120.54, 121.26)	121.09 (120.62, 121.59)
		RCS	18 (17, 18)	119.44 (119.18, 119.72)	119.59 (119.31, 119.86)
		FP	14 (11, 16)	120.11 (119.77, 120.61)	120.72 (120.26, 121.39)
10000	2.5	CAT3	0-20	121.00 (120.92, 121.09)	121.00 (120.92, 121.09)
		CAT5	0-22	120.99 (120.90, 121.10)	121.10 (120.99, 121.20)
		RCS	18 (17, 18)	119.46 (119.40, 119.53)	119.60 (119.54, 119.67)
		FP	14 (13, 15)	120.20 (119.92, 120.30)	120.88 (120.39, 120.98)
	5.0	CAT3	0-20	121.00 (120.82, 121.17)	121.00 (120.82, 121.17)
		CAT5	0-23	120.97 (120.78, 121.15)	121.09 (120.88, 121.31)

		RCS	18 (17, 18)	119.46 (119.33, 119.59)	119.60 (119.47, 119.74)
		FP	14 (13, 16)	120.14 (119.86, 120.36)	120.81 (120.35, 121.02)
	7.5	CAT3	0-20	121.00 (120.75, 121.24)	121.00 (120.75, 121.24)
		CAT5	0-24	120.96 (120.66, 121.20)	121.08 (120.75, 121.41)
		RCS	18 (17, 18)	119.46 (119.26, 119.65)	119.60 (119.40, 119.81)
		FP	14 (13, 16)	120.12 (119.81, 120.44)	120.77 (120.27, 121.12)

Table 4.7: Comparison of the average optimal alcohol intake and BP estimates obtained across 1000 simulations (replications) after fitting the Quadratic association datasets using different regression models. The 95% CI regions for the estimates are given in brackets. Datasets of various sample sizes ($n=200, 500, 1000, 5000, 10\ 000$) and standard deviations ($\sigma = 2.5, 5.0, 7.5$) are considered.

Quadratic association datasets			Estimates		
			Optimal alcohol intake	BP at optimum alcohol intake	BP at intake = 20 grams
Sample sizes	Std. dev	Methods			
200	2.5	CAT3	0-23	123.05 (122.44, 123.59)	123.05 (122.44, 123.59)
		CAT5	14-27	120.24 (119.49, 120.97)	120.24 (119.49, 120.97)
		RCS	23 (22, 23)	119.99 (119.53, 120.46)	120.24 (119.80, 120.70)
		FP	20 (19, 21)	119.61 (119.13, 120.16)	119.61 (119.13, 120.17)
	5.0	CAT3	0-23	123.05 (121.93, 123.59)	123.05 (121.93, 123.59)
		CAT5	14-27	120.23 (118.65, 121.61)	120.23 (118.65, 121.61)

		RCS	23 (22, 24)	119.97 (118.99, 121.01)	120.23 (119.30, 121.22)
		FP	20 (18, 21)	119.65 (118.46, 121.04)	119.67 (118.53, 121.08)
	7.5	CAT3	0-23	123.10 (121.28, 124.90)	123.10 (121.28, 124.90)
		CAT5	14-27	120.33 (117.81, 122.52)	120.33 (117.83, 122.58)
		RCS	23 (21, 24)	120.00 (118.44, 121.46)	120.29 (118.85, 121.69)
		FP	20 (17, 22)	119.74 (117.86, 121.51)	119.79 (117.91, 121.57)
500	2.5	CAT3	1-33	124.50 (124.14, 124.83)	124.50 (124.14, 124.88)
		CAT5	12-25	120.27 (119.82, 120.77)	120.27 (119.82, 120.77)
		RCS	23 (22, 23)	119.61 (119.29, 119.93)	119.88 (119.56, 120.17)
		FP	20 (20, 20)	119.61 (119.29, 119.91)	119.61 (119.29, 119.91)
	5.0	CAT3	1-39	124.50 (123.73, 125.05)	124.56 (123.73, 125.28)
		CAT5	12-25	120.31 (119.37, 121.21)	120.31 (119.37, 121.21)

		RCS	23 (22, 23)	119.63 (118.97, 120.24)	119.90 (119.29, 120.46)
		FP	20 (19, 21)	119.64 (118.98, 120.46)	119.64 (118.99, 120.48)
	7.5	CAT3	1-39	124.36 (123.37, 125.29)	124.50 (123.40, 125.69)
		CAT5	12-25	120.28 (118.87, 121.78)	120.28 (118.87, 121.78)
		RCS	23 (22, 23)	119.56 (118.67, 120.55)	119.83 (118.99, 120.77)
		FP	20 (18, 21)	119.61 (118.48, 121.00)	119.62 (118.55, 121.06)
1000	2.5	CAT3	0-21	123.73 (123.49, 124.01)	123.73 (123.49, 124.01)
		CAT5	14-25	120.07 (119.73, 120.42)	120.07 (119.73, 120.42)
		RCS	23 (23, 23)	119.76 (119.54, 119.97)	120.04 (119.85, 120.25)
		FP	20 (20, 20)	119.60 (119.41, 119.81)	119.60 (119.41, 119.81)
	5.0	CAT3	0-21	123.72 (123.19, 124.23)	123.72 (123.19, 124.23)
		CAT5	14-25	120.06 (119.36, 120.74)	120.06 (119.36, 120.74)

		RCS	23 (22, 23)	119.74 (119.25, 120.19)	120.03 (119.59, 120.43)
		FP	20 (19, 20)	119.59 (119.14, 120.07)	119.59 (119.14, 120.07)
	7.5	CAT3	0-21	123.76 (122.96, 124.57)	123.76 (122.96, 124.57)
		CAT5	14-25	120.08 (119.01, 121.19)	120.08 (119.01, 121.19)
		RCS	23 (22, 23)	119.75 (119.10, 120.45)	120.04 (119.42, 120.71)
		FP	20 (19, 21)	119.63 (118.92, 120.56)	119.64 (118.93, 120.56)
5000	2.5	CAT3	0-20	124.36 (124.23, 124.48)	124.36 (124.23, 124.48)
		CAT5	13-24	120.11 (119.95, 120.27)	120.11 (119.95, 120.27)
		RCS	23 (23, 23)	119.45 (119.35, 119.56)	119.74 (119.65, 119.84)
		FP	20 (20, 20)	119.60 (119.50, 119.70)	119.60 (119.50, 119.70)
	5.0	CAT3	0-22	124.36 (124.13, 124.59)	124.36 (124.13, 124.60)
		CAT5	13-24	120.12 (119.80, 120.42)	120.12 (119.80, 120.42)

		RCS	23 (23, 23)	119.46 (119.23, 119.66)	119.75 (119.55, 119.94)
		FP	20 (20, 20)	119.60 (119.40, 119.80)	119.60 (119.40, 119.80)
	7.5	CAT3	0-35	124.36 (124.02, 124.65)	124.37 (124.03, 124.71)
		CAT5	13-25	120.10 (119.64, 120.60)	120.10 (119.64, 120.60)
		RCS	23 (23, 23)	119.45 (119.15, 119.74)	119.74 (119.46, 120.01)
		FP	20 (20, 20)	119.60 (119.31, 119.87)	119.60 (119.31, 119.87)
10000	2.5	CAT3	0-20	124.32 (124.23, 124.41)	124.32 (124.23, 124.41)
		CAT5	13-24	120.19 (120.01, 120.22)	120.19 (120.01, 120.22)
		RCS	23 (23, 23)	119.50 (119.43, 119.58)	119.79 (119.72, 119.86)
		FP	20 (20, 20)	119.60 (119.54, 119.67)	119.60 (119.54, 119.67)
	5.0	CAT3	0-20	124.32 (124.14, 124.49)	124.32 (124.14, 124.49)
		CAT5	13-24	120.12 (119.90, 120.34)	120.12 (119.90, 120.34)

		RCS	23 (23, 23)	119.50 (119.36, 119.65)	119.79 (119.65, 119.92)
		FP	20 (20, 20)	119.60 (119.47, 119.74)	119.60 (119.47, 119.74)
	7.5	CAT3	0-20	124.32 (124.07, 124.55)	124.32 (124.07, 124.56)
		CAT5	13-24	120.11 (119.77, 120.43)	120.11 (119.77, 120.43)
		RCS	23 (23, 23)	119.50 (119.28, 119.71)	119.78 (119.58, 119.99)
		FP	20 (20, 20)	119.60 (119.40, 119.80)	119.60 (119.40, 119.80)

Appendix D

D.1 Additional tables in Chapter 5

Table 5.6: Comparison of net benefits and reduction of false positive results per 100 patients according to different statistical models in linear threshold datasets assuming various threshold probabilities.

Threshold probabilities (%)	Treat all	Categorisation (3 groups)			Categorisation (5 groups)			Linearisation		
		Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients
0.05	0.05	0.06 (0.04, 0.08)	0.00 (0.00, 0.01)	0 (0, 13)	0.06 (0.04, 0.08)	0.00 (0.00, 0.01)	4 (0, 16)	0.06 (0.04, 0.07)	0.00 (0.00, 0.01)	1 (0, 15)
0.10	0.00	0.03 (0.01, 0.04)	0.03 (0.01, 0.04)	24 (12, 36)	0.03 (0.02, 0.05)	0.03 (0.01, 0.04)	24 (13, 36)	0.03 (0.01, 0.04)	0.03 (0.01, 0.04)	24 (12, 36)
0.15	-0.06	0.01 (0.00, 0.03)	0.07 (0.05, 0.08)	38 (29, 46)	0.01 (0.00, 0.01)	0.07 (0.05, 0.09)	39 (29, 49)	0.01 (0.00, 0.03)	0.07 (0.05, 0.09)	39 (29, 49)
0.20	-0.12	0.00	0.12 (0.10, 0.14)	49 (41, 58)	0.00	0.13 (0.11, 0.15)	50 (42, 58)	0.00	0.13 (0.10, 0.15)	50 (42, 59)
0.25	-0.20	0.00	0.20 (0.17, 0.22)	59 (52, 66)	0.00	0.20 (0.17, 0.22)	59 (52, 66)	0.00	0.20 (0.17, 0.22)	59 (52, 66)

Fractional Polynomials			Restricted cubic splines		
Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients
0.06 (0.04, 0.08)	0.00 (0.00, 0.01)	2 (0, 16)	0.06 (0.04, 0.08)	0.00 (0.00, 0.01)	0 (0, 15)
0.03 (0.01, 0.05)	0.03 (0.01, 0.04)	24 (13, 36)	0.03 (0.01, 0.05)	0.03 (0.01, 0.04)	24 (12, 36)
0.01 (0.00, 0.03)	0.07 (0.05, 0.09)	39 (29, 49)	0.01 (0.00, 0.03)	0.07 (0.05, 0.09)	39 (29, 49)
0.00	0.13 (0.10, 0.15)	50 (42, 59)	0.00	0.13 (0.10, 0.15)	50 (42, 59)
0.00	0.20 (0.17, 0.22)	59 (52, 66)	0.00	0.20 (0.17, 0.22)	59 (52, 66)

Table 5.7: Comparison of net benefits and reduction of false positive results per 100 patients according to different statistical models in nonlinear threshold datasets assuming various threshold probabilities.

Threshold probabilities (%)	Treat all	Categorisation (3 groups)			Categorisation (5 groups)			Linearisation		
		Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients
0.05	0.05	0.06 (0.04, 0.07)	0.00 (0.00, 0.01)	1 (0, 15)	0.06 (0.04, 0.07)	0.00 (0.00, 0.01)	4 (0, 18)	0.05 (0.04, 0.07)	0.00 (-0.01, 0.01)	0 (0, 17)
0.10	0.00	0.03 (0.02, 0.05)	0.03 (0.02, 0.04)	29 (17, 40)	0.03 (0.02, 0.05)	0.03 (0.02, 0.05)	29 (16, 41)	0.03 (0.02, 0.05)	0.03 (0.01, 0.04)	27 (14, 40)
0.15	-0.06	0.02 (0.00, 0.03)	0.07 (0.06, 0.09)	41 (33, 49)	0.02 (0.01, 0.03)	0.08 (0.06, 0.09)	45 (35, 53)	0.02 (0.01, 0.03)	0.08 (0.06, 0.09)	44 (34, 53)
0.20	-0.12	0.00	0.13 (0.11, 0.14)	50 (43, 58)	0.01 (0.00, 0.02)	0.13 (0.12, 0.15)	54 (46, 61)	0.01 (0.00, 0.03)	0.13 (0.12, 0.15)	54 (46, 61)
0.25	-0.20	0.00	0.20 (0.17, 0.22)	60 (52, 66)	0.00	0.20 (0.18, 0.22)	60 (54, 67)	0.00	0.20 (0.18, 0.22)	60 (54, 67)
0.30	-0.28	0.00	0.28 (0.26, 0.31)	66 (60, 72)	0.00	0.28 (0.26, 0.31)	66 (60, 72)	0.00	0.28 (0.26, 0.31)	66 (60, 72)
0.35	-0.38	0.00	0.38 (0.35, 0.41)	71 (66, 76)	0.00	0.38 (0.35, 0.41)	71 (66, 76)	0.00	0.38 (0.35, 0.41)	71 (66, 76)

Fractional Polynomials			Restricted cubic splines		
Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 100 patients
0.06 (0.04, 0.07)	0.00 (0.00, 0.01)	2 (0, 18)	0.06 (0.04, 0.07)	0.00 (0.00, 0.01)	1 (0, 18)
0.03 (0.02, 0.05)	0.03 (0.02, 0.05)	30 (17, 43)	0.03 (0.02, 0.05)	0.03 (0.02, 0.05)	29 (16, 41)
0.02 (0.01, 0.04)	0.08 (0.06, 0.09)	45 (35, 54)	0.02 (0.01, 0.03)	0.08 (0.06, 0.09)	45 (34, 53)
0.01 (0.00, 0.03)	0.14 (0.12, 0.15)	54 (46, 62)	0.01 (0.00, 0.02)	0.14 (0.12, 0.15)	54 (46, 62)
0.01 (0.00, 0.02)	0.20 (0.18, 0.23)	61 (55, 68)	0.01 (0.00, 0.02)	0.20 (0.18, 0.23)	61 (55, 68)
0.00	0.29 (0.26, 0.31)	67 (61, 73)	0.00	0.29 (0.26, 0.31)	67 (61, 72)
0.00	0.38 (0.36, 0.41)	71 (66, 76)	0.00	0.38 (0.36, 0.41)	71 (66, 76)

Table 5.8: Comparison of net benefits and reduction of false positive results per 100 patients according to different statistical models in quadratic datasets assuming various threshold probabilities.

Threshold probabilities (%)	Treat all	Categorisation (3 groups)			Categorisation (5 groups)			Linearisation		
		Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 1000 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Predicted model (Net Benefit)	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 1000 patients
0.05	0.05	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 0)	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 2)	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 0)
0.10	0.00	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)	7 (0, 22)	0.01 (0.00, 0.02)	0.01 (0.00, 0.03)	9 (0, 23)	0.01 (0.00, 0.02)	0.01 (0.00, 0.02)	6 (0, 22)
0.15	-0.06	0.00	0.06 (0.04, 0.08)	33 (21, 46)	0.00	0.06 (0.04, 0.08)	34 (22, 46)	0.00	0.06 (0.04, 0.08)	33 (21, 46)
0.20	-0.13	0.00	0.13 (0.10, 0.15)	50 (41, 60)	0.00	0.13 (0.10, 0.15)	50 (41, 60)	0.00	0.13 (0.10, 0.15)	50 (46, 60)

Fractional Polynomials			Restricted cubic splines		
Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 1000 patients	Predicted model (Net Benefit)	Difference (Net Benefit)	Reduction in false positives per 1000 patients
0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 3)	0.05 (0.03, 0.07)	0.00 (0.00, 0.00)	0 (0, 0)
0.01 (0.00, 0.02)	0.01 (0.00, 0.03)	8 (0, 23)	0.01 (0.00, 0.02)	0.01 (0.00, 0.03)	8 (0, 23)
0.00	0.06 (0.04, 0.08)	34 (22, 47)	0.00	0.06 (0.04, 0.08)	34 (21, 46)
0.00	0.13 (0.10, 0.15)	50 (41, 60)	0.00	0.13 (0.10, 0.15)	50 (41, 60)

Appendix E

E.1 Justification of assumed associations in the DAG

Introduction

An observation that excessive alcohol consumption is associated with hypertension was first reported in 1915. While caring for military personnel in World War 1, Lian reported higher BP amongst soldiers consuming >2.5 L of wine per day (Lian, 1915). However, these findings were largely ignored until the 1970s. In the last 45 years, several studies and reviews have been performed to confirm the presence of the alcohol-hypertension relationship (Jackson et al., 1985, Moreira et al., 1998, Fuchs et al., 2001, Saremi et al., 2004, Steffens et al., 2006, Klatsky and Gunderson, 2008, Briasoulis et al., 2012, Mori et al., 2016). However, there are unresolved issues about the direction, shape, and factors explaining the alcohol-hypertension relationship. This might be due to several factors associated with the development of hypertension. Hypertension is multifactorial, therefore the alcohol-hypertension relationship is likely to be confounded by many variables (Lip and Beevers, 2003). A suitable alcohol-hypertension relationship requires consideration and identification of confounding variables for adjustment. In Figure 6.1, the DAG was used to identify suitable confounding variables for the alcohol-hypertension relationships in type 2 diabetes patients. Clinical and epidemiological knowledge from the literature was used to develop and explain the causal relationships in the DAG. The choice of variables in the DAG and their assumed relationships with the exposure (alcohol consumption) and outcome (hypertension) were explained focusing on competing exposures, confounders, mediators, and colliders in the diagram.

Competing Exposures

Diet. In Figure 6.1, diet was entered in the DAG as a competing exposure. The association between diet and hypertension has been reported in epidemiological and clinical studies (Reddy and Katan, 2004, Bazzano et al., 2013, Duman, 2013). There is no evidence suggesting that diet changes alcohol consumption (or vice-versa). Existing studies report lower hypertension risk amongst vegetarians (low fat diets with higher nutrients including potassium, magnesium, and fibre) or low-salt diets. For example, in a trial study on Dietary Approaches to Stop Hypertension (DASH), diet with low salt (or sodium), saturated fats, cholesterol and high in potassium, fibre, magnesium, protein, fruits & vegetables and calcium was recommended to reduce the risk of hypertension (Moore et al., 2001, Sacks et al., 2001). An observation study suggesting similar findings as the DASH trial reveal a positive association between dietary salt intake and BP. A positive linear relationship between salt intake and BP was reported in Turkey where each 100 mmol/day of salt intake was associated with a 5.8 mmHg increase of SBP (Erdem et al., 2010). Furthermore, dietary fats are suggested as the key link in the causal pathway connecting diet to hypertension. Healthier fats including monounsaturated fats and polyunsaturated fats (omega 3 and omega 6) have been linked with no incidence of hypertension (Duman, 2013). These fats play an important role in balancing the cholesterol levels in the blood - reducing bad cholesterol (LDL-C) and increasing good cholesterol (HDL-C). Apart from these studies, diets with high potassium intake, protein, fibre and fruits, and vegetables are associated with lower incidence of hypertension (Reddy and Katan, 2004, Bazzano et al., 2013). A diet with rich potassium intake has been suggested to be more beneficial amongst patients with higher salt intake (Van Bommel and Cleophas, 2012). On the other hand, fibre consumption tends to lower the LDL-C concentration without affecting the HDL-C (Reddy and Katan, 2004). Based on this evidence, the DAG provided in Figure 6.1

summarises the assumed relationships between diet measures and hypertension. In the DAG, diet intake measures including potassium intake, protein, fats, salts, fibre, fruits, and vegetable were treated as competing exposures.

Another competing exposure appearing in the DAG is coffee consumption. In the diagram, coffee and alcohol consumption were assumed to be independent and both were linked with high incidence of hypertension as suggested in the literature (Uiterwaal et al., 2007, Klatsky and Gunderson, 2008, Briasoulis et al., 2012)

Mediators

Obesity. The prevention or treatment of obesity is recommended as an important means to reduce the risk of CDV in adults population (British Cardiovascular Society et al., 2014). Evidence also suggests a causal link between obesity and hypertension (Fall et al., 2013). Amongst alcohol drinkers, a significant additive interaction between alcohol consumption and obesity was associated with the incidence of hypertension (Li et al., 2006, Luo et al., 2013). However, the mechanism between alcohol consumption and obesity in relation to hypertension remains unclear (Li et al., 2006, Buja et al., 2009). In Figure 6.1, alcohol consumption was assumed to be causally linked to obesity due to the energy derived from alcohol drinking (extra energy is stored as fat). A non-caloric mechanism linking alcohol consumption to obesity might causally be related to changes in steroid hormones favouring fat storage (Buja et al., 2009).

Insulin resistance. In type 2 diabetes patients, heavy alcohol drinking has been linked with insulin resistance because of disruption of glucose homeostasis (Kim and Kim, 2012). In Figure 6.1, insulin resistance was associated with hyperinsulinemia which relates with increased LDL and reduced HDL concentrations. Also, through hyperinsulinemia, insulin resistance was associated with the occurrence of hypertension. The assumed causal pathways are supported by Barnett (Barnett, 1994).

Poor glucose control. Maintaining tight blood control of blood glucose in alcoholics with type 2 diabetes is difficult. Alcohol consumption is associated with both incidents of hypoglycaemia (low blood glucose levels) and hyperglycaemia (high blood glucose levels). Hypoglycaemia is mostly prevalent in heavy alcoholics with insufficient dietary intake of glucose whilst hyperglycaemia is common in well fed heavy alcohol drinkers (Emanuele et al., 1998). In the two types of diabetes, hypoglycaemia is causally linked to hyperinsulinemia (Kim and Kim, 2012). Hyperglycaemia has strong but reversible oxidation linkages with LDL concentrates which are associated with hypertension through the renin-angiotensin system (RAS) (Leslie, 1993, Ji et al., 2014).

Colliders

Kidney function. The responsibilities of kidneys include (1) filtering harmful substances from the blood and (2) regulating the right quantity of water inside the body (deRibeaux, 1997). Alcohol drinking might disturb kidneys making them less able to perform their functions. A significantly reduced kidney function in alcohol fed animals was reported in an experimental study (Van Thiel et al., 1977). Alternatively, the harm on kidney function might occur indirectly through alcohol-induced hypertension. In observational studies, excessive alcohol drinking has been linked to the occurrence of high blood pressure (Parekh and Klag, 2001). Hence, kidney function was captured as a collider (with both arcs from the exposure and outcome variables pointing into it) in Figure 6.1. Furthermore, atherosclerosis (lining of fats in the arteries), excessive dietary salts and poor glycaemic control in patients with type 2 diabetes were causally linked to the poor functioning of the kidneys or CKD (Chade et al., 2005, Bash et al., 2008, Farquhar et al., 2015). Figure 6.1 show pathway linkages between these variables.

Stroke. Evidence suggests the alcohol and hypertension are associated with the risk of stroke (Hillbom et al., 2011). The mechanism defining the relationship between alcohol consumption and stroke is less clear. Alcohol consumption might directly reduce the risk of ischaemic stroke through the HDL-C. In contrast, the alcohol's antithrombotic action (clotting) might accelerate the risk of haemorrhagic stroke (Klatsky and Gunderson, 2008). Moreover, hypertension has long been established as a powerful predictor for both ischaemic and haemorrhagic strokes (Klatsky and Gunderson, 2008, Hillbom et al., 2011). Hypertension is linked to the occurrence of stroke through the multifactorial causation of atherosclerosis (Kannel et al., 1996). The causal pathway diagram showing details of these relations is provided in Figure 6.1.

Confounders

Smoking. Evidence in epidemiology link heavy drinking to smoking (Lip and Beevers, 2003) therefore, smoking may influence the association between alcohol consumption and hypertension. A negative relationship between smoking and blood pressure was reported in general population - with high BP levels observed amongst non-smoker compared to smokers (Lee et al., 1998, Can et al., 2009, Alomari and Al-Sheyab, 2016). So, if alcohol drinkers are more likely to be smokers and smoking reduces blood pressure, then smoking is an important variable to consider for confounding – it may change the alcohol-hypertension relationship amongst type 2 diabetes patients.

Physical Activity. Broadly, physical activity (PA) is defined as “any body movement produced by the contraction of skeletal muscles that substantially increases energy expenditure” (Colberg et al., 2010, Stump, 2011). In the thesis, the word PA is used interchangeably with ‘exercise’. Exercise is defined as “a subset of PA and done

with the intension of developing physical fitness (i.e. improvements in cardiovascular function, strength, and flexibility)” (Colberg et al., 2010, Stump, 2011).

Evidence suggests that exercise improves many risk factors associated with cardiovascular disease (CVD). Amongst patients with diabetes, the positive effects of regular exercise include greater insulin sensitivity and improvements in common risk factors of inflammation, abnormal adiposity, dyslipidaemia and hypertension (Stessman and Jacobs, 2014). A recent meta-analysis recommends 150 min/week of regular PA in moderate intensity amongst young and middle-aged diabetes patients (Kodama et al., 2013). However, it is also important to note that the mechanism at which PA affects individuals could be different depending on the behaviour of individuals and the severity of the disease. In the literature, much is not known about risk management of hypertension in alcohol drinkers who are diabetic taking into consideration the issue of physical fitness. However, if exercising improves many risk factors associated with CVD including hypertension, patients who are physically active may consume more alcohol compared to non-active or sedentary subjects thus influencing the alcohol-hypertension relationship.

Age. Age is another important variable to consider for confounding when investigating the relationship between alcohol and hypertension. It is suggested that blood pressure and the risk of hypertension increases with age (Hart et al., 2012, Buford, 2016). Furthermore, a strong association between alcohol consumption and hypertension was reported in older people (van Leer et al., 1994). Based on these findings, older patients with type 2 diabetes may stop drinking alcohol as a way of controlling their blood pressure control thus influencing the alcohol-hypertension relationship.

Sex. Previous alcohol studies suggest females are more prone to elevated blood pressure compared to males (Weissfeld et al., 1988, Wakabayashi, 2008). However, the mechanism responsible for the sex-related difference in blood pressure is not fully understood. Hormonal factors and the use of contraceptive pills in females might explain the difference (Oparil and Miller, 2005). Consequently, to control the harm associated with alcohol, females may be encouraged to drink less compared to males. The latter may, therefore, change the alcohol-hypertension relationship.

Ethnicity. It is important to consider ethnicity when investigating the relationship between alcohol consumption and hypertension. Evidence in alcohol studies suggests higher mean blood pressure and higher rates of hypertension in black ethnic groups compared to other ethnic groups (Fuchs et al., 2001, Ikeda et al., 2013). To control the risk of hypertension, black people may be advised to reduce alcohol drinking and this could change the association between alcohol and hypertension.

The family history of hypertension. Evidence suggests greater mean blood pressures and a higher risk of hypertension amongst subjects with family history of hypertension compared to those without (Tozawa et al., 2001, Ranasinghe et al., 2015). For example in a study by Tozawa and colleagues where mean SBP levels for study participants were computed and assigned to groups according to the number of family members with history of hypertension, the results showed higher mean SBP levels in probands (individuals) with a family history of hypertension compared to those without. The proband with a mean SBP of 121 ± 17 mmHg was reported when 1 family member was hypertensive (n=1, 760). If 2 family members were hypertensive (n=280), the proband mean SBP was 124 ± 18 mmHg. If 3 or more family members were hypertensive (n=43), the proband mean SBP was 127 ± 17 mmHg. In contrast, the mean SBP in probands without a family history of hypertension was 119 ± 15 mmHg

(Tozawa et al., 2001). Based on these findings, a positive family history of hypertension may be linked to an increased risk of hypertension through family sharing of cultural/environmental and lifestyle factors such as alcohol consumption. Thus, a family history of hypertension may influence the relationship between alcohol consumption and hypertension in type 2 diabetes patients.

Antihypertension medication use. Antihypertensive(s) are important for lowering blood pressure levels and controlling the risk of hypertension (Bandi et al., 2017). Amongst alcoholics, the use of antihypertensive(s) may influence the association between alcohol consumption and hypertension. In a cross-sectional study of males aged ≥ 65 years, Wakabayashi (2010) found significantly higher odds ratios (ORs) of hypertension in heavy (≥ 22 and < 44 g/day) and very heavy (≥ 44 g/day) alcohol drinkers compared to non-drinkers in patients who did not receive therapy for hypertension. In contrast, the ORs of hypertension in heavy and very heavy alcohol drinkers receiving therapy for hypertension were insignificant. In addition, the ORs of hypertension for light drinkers (< 22 g/day) vs non-drinkers in the two groups (receiving therapy vs not receiving therapy for hypertension) were both insignificant. These findings suggest that alcohol-induced elevation of blood pressure may be suppressed by the use of antihypertensive therapy. Thus, antihypertensive therapy has potential to influence the alcohol-hypertension relationships.

Appendix F

F.1 Additional tables and figures in Chapter 6

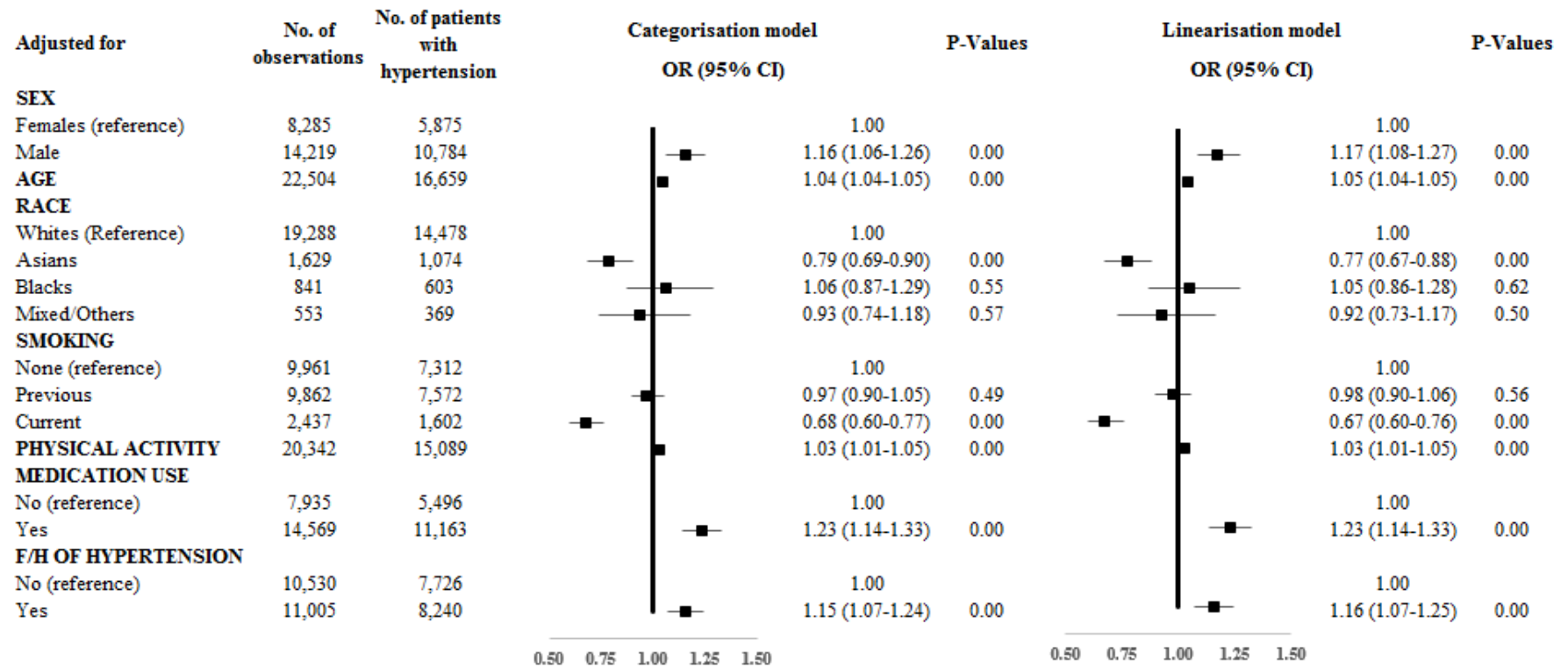


Figure 6.11: Odds ratios for several covariates adjusted for an alcohol-hypertension relationship and their 95% CIs obtained by categorising and linearising the alcohol consumption measures using logistic regression models.

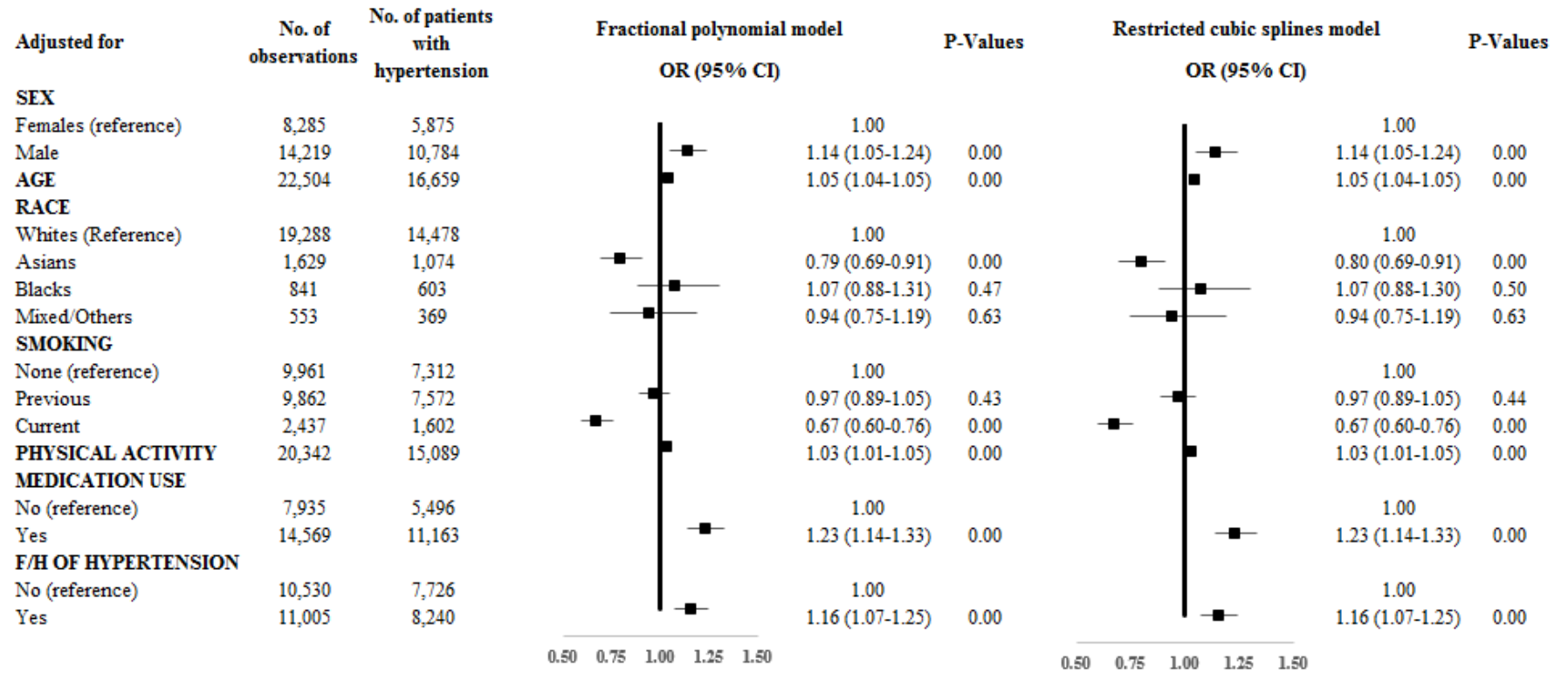


Figure 6.12: Odds ratios of several covariates adjusted for the alcohol-hypertension relationships and their 95% CIs obtained by fitting logistic regression using first-order degree fractional polynomial (FP1) and three knots restricted cubic spline (RCS3) models

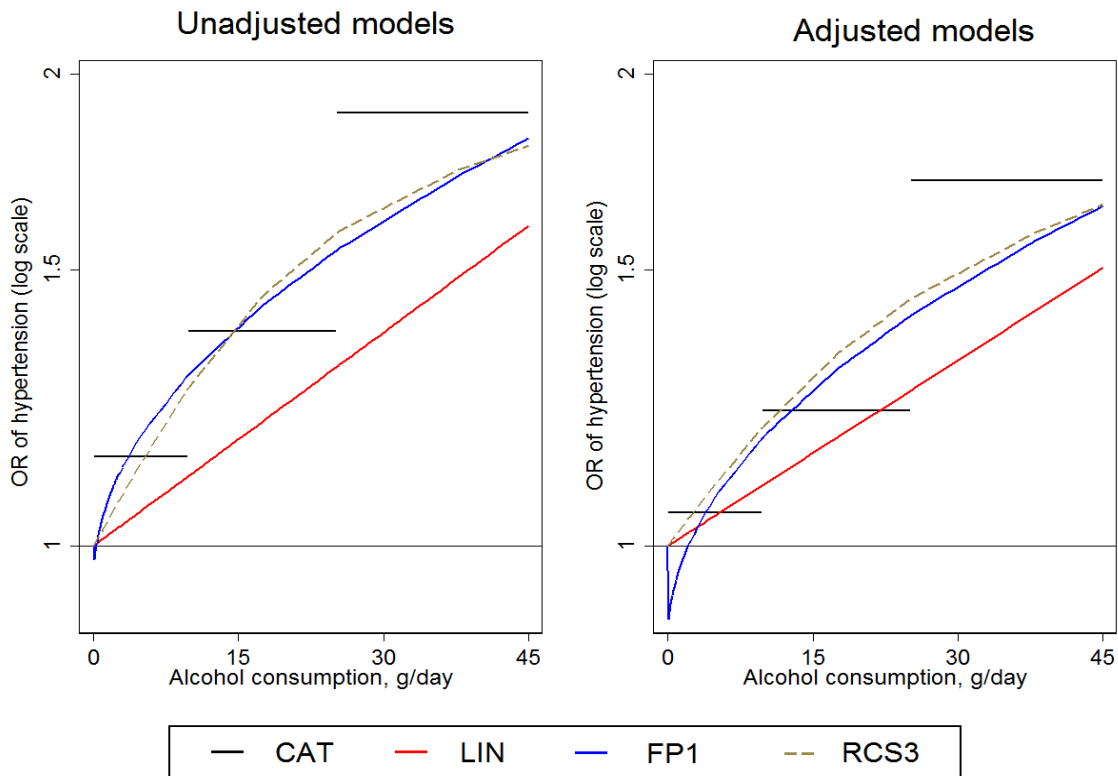


Figure 6.13: The unadjusted and adjusted odds of hypertension (on log scales) estimated using categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption (in g/day).

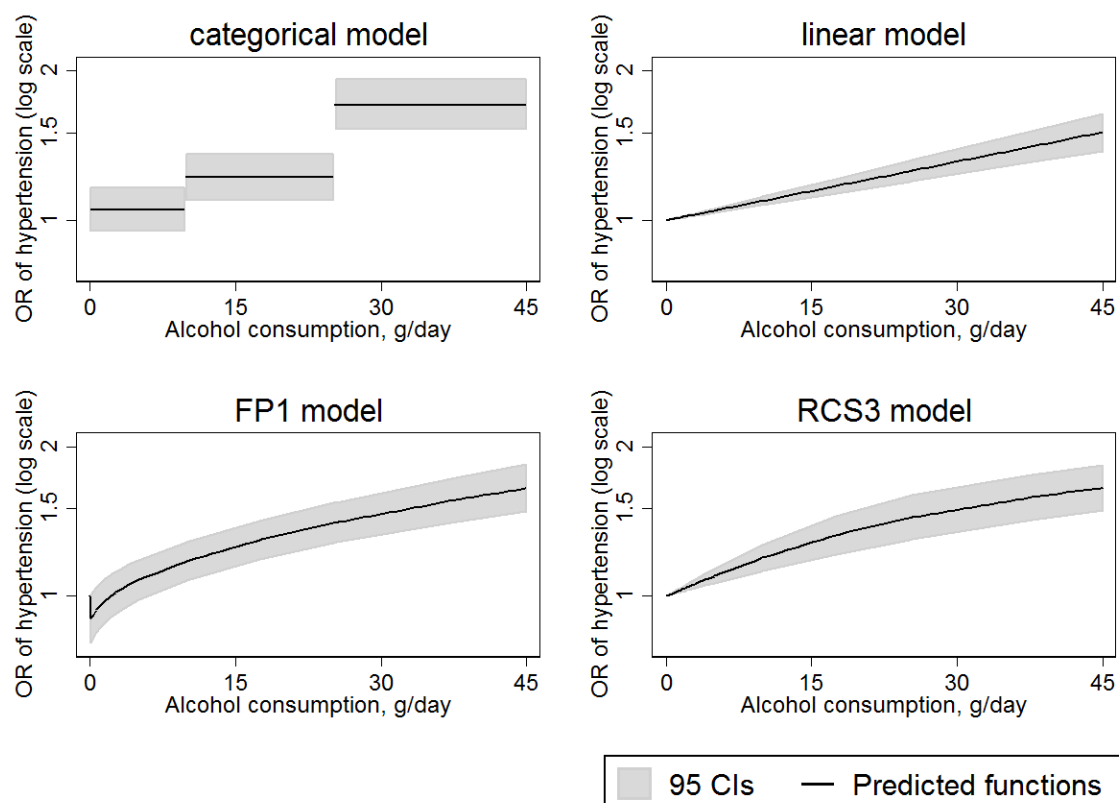


Figure 6.14: The adjusted odds of hypertension (on log scales) together with their 95% CIs estimated using the categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption, g/day

Table 6.6: The unadjusted and adjusted Odds ratios (ORs) of hypertension and their 95% confidence intervals obtained using the method of categorisation (CAT).

Alcohol consumption, g/day	No. of observations	No. of hypertensi on cases	Unadjusted OR (CAT Model)		Adjusted OR (CAT Model)	
			Estimate	95% CI	Estimate	95% CI
0	3,834	2,637	1.00	-	1.00	-
0-9.7	4,094	2,965	1.19	1.08 - 1.31	1.08	0.97 - 1.21
9.7-25.1	3,977	3,014	1.42	1.29 - 1.57	1.24	1.10 - 1.39
25.1+	3,995	3,255	2.00	1.80 - 2.22	1.79	1.58 - 2.03

Table 6.7: The unadjusted and adjusted odds ratios (ORs) of hypertension & their 95% confidence intervals obtained from the best fitting linearisation (LIN), fractional polynomials - first order degree (FP1) and the restricted cubic spline with 3 knots (RCS3) models. The odds of hypertension was modelled as a function of alcohol consumption, g/day.

Alcohol consumption, g/day	No. of observations	No. of hypertension cases	Ref. points	OR Estimates (LIN Based Model)		OR Estimates (FP1 Based Model)		OR Estimates (RCS3 Based Model)	
				Unadjusted	Adjusted	Unadjusted	Adjusted	Unadjusted	Adjusted
				OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)	OR (95% CI)
0	3,834	2,637	0 (ref)	1.00	1.00	1.00	1.00	1.00	1.00
0-9.7	4,094	2,965	5.0	1.05 (1.05-1.06)	1.05 (1.04-1.06)	1.23 (1.13-1.34)	1.10 (1.00-1.22)	1.14 (1.11-1.18)	1.10 (1.07-1.14)
9.7-25.1	3,977	3,014	17.5	1.20 (1.17-1.24)	1.18 (1.13-1.22)	1.49 (1.37-1.61)	1.33 (1.21-1.47)	1.49 (1.38-1.61)	1.35 (1.23-1.49)
25.1-49.9	2,541	2,054	37.5	1.49 (1.40-1.60)	1.42 (1.31-1.53)	1.79 (1.64-1.94)	1.61 (1.44-1.80)	1.81 (1.65-1.99)	1.63 (1.45-1.82)
49.9-74.9	930	760	62.5	1.95 (1.74-2.18)	1.78 (1.57-2.03)	2.13 (1.90-2.37)	1.91 (1.67-2.18)	1.99 (1.79-2.22)	1.87 (1.64-2.12)
74.9+	524	441	87.5	2.54 (2.17-2.97)	2.24 (1.87-2.69)	2.44 (2.14-2.78)	2.19 (1.87-2.57)	2.19 (1.88-2.55)	2.13 (1.79-2.54)

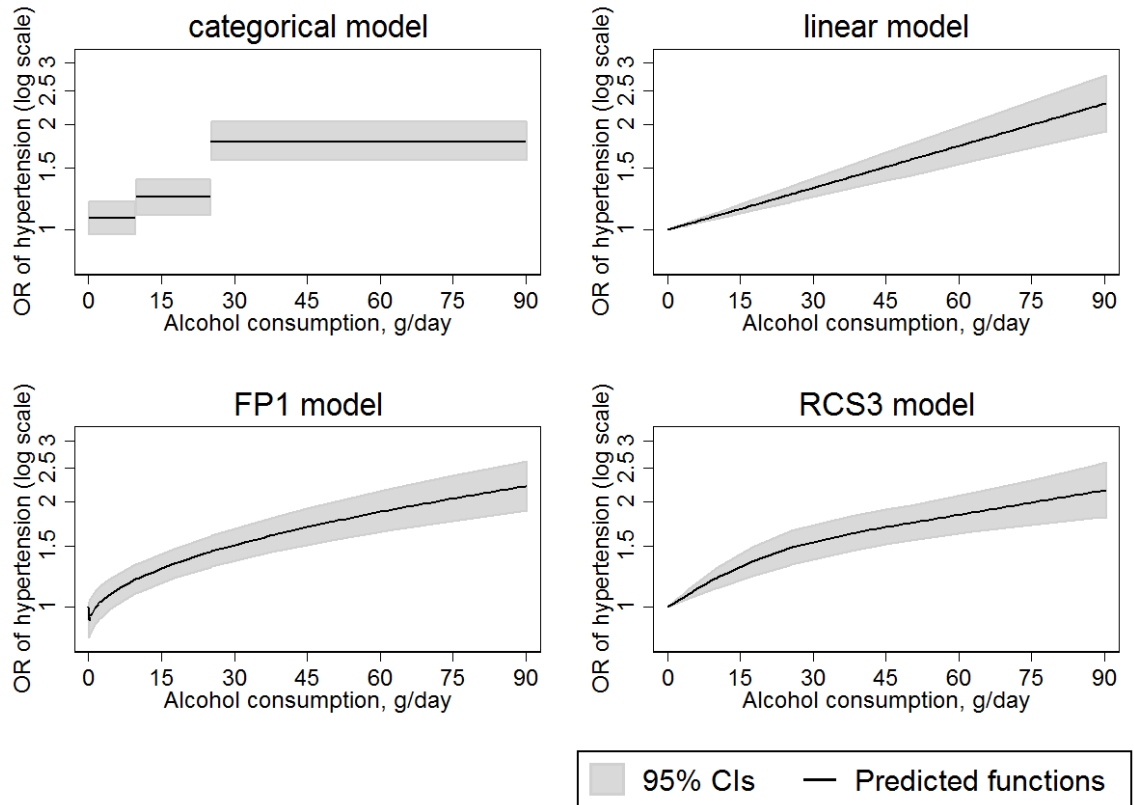



Figure 6.15: The adjusted odds of hypertension (on log scales) together with their 95% CIs estimated using the categorisation, linearisation, first order degree fractional polynomials (FP1), and restricted cubic splines with three knots (RCS3) models at different units of alcohol consumption, g/day.

Appendix G

G.1 Research protocol submitted at the UK Biobank

 Project Details		24/03/2017	
Name:	Onkabetse Mabikwa	Institute Name:	University of Leeds
Application:	26883	Current Stage:	Main - Adjudication in Progress
Application Name:	Modelling the alcohol-blood pressure associations in type 2 diabetes patients: UK Biobank		
Preliminary Application Details			
1a. The aims of the proposed research including the research question(s) that you are aiming to answer and the health condition(s) under investigation			
<p>The balance of effects of alcohol consumption on cardiovascular risk factors in people with type 2 diabetes is of considerable interest. Most studies reporting the effects of alcohol in type 2 diabetes has focused on insulin sensitivity, lipids/high density lipoprotein-cholesterol (HDL-C) and haemostatic factors. We aim to investigate the effects of alcohol intake on blood pressure (BP) in patients with type 2 diabetes. Various statistical approaches are applied to determine the direction, turning points and magnitudes of different doses of alcohol on BP. Interventions aimed at managing BP in alcohol consumers with type 2 diabetes are lacking.</p>			
1b. How does the proposed research meet UK Biobank's stated purpose?			
<p>Our research meets the UK Biobank's aim of improving and promoting healthy living within our societies. Knowledge gained in the research will support and promote the administration of patients living with type 2 diabetes hence improving care and quality of life. Further, by comparing different application methods, the research also have the potential to inform future studies on appropriate statistical approaches for analysing the exposure-outcome relationships</p>			
1c. Please give a non-technical description of how the research will be undertaken			
<p>Our study will examine the association of alcohol intake as an exposure and blood pressure outcomes in people with type 2 diabetes. Different statistical models will be fitted to establish a suitable exposure-outcome association with biologically meaningful interpretations. To account for potential risk factors, statistical models with multiple possible predictors will also be built for comparison. This approach enables us to examine the association models with predictor variables previously studied elsewhere and also examine other unknown variables supported by evidence found in the UK Biobank.</p>			
1d. Please state the approximate number of participants to be included (i.e. whether the full cohort or a subset)			
<p>We request the full cohort of type 2 diabetes patients (n=23,843) identified using the self reported baseline assessment data.</p>			
<p>A full reference to this cohort is described in an article by Eastwood et al (2016) below:</p>			
<p>Eastwood, S. V., Mathur, R., Atkinson, M., Brophy, S., Sudlow, C., Flaig, R., . . . Chaturvedi, N. (2016). Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. PLoS One, 11(9), e0162388. doi: 10.1371/journal.pone.0162388</p>			
2. Please describe the methodology and timetable for the proposed Research Project. Please include a description of the data and/or the quantity and type of samples required			
<p>The association between alcohol intake and blood pressure in people with</p>			

type 2 diabetes will be investigated using different statistical regression models based on fractional polynomials, restricted cubic splines, categorisation and linearisation approaches. The four methods vary in their application. As part of Ph.D work, the researcher proposes to demonstrate application of these methods using existing data collected by the UK Biobank.

The model fitted under the categorisation method compares systolic blood pressures (SBP) amongst the non-alcohol drinkers (0 g/day) and alcohol drinking participants classified in tertiles of alcohol consumption. For the linearisation method, the association between alcohol intake and SBP is assumed to be linear thus a simple linear regression model is fitted without transforming the alcohol intake measures. The two approaches are very popular in public health research.

To examine nonlinearity or an asymmetric relationship, we use fractional polynomials (FP) and restricted cubic splines (RCS) regression models. The two approaches allow flexibility and maintain the alcohol intake as a continuous variable in the analyses.

For multivariable regression models, we compare multiple linear regression, multivariable fractional polynomial regression (MFP) and multivariable regression spline model (MVRS) procedures for functional estimation. Thus the following existing UK Biobank data variables are considered and required for analysis.

Data required:

Biomarkers of a) diabetes: HbA1c, glucose; b) cardiovascular: blood pressure, lipids; c) renal: creatinine, potassium, sodium, urine albumin concentrations

Behavioral data: smoking, alcohol consumption, daily exercises

Dietary measures: calories, glucose fasting time, polyunsaturated fats, dietary fibre, total sugar and saturated fats

Socio-demographic data: age, sex, ethnicity, education, employment status

Physical measures: hand grip strength, weight, height, Body mass index, birth weight

Family history of diseases such as diabetes, cholesterol, hypertension, heart attack, angina, stroke, cancer

Medical history of diabetes, cholesterol, hypertension, heart attack, angina, stroke, cancer

Self reported regular medication for diabetes, cholesterol, hypertension, heart attack, angina, stroke, cancer

The proposed timetable for this project is as follows:

11/2016 - Registration with the UK Biobank
 12/2016 - Preliminary data application
 01/2017 - Main application of data
 02/2017 - Release of data by the UK Biobank
 03/2017 - Data cleaning and coding
 04/2017 - 05/2017 - Data Analysis and write-up
 06/2017 - Publications and presentation of results

The proposed timelines are based on the author's PhD schedules approved by the University of Leeds.

2.1. What results will you return to UK Biobank, i.e. what will you produce that will be useful to other researchers?

The model codes and publications shall be made available to the UK Biobank. We aim to publish the research in peer review journals such as diabetologia, or diabetes/metabolism research and reviews. Additionally, we will make presentations at national and international conferences or meetings organised by the Society for Social Medicine and Faculty of Medicine and Health at the University of Leeds.

Research is in the Public's Interest? **Y** Research is Health-Related? **Y**

Did the Principal Applicant foresee any Ethical issues with the Application? **N**

Did the Principal Applicant believe that Re-contact was potentially required? **N**

Did the Principal Applicant indicate that the project requires sample analysis? **N**

"Tags" provided by Principal Applicant

type 2 diabetes, alcohol, blood pressure.

Main Application Details

1a. The aims of the proposed research including the research question(s) that you are aiming to answer and the health condition(s) under investigation

The balance of effects of alcohol consumption on cardiovascular risk factors in people with type 2 diabetes is of considerable interest. Most studies reporting the effects of alcohol in type 2 diabetes has focused on insulin sensitivity, lipids/high density lipoprotein-cholesterol (HDL-C) and haemostatic factors. We aim to investigate the effects of alcohol intake on blood pressure (BP) in patients with type 2 diabetes. Various statistical approaches are applied to determine the direction, turning points and magnitudes of different doses of alcohol on BP. Interventions aimed at managing BP in alcohol consumers with type 2 diabetes are lacking.

1b. How does the proposed research meet UK Biobank's stated purpose?

Our research meets the UK Biobank's aim of improving and promoting healthy living within our societies. Knowledge gained in the research will support and promote the administration of patients living with type 2 diabetes hence improving care and quality of life. Further, by comparing different application methods, the research also have the potential to inform future studies on appropriate statistical approaches for analysing the exposure-outcome relationships

1c. Please give a non-technical description of how the research will be undertaken

Our study will examine the association of alcohol intake as an exposure and blood pressure outcomes in people with type 2 diabetes. Different statistical models will be fitted to establish a suitable exposure-outcome association with biologically meaningful interpretations. To account for potential risk factors, statistical models with multiple possible predictors will also be built for comparison. This approach enables us to examine the association models with predictor variables previously studied elsewhere and also examine other unknown variables supported by evidence found in the UK Biobank.

1d. Please state the approximate number of participants to be included (i.e. whether the full cohort or a subset)

We request the full cohort of type 2 diabetes patients (n=23,843) identified using the self reported baseline assessment data.

A full reference to this cohort is described in an article by Eastwood et al (2016) below:

Eastwood, S. V., Mathur, R., Atkinson, M., Brophy, S., Sudlow, C., Flaig, R., . . . Chaturvedi, N. (2016). Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One*, 11(9), e0162388. doi: 10.1371/journal.pone.0162388

Provided scientific rationale

Diabetes mellitus (MD) is amongst the serious non-communicable diseases targeted by world leaders for action due to steadily increasing numbers of cases reported over the past decades (WHO, 2016). The consequence of diabetes includes morbidity and mortality and is accelerated by its related complications such as kidney disease, cardiovascular disease, neuropathy, blindness and lower extremity amputations (Bebb et al., 2007; Deshpande, Harris-Hayes, & Schootman, 2008). Hypertension or high blood pressure is another cause of morbidity and mortality in patients with diabetes (Bebb et al., 2007). Hypertension is more prevalent in patients with type 2 diabetes. Therefore tighter and lower blood pressure levels are required and recommended to control and manage patients with type 2 diabetes. However, this is not easy. Lifestyle behaviours such as alcohol consumption increases the risk of type 2 diabetes. Heavy or excess alcohol consumption elevates the blood pressure and this could cause harm in patients with type 2 diabetes (Mori et al., 2016; Stamler et al., 2003).

Studies (Blomster et al., 2014; Gepner et al., 2015; Razay, Heaton, Bolton, & Hughes, 1992) recommend moderate alcohol consumption. The consumption of alcohol in moderate quantity has been associated with reduced incidence of risk amongst cardiovascular diseases and mortality. However, the risk-benefits of moderate alcohol consumption in patients with type 2 diabetes is controversial (Briasoulis, Agarwal, & Messerli, 2012; Gepner et al., 2015). The association between moderate alcohol consumption and blood pressure is not clear (Gepner et al., 2016). In some studies, linear, thresholds or J-shaped, and U-shaped alcohol-blood pressure relationships have been reported (Briasoulis et al., 2012). The contradictions complicate the potential benefits of alcohol consumption. Therefore, this area needs to be studied further. To do so, we propose using the UK Biobank study. The UK Biobank provides a large cohort of diabetes patients with proportions that compare to the general UK population. However, no analyses have yet being carried out in diabetes patients examining these relationships (Eastwood et al., 2016).

The study is important due to public and clinical interest in the area. Investigating the risk-benefits of alcohol intake in patients with type 2 diabetes is required to support clinicians in developing appropriate strategies to control the effects of the disease. Further, the potential risks and benefits of alcohol consumption can only be evaluated when the actual alcohol-blood pressure relationships are clear.

Detail of pilot studies undertaken

N/A

Please provide details of the methodology to be used

Study participants:

Patients with type 2 diabetes are defined based on an algorithm developed by Eastwood and colleagues (Eastwood et al., 2016) for UK Biobank users. In the algorithm, a total of 23,843 patients with 'probable' or 'possible' type 2 diabetes were identified from the self reported baseline data

obtained through the online touch screen questionnaires and the nurses interviews. Patients were assigned 'probable' status when there was greater certainty of type 2 diabetes, and 'possible' when there was less certainty.

The eligibility criteria (from the algorithm) for the 23,843 type 2 patients are provided below.

Inclusion criteria

- (i). Self - reported type 2 diabetes (nurse interviews (NI))
- (ii). Self-reported type 2 diabetes medications (online touch screen (TS) + (NI))
- (iii). Age of diagnosis for diabetes ≥ 36 years (amongst European origin) or ≥ 30 years (amongst South Asian or African-Caribbean origin)
- (iv). Currently on any other (non-metformin) oral diabetes medications (NI)

Exclusion criteria

- (i). Non-diabetes participants (i.e. all patients not reporting any diabetes from the NI, or gestational diabetes in both the NI and TS or any diabetes medication in both NI and TS)
- (ii). Self-reported type 1 diabetes patients (NI) which also includes self-reported insulin use < 12 months post diagnosis (TS and NI) or self-report current insulin use (TS and NI)
- (iii). self-reported gestational diabetes amongst the females who are pregnant (TS and NI)

The 23,843 patients with type 2 diabetes are all considered in the analysis investigating the alcohol-blood pressure relationships. However, for sensitivity analysis we will exclude 'possible' type 2 diabetes patients for model verification and to check conformity.

Blood pressure and Alcohol consumption measures

At the baseline, two systolic blood pressure measurements (in mmHg) were collected from each participant using the Omron digital blood pressure monitor and entered directly into a data system of the Assessment Centre Environment (ACE). In the analysis we estimate and use the average systolic blood pressure measures. The average BP measurement (in mmHg) is obtained using the two measures obtained at the baseline. If only one measurement of BP is available, we treat it as the average. Otherwise, participant with all two missing values are excluded in the analysis.

The baseline data on alcohol consumption was collected using the touch screens and web-based questionnaires. In the questionnaire, the alcohol intake was assessed with the following questions; "Did you have any alcoholic drinks yesterday? For instance, beer, wine or spirits". If the answer was yes, the participants were asked: "How many alcoholic beverages they drank?" The amount of intake or consumption was measured using a wine glass, pint, sherry/port glass, and other units of alcohol drinks. Based on participant's responses, the total alcohol intake (in grams) from both beverages in the last 24 hour were estimated and provided by the UK Biobank.

The other questions for participants who indicated that they drink alcohol is based on the "average weekly and monthly consumption of beer, wine, or spirits" and the alcohol intake frequency, "how often they drink alcohol in day/week/month?". The responses for the latter were recorded as daily or almost daily, three or four times a week, once or twice a week, one to three times a month, special occasions only, never and prefer not to answer.

Using the weekly/monthly consumption, we will estimate the average daily consumption (in grams) based on the standard conversion that one drink, i.e. half a pint (284 ml) contains 9g of alcohol, a glass of wine (144 ml), 10g a glass of port or sherry (57 ml), 11 g and a measure of spirits (23-28 ml), 8g (Jackson, Stewart, Beaglehole, & Scragg, 1985; Razay et al., 1992).

Further, the average weekly and monthly consumption (in grams) will also be considered and compared to the average daily intakes in the analysis. The average consumption will also be stratified according to the types of beverages for sub-group analysis. This is necessary to measure the effect of each beverage. We will exclude all the participants who did not answer or with missing values in the alcohol question, and those who drink occasionally. Alcohol measures which looks like outliers will also excluded in the analysis.

Statistical methods

The association between alcohol intake and blood pressure in people with type 2 diabetes will be investigated using different statistical regression models based on categorisation, linearisation, fractional polynomials, restricted cubic splines approaches. The four methods vary in their application.

The model fitted under the categorisation method compares systolic blood pressures (SBP) amongst the non-alcohol drinkers (0 g/day) and alcohol drinking participants classified in tertiles of alcohol consumption. For the linearisation method, the association between alcohol intake and SBP is assumed to be linear thus a simple linear regression model is fitted without transforming the alcohol intake measures. The two approaches are very popular in public health research (Royston & Sauerbrei, 2008).

To examine possibilities of nonlinear and asymmetric relationships, we use the fractional polynomials (FP) (Royston & Sauerbrei, 2008; Sauerbrei et al., 2006) and restricted cubic splines (RCS) (Royston & Sauerbrei, 2007) regression models. The two approaches are flexible in fitting complex nonlinear functions and maintains the alcohol intake as a quantitative variable in the analyses.

For multivariable regression models, we compare the multiple linear regression, multivariable fractional polynomial regression (MFP) and multivariable regression spline model (MVRS) procedures for functional estimation. These models account for other predictor variables in the data.

With application of the proposed methods, we anticipate different association results from each approach. This will help us establish best practices and also to explain some inconsistencies found when reporting the alcohol-blood pressure relationships in patients with type 2 diabetes. Finally, the findings will be shared with others through conference presentations and by publishing the research in peer reviewed journals.

Expected value of results? Results explaining the the effects of alcohol intake in patients with type 2 diabetes are necessary to guide the public and support clinicians in developing appropriate strategies to control and manage the disease.

Were references supporting justification for the project provided? Y

Will any Sample analysis be performed as part of the research project? N

If yes, which sample(s), how many, what volume and what assay?

Details of any power calculations made to support this application

Confirmed procedures in place to manage access to samples? N

Proposal for managing access to samples

Confirmed that all samples will be stored in a secure manner at all times? N

Proposal for securing samples storage

Confirmed that suitable storage facilities for samples are available N

Proposal for providing suitable sample storage

Please confirm that the data will be stored in a secure manner at all times (e.g., behind a firewall, use of anti-virus software) Y

If no, please describe how you will rectify this

Expected Project Start Date: 03/04/2017

Expected Project End Date: 31/03/2018

Estimated publication date? 31/08/2018

Has funding been granted? Y

Showcase Notes:

Since the algorithm is not yet implemented, we propose to run the outcomes ourselves based on the definitions provided by Eastwood et al (2016). Our results can later be validated with your incident cases after implementation of the algorithm by your team in the summer.

Additional data required (if available) includes:

Biomarkers of diabetes: HbA1c and glucose

Biomarkers of cardiovascular: lipids

Collaborators

Full Name	Institute	Email Address	Status
Dr. Darren Greenwood	University of Leeds	d.c.greenwood@leeds.ac.uk	Approved for App
Dr. Paul Baxter	University of Leeds	p.d.baxter@leeds.ac.uk	Approved for App
Dr. Sarah Fleming	University of Leeds	s.j.fleming@leeds.ac.uk	Approved for App

List of References

- Adkins, L. C. & Gade, M. N. 2012. Monte Carlo experiments using stata: a primer with examples. *30th Anniversary Edition*. Emerald Group Publishing Limited.
- Akobeng, A. K. 2007. Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatrica*, 96, 644-7.
- Allen, N., Sudlow, C., Downey, P., et al. 2012. UK Biobank: Current status and what it means for epidemiology. *Health Policy and Technology*, 1, 123-126.
- Alomari, M. A. & Al-Sheyab, N. A. 2016. Cigarette smoking lowers blood pressure in adolescents: the Irbid-TRY. *Inhalation Toxicology*, 28, 140-144.
- Altman, D. G., Lausen, B., Sauerbrei, W., et al. 1994. Dangers of using optimal cutpoints in the evaluation of prognostic factors. *Journal of the National Cancer Institute*, 86, 829-835.
- Altman, D. G., Vergouwe, Y., Royston, P., et al. 2009. Prognosis and prognostic research: validating a prognostic model. *BMJ: British Medical Journal*, 338, 1432-1435.
- Ambler, G. & Royston, P. 2001. Fractional polynomial model selection procedures: investigation of type I error rate. *Journal of Statistical Computation and Simulation*, 69, 89-108.
- Austin, P. C. & Brunner, L. J. 2004. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medicine*, 23, 1159-1178.
- Austin, P. C. & Steyerberg, E. W. 2014. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in Medicine*, 33, 517-535.
- Bakhshi, E., Eshraghian, M. R., Mohammad, K., et al. 2008. A comparison of two methods for estimating odds ratios: Results from the National Health Survey. *BMC Medical Research Methodology*, 8.
- Bakhshi, E., McArdle, B., Mohammad, K., et al. 2012. Let continuous outcome variables remain continuous. *Computational and Mathematical Methods in Medicine*.
- Bandi, P., Goldmann, E., Parikh, N. S., et al. 2017. Age-Related Differences in Antihypertensive Medication Adherence in Hispanics: A Cross-Sectional Community-Based Survey in New York City, 2011–2012. *Preventing Chronic Disease*, 14, E57.
- Baneshi, M. R. & Talei, A. R. 2011. Dichotomisation of continuous data: Review of methods, advantages and disadvantages. *Iranian Journal of Cancer Prevention*, 1, 26-32.
- Bantle, A. E., Thomas, W. & Bantle, J. P. 2008. Metabolic effects of alcohol in the form of wine in persons with type 2 diabetes mellitus. *Metabolism: Clinical & Experimental*, 57, 241-5.
- Barnett, A. H. 1994. Diabetes and hypertension. *British Medical Bulletin*, 50, 397-407.

- Bartlett, J. W., Seaman, S. R., White, I. R., et al. 2015. Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, 24, 462-487.
- Bash, L. D., Selvin, E., Steffes, M., et al. 2008. Poor Glycemic Control in Diabetes and The Risk of Incident Chronic Kidney Disease Even in The Absence of Albuminuria and Retinopathy: The Atherosclerosis Risk in Communities (Aric) Study. *Arch Intern Med*, 168, 2440-2447.
- Bazzano, L. A., Green, T., Harrison, T. N., et al. 2013. Dietary Approaches to Prevent Hypertension. *Current Hypertension Reports*, 15, 694-702.
- Bebb, C., Coupland, C., Stewart, J., et al. 2007. Practice and patient characteristics related to blood pressure in patients with type 2 diabetes in primary care: a cross-sectional study. *Family Practice*, 24, 547-54.
- Becher, H. 1992. The concept of residual confounding in regression models and some applications. *Statistics in Medicine*, 11, 1747-1758.
- Beck, N. & Jackman, S. 1998. Beyond linearity by default: Generalized additive models. *American Journal of Political Science*, 596-627.
- Beevers, D. G., Maheswaran, R. & Potter, J. F. 1990. Alcohol, blood pressure and antihypertensive drugs. *Journal of Clinical Pharmacy and Therapeutics*, 15, 395-397.
- Beilin, L. J., Puddey, I. B. & Burke, V. 1996. Alcohol and hypertension--kill or cure? *Journal of Human Hypertension*, 10 Suppl 2, S1-5.
- Bender, R. 2009. Introduction to the use of regression models in epidemiology. *Methods in Molecular Biology*, 471, 179-95.
- Benedetti, A., Abrahamowicz, M., Leffondre, K., et al. 2009. Using Generalized Additive Models to Detect and Estimate Threshold Associations. *International Journal of Biostatistics*, 5.
- Bennette, C. & Vickers, A. 2012. Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology*, 12.
- Bergmann, M. M., Jacobs, E. J., Hoffmann, K., et al. 2004. Agreement of self-reported medical history: comparison of an in-person interview with a self-administered questionnaire. *Eur J Epidemiol*, 19, 411-6.
- Beulens, J. W., Stolk, R. P., van der Schouw, Y. T., et al. 2005. Alcohol consumption and risk of type 2 diabetes among older women. *Diabetes Care*, 28, 2933-8.
- Binder, H., Sauerbrei, W. & Royston, P. 2013. Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response. *Statistics in Medicine*, 32, 2262-2277.
- Blanchet, F. G., Legendre, P. & Borcard, D. 2008. Forward selection of explanatory variables. *Ecology*, 89, 2623-2632.
- Blomster, J. I., Zoungas, S., Chalmers, J., et al. 2014. The relationship between alcohol consumption and vascular complications and mortality in individuals with type 2 diabetes. *Diabetes Care*, 37, 1353-1359.

- Breitling, L. P. 2015. Calcium intake and bone mineral density as an example of non-linearity and threshold analysis. *Osteoporosis International*, 26, 1271-1281.
- Breitling, L. P. & Brenner, H. 2010. Odd odds interactions introduced through dichotomisation of continuous outcomes. *Journal of Epidemiology and Community Health*, 64, 300-303.
- Brenner, H. 1998. A Potential Pitfall in Control of Covariates in Epidemiologic Studies. *Epidemiology*, 9, 68-71.
- Brenner, H. & Blettner, M. 1997. Controlling for continuous confounders in epidemiologic research. *Epidemiology*, 8, 429-434.
- Brenner, H. & Loomis, D. 1994. Varied forms of bias due to nondifferential error in measuring exposure. *Epidemiology*, 5, 510-517.
- Briasoulis, A., Agarwal, V. & Messerli, F. H. 2012. Alcohol Consumption and the Risk of Hypertension in Men and Women: A Systematic Review and Meta-Analysis. *The Journal of Clinical Hypertension*, 14, 792-798.
- British Cardiovascular Society, Association of British Clinical Diabetologists, British Association for Cardiovascular Prevention & Rehabilitation, et al. 2014. Joint British Societies' consensus recommendations for the prevention of cardiovascular disease (JBS3). *Heart*, 100, ii1-ii67.
- Buford, T. W. 2016. Hypertension and aging. *Ageing Research Reviews*, 26, 96-111.
- Buja, A., Scafato, E., Sergi, G., et al. 2009. Alcohol consumption and metabolic syndrome in the elderly: results from the Italian longitudinal study on aging. *European Journal Of Clinical Nutrition*, 64, 297.
- Burton, A., Altman, D. G., Royston, P., et al. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25, 4279-4292.
- Campbell, M. J. & Gardner, M. J. 1988. Statistics in Medicine: Calculating confidence intervals for some non-parametric analyses. *British Medical Journal (Clinical Research Ed.)*, 296, 1454-1456.
- Can, G., Schwandt, P., Onat, A., et al. 2009. Body fat, dyslipidemia, blood pressure and the effects of smoking in Germans and Turks. *Turkish Journal of Medical Sciences*, 39, 579-589.
- Chade, A. R., Lerman, A. & Lerman, L. O. 2005. Kidney in Early Atherosclerosis. *Hypertension*, 45, 1042-1049.
- Chai, T. & Draxler, R. R. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7, 1247-1250.
- Chang, H. J. & Park, C. Y. 1991. Effect of alcohol intake on blood pressure. *Journal of Catholic Medical College*, 44, 685-695.
- Chen, H., Cohen, P. & Chen, S. 2007. Biased odds ratios from dichotomization of age. *Statistics in Medicine*, 26, 3487-3497.
- Choudhury, S. R., Okayama, A., Kita, Y., et al. 1995. The associations between alcohol-drinking and dietary habits and blood-pressure in Japanese men. *Journal of Hypertension*, 13, 587-593.

- Cleveland, W. S. & Devlin, S. J. 1988. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.
- Cohen, J. 1983. The cost of dichotomization. *Applied Psychological Measurement*, 7, 249-253.
- Cohen, P. & Chen, H. 2009. How the Reflection of Linear Correlation in Odds Ratios Depends on the Cut-Off Points. *Communications in Statistics-Simulation and Computation*, 38, 610-620.
- Colberg, S. R., Sigal, R. J., Fernhall, B., et al. 2010. Exercise and type 2 diabetes: the American College of Sports Medicine and the American Diabetes Association: joint position statement. *Diabetes Care*, 33, e147-67.
- Colhoun, H. M., Dong, W., Barakat, M. T., et al. 1999. The scope for cardiovascular disease risk factor intervention among people with diabetes mellitus in England: a population-based analysis from the Health Surveys for England 1991-94. *Diabetic Medicine*, 16, 35-40.
- Collins, G. S., Ogundimu, E. O., Cook, J. A., et al. 2016. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. *Statistics in Medicine*, 35, 4124-4135.
- Colosia, A. D., Palencia, R. & Khan, S. 2013. Prevalence of hypertension and obesity in patients with type 2 diabetes mellitus in observational studies: a systematic literature review. *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, 6, 327-338.
- Crowther, M. J. & Lambert, P. C. 2013. Simulating biologically plausible complex survival data. *Statistics in Medicine*, 32, 4118-34.
- Culleton, B. F., Larson, M. G., Evans, J. C., et al. 1999. Prevalence and correlates of elevated serum creatinine levels: The framingham heart study. *Archives of Internal Medicine*, 159, 1785-1790.
- Cumsille, F., Bangdiwala, S. I., Sen, P. K., et al. 2000. Effect of dichotomizing a continuous variable on the model structure in multiple linear regression models. *Communications in Statistics-Theory and Methods*, 29, 643-654.
- Dalen, I., Buonaccorsi, J. P., Sexton, J. A., et al. 2009. Correction for misclassification of a categorized exposure in binary regression using replication data. *Statistics in Medicine*, 28, 3386-3410.
- de Gonzalez, A. B., Hartge, P., Cerhan, J. R., et al. 2010. Body-Mass Index and Mortality among 1.46 Million White Adults. *New England Journal of Medicine*, 363, 2211-2219.
- deRibeaux, M. B. 1997. Kidney structure and function. *Alcohol Health & Research World*, 21, 91-92.
- Deshpande, A. D., Harris-Hayes, M. & Schootman, M. 2008. Epidemiology of Diabetes and Diabetes-Related Complications. *Physical Therapy*, 88, 1254-1264.
- Desquilbet, L. & Mariotti, F. 2010. Dose-response analyses using restricted cubic spline functions in public health research. *Statistics in Medicine*, 29, 1037-1057.
- Diabetes UK. 2015. Available: <https://www.mrc.ac.uk/documents/pdf/diabetes-uk-facts-and-stats-june-2015/>.

- Diamond, G. A. 1992. What price perfection? Calibration and discrimination of clinical prediction models. *Journal of Clinical Epidemiology*, 45, 85-89.
- Diciccio, T. J. & Romano, J. P. 1988. A Review of Bootstrap Confidence Intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 50, 338-354.
- Duman, S. 2013. Rational approaches to the treatment of hypertension: diet. *Kidney International Supplements*, 3, 343-345.
- Durrleman, S. & Simon, R. 1989. Flexible regression models with cubic splines. *Statistics in Medicine*, 8, 551-561.
- Eastwood, S. V., Mathur, R., Atkinson, M., et al. 2016. Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. *PLoS One*, 11, e0162388.
- Emanuele, N. V., Swade, T. F. & Emanuele, M. A. 1998. Consequences of alcohol use in diabetics. *Alcohol Research and Health*, 22, 211.
- Erdem, Y., Arici, M., Altun, B., et al. 2010. The relationship between hypertension and salt intake in Turkish population: SALTURK study. *Blood Press*, 19, 313-8.
- Fall, T., Hägg, S., Mägi, R., et al. 2013. The role of adiposity in cardiometabolic traits: a Mendelian randomization analysis. *PLoS Med*, 10, e1001474.
- Farquhar, W. B., Edwards, D. G., Jurkowitz, C. T., et al. 2015. Dietary Sodium and Health More Than Just Blood Pressure. *Journal of the American College of Cardiology*, 65, 1042-1050.
- Fat, L. N. & Fuller, E. 2012. Drinking patterns. *In: The Health and Social Care Information Centre (ed.)*. England: National Statistics Office.
- Fedorov, V., Mannino, F. & Zhang, R. 2009. Consequences of dichotomization. *Pharmaceutical Statistics*, 8, 50-61.
- Figueiras, A. & Cadarso-Suárez, C. 2001. Application of Nonparametric Models for Calculating Odds Ratios and Their Confidence Intervals for Continuous Exposures. *American Journal of Epidemiology*, 154, 264-275.
- Friedman, J. H. & Silverman, B. W. 1989. Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31, 3-21.
- Froslic, K. F., Roislien, J., Laake, P., et al. 2010. Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) Study. *BMC Medical Research Methodology*, 10.
- Fry, A., Littlejohns, T. J., Sudlow, C., et al. 2017. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American journal of epidemiology*, 186, 1026-1034.
- Fuchs, F. D., Chambless, L. E., Whelton, P. K., et al. 2001. Alcohol consumption and the incidence of hypertension. *Hypertension*, 37, 1242-1250.
- Gardner, R. M., Lee, B. K., Magnusson, C., et al. 2015. Maternal body mass index during early pregnancy, gestational weight gain, and risk of autism spectrum disorders: Results from a Swedish total population and discordant sibling study. *International Journal of Epidemiology*, 44, 870-883.
- Gauffin, K., Vinnerljung, B. & Hjern, A. 2015. School performance and alcohol-related disorders in early adulthood: a Swedish national cohort study. *International Journal of Epidemiology*, 44, 919-927.

- Gepner, Y., Golan, R., Harman-Boehm, I., et al. 2015. Effects of Initiating Moderate Alcohol Intake on Cardiometabolic Risk in Adults With Type 2 Diabetes: A 2-Year Randomized, Controlled Trial.[Summary for patients in *Ann Intern Med*. 2015 Oct 20;163(8):I-34; PMID: 26457408]. *Annals of Internal Medicine*, 163, 569-79.
- Gepner, Y., Henkin, Y., Schwarzfuchs, D., et al. 2016. Differential Effect of Initiating Moderate Red Wine Consumption on 24-h Blood Pressure by Alcohol Dehydrogenase Genotypes: Randomized Trial in Type 2 Diabetes. *American journal of hypertension*, 29, 476-83.
- Giancristofaro, R. A. & Salmaso, L. 2007. Model performance analysis and model validation in logistic regression. *Statistica*, 63, 375-396.
- Gillman, M. W., Cook, N. R., Evans, D. A., et al. 1995. Relationship of alcohol intake with blood pressure in young adults. *Hypertension*, 25, 1106-1110.
- González-Ferrer, V., González-Ferrer, Y. & Ramírez-Marino, M. 2017. Statistical modeling in health research: Purpose drives approach. *MEDICC Review*, 19, 71-74.
- Govindarajulu, U. S., Malloy, E. J., Ganguli, B., et al. 2009. The Comparison of Alternative Smoothing Methods for Fitting Non-Linear Exposure-Response Relationships with Cox Models in a Simulation Study. *International Journal of Biostatistics*, 5.
- Greenland, S. 1995a. Avoiding Power Loss Associated with Categorization and Ordinal Scores in Dose-Response and Trend Analysis. *Epidemiology*, 6, 450-454.
- Greenland, S. 1995b. Dose-response and trend analysis in epidemiology: Alternatives to categorical analysis. *Epidemiology*, 6, 356-365.
- Greenland, S. & Neutra, R. 1980. Control of Confounding in the Assessment of Medical Technology. *International Journal of Epidemiology*, 9, 361-367.
- Greenland, S., Pearl, J. & Robins, J. M. 1999. Causal diagrams for epidemiologic research. *Epidemiology*, 37-48.
- Groenwold, R. H. H., Klungel, O. H., Altman, D. G., et al. 2013. Adjustment for continuous confounders: an example of how to prevent residual confounding. *CMAJ : Canadian Medical Association Journal*, 185, 401-406.
- Guertin, K. A., Freedman, N. D., Loftfield, E., et al. 2015. Coffee consumption and incidence of lung cancer in the NIH-AARP Diet and Health Study. *International Journal of Epidemiology*, 45, 929-939.
- Harrell, F. E. 2001. *Regression modeling strategies: with applications to linear models, logistic regression and survival analysis*, USA, Springer series in statistics.
- Harrell, F. E., Lee, K. L. & Mark, D. B. 1996. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.
- Hart, E. C., Joyner, M. J., Wallin, B. G., et al. 2012. Sex, ageing and resting blood pressure: gaining insights from the integrated balance of neural and haemodynamic factors. *Journal of Physiology-London*, 590, 2069-2079.
- Hastie, T. J. & Tibshirani, R. J. 1990. *Generalized additive models*, CRC Press.

- Haukoos, J. S. & Lewis, R. J. 2005. Advanced statistics: bootstrapping confidence intervals for statistics with "difficult" distributions. *Academic Emergency Medicine*, 12, 360-5.
- Heinzel, H. & Kaider, A. 1997. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine*, 54, 201-208.
- Higashiyama, A., Okamura, T., Watanabe, M., et al. 2013. Alcohol consumption and cardiovascular disease incidence in men with and without hypertension: the Suita study. *Hypertension Research*, 36, 58-64.
- Hillbom, M., Saloheimo, P. & Juvela, S. 2011. Alcohol consumption, blood pressure, and the risk of stroke. *Curr Hypertens Rep*, 13, 208-13.
- Hollander, N. & Schumacher, M. 2006. Estimating the functional form of a continuous covariate's effect on survival time. *Computational Statistics & Data Analysis*, 50, 1131-1151.
- House of Commons Science and Technology Committee 2012. Alcohol guidelines. The Stationary Office Limited, London.
- Humphries, M. 2013. Missing Data & How to Deal: An overview of missing data. *Population Research Center. University of Texas*. Recuperado de: <http://www.google.com/url>.
- Hunink, M. M., Weinstein, M. C., Wittenberg, E., et al. 2014. *Decision making in health and medicine: integrating evidence and values*, Cambridge University Press.
- Husain, K., Ansari, R. A. & Ferder, L. 2014. Alcohol-induced hypertension: Mechanism and prevention. *World Journal of Cardiology*, 6, 245-252.
- Ikeda, M. L. R., Barcellos, N. T., Alencastro, P. R., et al. 2013. Association of Blood Pressure and Hypertension with Alcohol Consumption in HIV-Infected White and Nonwhite Patients. *Scientific World Journal*.
- Jackson, R., Stewart, A., Beaglehole, R., et al. 1985. Alcohol consumption and blood pressure. *American journal of epidemiology*, 122, 1037-1044.
- Jee, S. H., Sull, J. W., Park, J., et al. 2006. Body-mass index and mortality in Korean men and women. *New England Journal of Medicine*, 355, 779-787.
- Jenkner, C., Lorenz, E., Becher, H., et al. 2016. Modeling continuous covariates with a "spike" at zero: Bivariate approaches. *Biometrical Journal*, 58, 783-796.
- Ji, L., Zhi, X., Lu, J., et al. 2014. Hyperglycemia and Blood Pressure Treatment Goal: A Cross Sectional Survey of 18350 Patients with Type 2 Diabetes in 77 Tertiary Hospitals in China. *PLOS ONE*, 9, e103507.
- Judd, S. E., McClure, L. A., Howard, V. J., et al. 2011. Heavy drinking is associated with poor blood pressure control in the REasons for Geographic and Racial Differences in Stroke (REGARDS) study. *International journal of environmental research and public health*, 8, 1601-12.
- Kang, H. 2013. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64, 402-406.
- Kannel, W. B., Wolf, P. A., Verter, J., et al. 1996. Epidemiologic assessment of the role of blood pressure in stroke: the Framingham Study. 1970. *Jama*, 276, 1269-78.

- Kaukonen, K., Bailey, M., Pilcher, D., et al. 2015. Systemic inflammatory response syndrome criteria in defining severe sepsis. *New England Journal of Medicine*, 372, 1629-1638.
- Kearney, P. M., Whelton, M., Reynolds, K., et al. 2004. Worldwide prevalence of hypertension: a systematic review. *Journal of Hypertension*, 22, 11-9.
- Keele, L. J. 2008. *Semiparametric Regression for the social sciences*, Chichester, England, John Wiley & Sons.
- Keil, U., Liese, A., Filipiak, B., et al. 1998. Alcohol, blood pressure and hypertension. *Novartis Foundation Symposium*, 216, 125-44; discussion 144-51.
- Keogh, R. H., Strawbridge, A. D. & White, I. R. 2012. Effects of classical exposure measurement error on the shape of exposure-disease associations. *Epidemiologic Methods*, 1, 13.
- Kerr, K. F., Brown, M. D., Zhu, K., et al. 2016. Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *Journal of Clinical Oncology*, 34, 2534-40.
- Kim, S.-J. & Kim, D.-J. 2012. Alcoholism and Diabetes Mellitus. *Diabetes & Metabolism Journal*, 36, 108-115.
- Klatsky, A. L., Friedman, G. D., Siegelau, A. B., et al. 1977. Alcohol consumption and blood pressure: Kaiser-Permanente multiphasic health examination data. *New England Journal of Medicine*, 296, 1194-1200.
- Klatsky, A. L. & Gunderson, E. 2008. Alcohol and hypertension: a review. *J Am Soc Hypertens*, 2, 307-17.
- Kodama, S., Tanaka, S., Heianza, Y., et al. 2013. Association Between Physical Activity and Risk of All-Cause Mortality and Cardiovascular Disease in Patients With Diabetes. *A meta-analysis*, 36, 471-479.
- Kodell, R. L. & Chen, J. J. 1991. Characterization of dose-response relationships inferred by statistically significant trend tests. *Biometrics*, 47, 139-146.
- Koehler, E., Brown, E. & Haneuse, S. 2009. On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *American Statistician*, 63, 155-162.
- Lago, R. M., Singh, P. P. & Nesto, R. W. 2007. Diabetes and hypertension. *Nat Clin Pract Endocrinol Metab*, 3, 667.
- Lamina, C., Sturm, G., Kollerits, B., et al. 2012. Visualizing interaction effects: a proposal for presentation and interpretation. *Journal of Clinical Epidemiology*, 65, 855-862.
- Langan, S., Schmitt, J., Coenraads, P. J., et al. 2010. The Reporting of Observational Research Studies in Dermatology Journals A Literature-Based Study. *Archives of Dermatology*, 146, 534-541.
- Law, G. R., Green, R. & Ellison, G. T. H. 2012. Confounding and Causal Path Diagrams. In: Tu, Y.-K. & Greenwood, D. C. (eds.) *Modern Methods for Epidemiology*. Dordrecht: Springer Netherlands.
- Lawlor, D. A., Emberson, J. R., Ebrahim, S., et al. 2003. Is the association between parity and coronary heart disease due to biological effects of pregnancy or adverse lifestyle risk factors associated with child-rearing? - Findings from the

- British women's heart and health study and the British regional heart study. *Circulation*, 107, 1260-1264.
- Lee, K.-S., Park, C.-Y., Meng, K.-H., et al. 1998. The Association of Cigarette Smoking and Alcohol Consumption with Other Cardiovascular Risk Factors in Men from Seoul, Korea. *Annals of Epidemiology*, 8, 31-38.
- Lemmens, P., Tan, E. S. & Knibbe, R. A. 1992. Measuring quantity and frequency of drinking in a general population survey: a comparison of five indices. *Journal of Studies on Alcohol*, 53, 476-86.
- Leslie, R. D. G. 1993. Metabolic changes in diabetes. *Eye*, 7, 205.
- Li, L., Hardy, R., Kuh, D., et al. 2015. Life-course body mass index trajectories and blood pressure in mid life in two British birth cohorts: stronger associations in the later-born generation. *International Journal of Epidemiology*, 44, 1018-1026.
- Li, Y., Wang, J.-G., Gao, P.-J., et al. 2006. Interaction Between Body Mass Index and Alcohol Intake in Relation to Blood Pressure in HAN and SHE Chinese*. *American journal of hypertension*, 19, 448-453.
- Lian, C. 1915. L'alcoolisme cause d'hypertension arterielle. *Bull Acad Med (Paris)*, 74, 525-28.
- Lip, G. Y. & Beevers, D. G. 2003. Alcohol and hypertension-does it matter?(no!). *Journal of Cardiovascular Risk*, 10, 11-14.
- Little, J., Higgins, J. P., Ioannidis, J. P., et al. 2009. Strengthening the reporting of genetic association studies (STREGA)-An extension of the STROBE statement. *Genetic Epidemiology*, 33, 581-98.
- Liu, Z., Fang, F., Chang, E., et al. 2015. Sibship size, birth order and risk of nasopharyngeal carcinoma and infectious mononucleosis: A nationwide study in Sweden. *International Journal of Epidemiology*.
- Long, J. & Ryoo, J. 2010. Using fractional polynomials to model non-linear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 63, 177-203.
- Lorenz, E., Jenkner, C., Sauerbrei, W., et al. 2017. Modeling Variables With a Spike at Zero: Examples and Practical Recommendations. *American journal of epidemiology*, 185, 650-660.
- Luo, W., Guo, Z., Hao, C., et al. 2013. Interaction of current alcohol consumption and abdominal obesity on hypertension risk. *Physiology & Behavior*, 122, 182-186.
- Luo, Z. & Wahba, G. 1997. Hybrid Adaptive Splines. *Journal of the American Statistical Association*, 92, 107-116.
- MacCallum, R. C., Zhang, S. B., Preacher, K. J., et al. 2002. On the practice of dichotomization of quantitative variables. *Psychological Methods*, 7, 19-40.
- Maclure, M. & Greenland, S. 1992. Tests for trend and dose-response - Misinterpretations and alternatives. *American Journal of Epidemiology*, 135, 96-104.
- Maheswaran, R., Beevers, D. G., Kendall, M. J., et al. 1990. The interaction of alcohol and b-blockers in arterial hypertension. *Journal of Clinical Pharmacy and Therapeutics*, 15, 405-410.

- Makridakis, S. G. & Hibon, M. 1995. *Evaluating Accuracy (or Error) Measures*, Fontainebleau, France, INSEAD.
- Marill, K. A. 2004. Advanced Statistics: Linear Regression, Part II: Multiple Linear Regression. *Academic Emergency Medicine*, 11, 94-102.
- Marmot, M. G., Elliott, P., Shipley, M. J., et al. 1994. Alcohol and blood pressure - the intersalt study. *British Medical Journal*, 308, 1263-1267.
- Matsumoto, C., Tomiyama, H., Yamada, J., et al. 2009. Association of blood pressure levels with the effects of alcohol intake on the vasculature in Japanese men. *Hypertension Research*, 32, 127-132.
- May, S. & Bigelow, C. 2005. Modeling nonlinear dose-response relationships in epidemiologic studies: Statistical approaches and practical challenges. *Dose-Response*, 3, 474-490.
- McConnell, B. & Vera-Hernández, M., 2015, *Going beyond simple sample size calculations: a practitioner's guide*: IFS Working Papers.
- McCullagh, P. & Nelder, J. A. 1989. *Generalized linear models*, CRC press.
- Mckee, M. & Britton, A. 1998. The positive relationship between alcohol and heart disease in eastern Europe: potential physiological mechanisms. *Journal of the Royal Society of Medicine*, 91, 402-407.
- Min, Y. & Agresti, A. 2002. Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, 1, 7-33.
- Moons, K. G., Altman, D. G., Reitsma, J. B., et al. 2015. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and ElaborationThe TRIPOD Statement: Explanation and Elaboration. *Annals of Internal Medicine*, 162, W1-W73.
- Moore, T. J., Conlin, P. R., Ard, J., et al. 2001. DASH (Dietary Approaches to Stop Hypertension) diet is effective treatment for stage 1 isolated systolic hypertension. *Hypertension*, 38, 155-8.
- Moreira, L. B., Fuchs, F. D., Moraes, R. S., et al. 1998. Alcohol intake and blood pressure: the importance of time elapsed since last drink. *Journal of hypertension*, 16, 175-180.
- Mori, T. A., Burke, V., Zilkens, R. R., et al. 2016. The effects of alcohol on ambulatory blood pressure and other cardiovascular risk factors in type 2 diabetes: a randomized intervention. *Journal of hypertension*, 34, 421-8; discussion 428.
- Morton, W. T. 1988. Knot positions in least-squares fitting of data using cubic splines. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 272, 861-865.
- Muggeo, V. M. R. 2003. Estimating regression models with unknown break-points. *Statistics in Medicine*, 22, 3055-3071.
- Müller, M. 2012. Generalized linear models. *Handbook of Computational Statistics*. Germany: Springer-Verlag.
- National Statistics 2016. Health Survey for England. In: Statistics, N. (ed.). England: NHS Digital.
- National Statistics, 2017, *Overview of the UK population* England: Office of National Statistics. Available:

<https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/july2017>.

- Nelder, J. A. & Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society*, 135, 370-384.
- NICE 2014. Obesity: identifying, assessing and managing obesity in adults, young people and children. *In: Excellence*, N. I. f. H. a. C. (ed.). United Kingdom: National Institute for Health and Care Excellence
- NICE. 2015. Type 2 diabetes in adults: management (NG28). Available: <http://nice.org.uk/guidance/ng28>.
- Nieboer, D., Vergouwe, Y., Roobol, M. J., et al. 2015. Non-linear modeling was applied thoughtfully for risk prediction: the Prostate Biopsy Collaborative group. *Journal of Clinical Epidemiology*, 68, 426-434.
- Núñez, E., Steyerberg, E. W. & Núñez, J. 2011. Regression Modeling Strategies. *Revista Española de Cardiología (English Edition)*, 64, 501-507.
- O'Brien, S. M. 2004. Cutpoint selection for categorizing a continuous predictor. *Biometrics*, 60, 504-509.
- Obermeyer, Z. & Emanuel, E. J. 2016. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine*, 375, 1216-1219.
- Okubo, Y., Sairenchi, T., Irie, F., et al. 2014. Association of Alcohol Consumption With Incident Hypertension Among Middle-Aged and Older Japanese Population The Ibarakai Prefectural Health Study (IPHS). *Hypertension*, 63, 41-47.
- Oparil, S. & Miller, A. P. 2005. Gender and blood pressure. *Journal of clinical hypertension (Greenwich, Conn.)*, 7, 300-9.
- Paoletti, R., Sies, H., Bug, J., et al. 1998. *Vitamin C: The state of art in disease prevention sixty years after the Nobel Prize*, Italy, Springer.
- Parekh, R. S. & Klag, M. J. 2001. Alcohol: role in the development of hypertension and end-stage renal disease. *Current Opinion in Nephrology and Hypertension*, 10, 385-390.
- Pastor, R. & Guallar, E. 1998. Use of two-segmented logistic regression to estimate change-points in epidemiologic studies. *American Journal of Epidemiology*, 148, 631-42.
- Paul, P., Pennell, M. L. & Lemeshow, S. 2013. Standardizing the power of the Hosmer–Lemeshow goodness of fit test in large data sets. *Statistics in Medicine*, 32, 67-80.
- Peacock, J. L., Sauzet, O., Ewings, S. M., et al. 2012. Dichotomising continuous data while retaining statistical power using a distributional approach. *Statistics in Medicine*, 31, 3089-3103.
- Pearson, T. A. 1996. Alcohol and heart disease. *Circulation*, 94, 3023-3025.
- Philippe, P. & Mansi, O. 1998. Nonlinearity in the epidemiology of complex health and disease processes. *Theoretical Medicine and Bioethics*, 19, 591-607.

- Pinheiro, J. C. & Bates, D. M. 2000. *Mixed-Effects Models in S and S-Plus*, New York, USA, Springer-Verlag New York, Inc.
- Pocock, S. J., Collier, T. J., Dandreo, K. J., et al. 2004. Issues in the reporting of epidemiological studies: A survey of recent practice. *British Medical Journal*, 329, 883-887.
- Ranasinghe, P., Cooray, D. N., Jayawardena, R., et al. 2015. The influence of family history of Hypertension on disease prevalence and associated metabolic risk factors among Sri Lankan adults. *BMC Public Health*, 15, 576.
- Razay, G., Heaton, K. W., Bolton, C. H., et al. 1992. Alcohol consumption and its relation to cardiovascular risk factors in British women. *British Medical Journal*, 304, 80-83.
- Reddy, K. S. & Katan, M. B. 2004. Diet, nutrition and the prevention of hypertension and cardiovascular diseases. *Public Health Nutrition*, 7, 167-186.
- Richardson, D. B. & Loomis, D. 2004. The impact of exposure categorisation for grouped analyses of cohort data. *Occupational and Environmental Medicine*, 61, 930-935.
- Rosenberg, P. S., Katki, H., Swanson, C. A., et al. 2003. Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Statistics in Medicine*, 22, 3369-3381.
- Royston, P. 2013. marginscontplot: Plotting the marginal effects of continuous predictors. *Stata Journal*, 13, 510-527.
- Royston, P. & Altman, D. G. 1994. Regression using fractional polynomials of continuous covariates - Parsimonious parametric modeling. *Applied Statistics-Journal of the Royal Statistical Society Series C*, 43, 429-467.
- Royston, P. & Altman, D. G. 2010. Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine*, 29, 2508-2520.
- Royston, P., Altman, D. G. & Sauerbrei, W. 2006. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, 25, 127-141.
- Royston, P., Ambler, G. & Sauerbrei, W. 1999. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology*, 28, 964-974.
- Royston, P. & Sauerbrei, W. 2005. Building multivariable regression models with continuous covariates in clinical epidemiology - With an emphasis on fractional polynomials. *Methods of Information in Medicine*, 44, 561-571.
- Royston, P. & Sauerbrei, W. 2007. Multivariable modeling with cubic regression splines: A principled approach. *Stata Journal*, 7, 45-70.
- Royston, P. & Sauerbrei, W. 2008. *Multivariable Model-building: A pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables*, Wiley Series in Probability and Statistics.
- Rubin, D. B. 2004. *Multiple imputation for nonresponse in surveys*, USA, John Wiley & Sons.
- Rutherford, M. J., Crowther, M. J. & Lambert, P. C. 2015. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event

- data: a simulation study. *Journal of Statistical Computation and Simulation*, 85, 777-793.
- Sacks, F. M., Svetkey, L. P., Vollmer, W. M., et al. 2001. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. *New England journal of medicine*, 344, 3-10.
- Saremi, A., Hanson, R. L., Tulloch-Reid, M., et al. 2004. Alcohol consumption predicts hypertension but not diabetes. *Journal of Studies on Alcohol*, 65, 184-90.
- Sauer, B. & VanderWeele, T. J. 2013. Use of directed acyclic graphs. Rockville (MD): Agency for Healthcare Research and Quality (US).
- Sauerbrei, W., Abrahamowicz, M., Altman, D. G., et al. 2014. Strengthening analytical thinking for observational studies: The STRATOS initiative. *Statistics in Medicine*, 33, 5413-5432.
- Sauerbrei, W., Meier-Hirmer, C., Benner, A., et al. 2006. Multivariable regression model building by using fractional polynomials: Description of SAS, STATA and R programs. *Computational Statistics & Data Analysis*, 50, 3464-3485.
- Sauerbrei, W. & Royston, P. 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society Series a-Statistics in Society*, 162, 71-94.
- Sauerbrei, W., Royston, P. & Binder, H. 2007. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. *Statistics in Medicine*, 26, 5512-5528.
- Saunders, J. B., Beevers, D. G. & Paton, A. 1981. Alcohol-induced hypertension. *The Lancet*, 318, 653-656.
- Sauzet, O. & Peacock, J. L. 2014. Estimating dichotomised outcomes in two groups with unequal variances: a distributional approach. *Statistics in Medicine*, 33, 4547-4559.
- Schimek, M. G. & Turlach, B. A. 2000. Additive and Generalized Additive Models. *Smoothing and Regression*. John Wiley & Sons, Inc.
- Schmidt, A. F. & Finan, C. 2017. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*.
- Schmidt, C. O., Ittermann, T., Schulz, A., et al. 2013a. Linear, nonlinear or categorical: How to treat complex associations in regression analyses? Polynomial transformations and fractional polynomials. *Int J Public Health*, 58, 157-160.
- Schmidt, C. O., Ittermann, T., Schulz, A., et al. 2013b. Linear, nonlinear or categorical: How to treat complex associations? Splines and nonparametric approaches. *Int J Public Health*, 58, 161-165.
- Schneider, A., Hommel, G. & Blettner, M. 2010. Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications. *Deutsches Ärzteblatt International*, 107, 776-782.
- Shmueli, G. 2010. To explain or to predict? *Statistical science*, 25, 289-310.
- Siegmund, D. O. 2005. *Probability theory* [Online]. Encyclopædia Britannica, inc. . Available: <https://www.britannica.com/topic/probability-theory/An-alternative-interpretation-of-probability> [Accessed March 08, 2017].

- Smith, P. L. 1979. Splines As a Useful and Convenient Statistical Tool. *The American Statistician*, 33, 57-62.
- Stamler, J., Elliott, P., Dennis, B., et al. 2003. INTERMAP: background, aims, design, methods, and descriptive statistics (nondietary). *Journal of Human Hypertension*, 17, 591-608.
- StataCorp LP 2013. Stata: Release 13 - Statistical software. 13 ed. College Station, Texas: Stata Press.
- Steenland, K. & Deddens, J. A. 2004. A practical guide to dose-response analyses and risk assessment in occupational epidemiology. *Epidemiology*, 15, 63-70.
- Steffens, A. A., Moreira, L. B., Fuchs, S. C., et al. 2006. Incidence of hypertension by alcohol consumption: is it modified by race? *J Hypertens*, 24, 1489-92.
- Stessman, J. & Jacobs, J. M. 2014. Diabetes Mellitus, Physical Activity, and Longevity Between the Ages of 70 and 90. *Journal of the American Geriatrics Society*, 62, 1329-1334.
- Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., et al. 2013. Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research. *PLoS Medicine*, 10, e1001381.
- Steyerberg, E. W., Vickers, A. J., Cook, N. R., et al. 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)*, 21, 128-138.
- Stockwell, T., Donath, S., Cooper-Stanbury, M., et al. 2004. Under-reporting of alcohol consumption in household surveys: a comparison of quantity-frequency, graduated-frequency and recent recall. *Addiction*, 99, 1024-33.
- Stone, C. J. 1986. [Generalized Additive Models]: Comment. *Statist. Sci.*, 1, 312-314.
- Strasak, A. M., Umlauf, N., Pfeiffer, R. M., et al. 2011. Comparing penalized splines and fractional polynomials for flexible modelling of the effects of continuous predictor variables. *Computational Statistics & Data Analysis*, 55, 1540-1551.
- Streiner, D. L. 2002. Breaking up is hard to do: The heartbreak of dichotomizing continuous data. *Canadian Journal of Psychiatry-Revue Canadienne De Psychiatrie*, 47, 262-266.
- Strelan, J. C., Kerkhoffs, E. & Snorek, M. 2001. Median confidence intervals. *Modelling and Simulation 2001-Proceedings of the ESM 2001*.
- Stump, C. S. 2011. Physical Activity in the Prevention of Chronic Kidney Disease. *Cardiorenal Medicine*, 1, 164-173.
- Suissa, S. 1991. Binary methods for continuous outcomes - a parametric alternative. *Journal of Clinical Epidemiology*, 44, 241-248.
- Taylor, J. M. G. & Yu, M. G. 2002. Bias and efficiency loss due to categorizing an explanatory variable. *Journal of Multivariate Analysis*, 83, 248-263.
- Textor, J. 2013. *Drawing and analyzing causal DAGs with DAGitty: User manual for version 2.0* [Online]. Available: <http://www.dagitty.net/manual-2.x.pdf>.
- Therneau, T. M. & Grambsch, P. M. 2000. *Modeling survival data: extending the Cox model*, Springer Science & Business Media.

- Tozawa, M., Oshiro, S., Iseki, O., et al. 2001. Family history of hypertension and blood pressure in a screened cohort. *Hypertension Research*, 24, 93-98.
- Turner, E. L., Dobson, J. E. & Pocock, S. J. 2010. Categorisation of continuous risk factors in epidemiological publications: A survey of current practice. *Epidemiologic perspectives & innovations : EP+I*, 7, 9-9.
- Turner, H. 2008. Introduction to generalized linear models. *Rapport technique, Vienna University of Economics and Business*.
- Uiterwaal, C., Verschuren, W. M. M., Bueno-De-Mesquita, H. B., et al. 2007. Coffee intake and incidence of hypertension. *American Journal of Clinical Nutrition*, 85, 718-723.
- UK Biobank, 2011a, *Access procedures: Application and review procedures for access to the UK Biobank Resource* England: UK Biobank.
- UK Biobank 2011b. Blood pressure. England: UK Biobank.
- Van Bommel, E. & Cleophas, T. 2012. Potassium treatment for hypertension in patients with high salt intake: a meta-analysis.
- van Leer, E. M., Seidell, J. C. & Kromhout, D. 1994. Differences in the Association between Alcohol Consumption and Blood Pressure by Age, Gender, and Smoking. *Epidemiology*, 5, 576-582.
- Van Thiel, D. H., Gavaler, J. S., Little, J. M., et al. 1977. Alcohol: its effect on the kidney. *Metabolism*, 26, 857-66.
- Vanleer, E. M., Seidell, J. C. & Kromhout, D. 1994. Differences in the association between alcohol-consumption and blood-pressure by age, gender, and smoking. *Epidemiology*, 5, 576-582.
- Vesey, C., Saloojee, Y., Cole, P., et al. 1982. Blood carboxyhaemoglobin, plasma thiocyanate, and cigarette consumption: implications for epidemiological studies in smokers. *British Medical Journal (Clinical Research Ed.)*, 284, 1516-1518.
- Vickers, A. J. & Elkin, E. B. 2006. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26, 565-74.
- Victora, C. G., Horta, B. L., de Mola, C. L., et al. 2015. Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. *The Lancet Global Health*, 3, e199-e205.
- Von Elm, E., Altman, D. G., Egger, M., et al. 2007. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Epidemiology*, 18, 800-4.
- Wakabayashi, I. 2008. Influence of gender on the association of alcohol drinking with blood pressure. *American Journal of Hypertension*, 21, 1310-1317.
- Wakabayashi, I. 2010. History of Antihypertensive Therapy Influences the Relationships of Alcohol With Blood Pressure and Pulse Pressure in Older Men. *American journal of hypertension*, 23, 633-638.
- Wakeford, R. & McElvenny, D. 2007. From epidemiological association to causation. *Occupational Medicine*, 57, 464-465.

- Wang, L., Cui, L., Wang, Y., et al. 2015. Resting heart rate and the risk of developing impaired fasting glucose and diabetes: The Kailuan prospective study. *International Journal of Epidemiology*, 44, 689-699.
- Web of Science. 2015. Journal Citation Reports. Thomson Reuters.
- Weissfeld, J. L., Johnson, E. H., Brock, B. M., et al. 1988. Sex and age interactions in the association between alcohol and blood-pressure. *American journal of epidemiology*, 128, 559-569.
- White, I. 2015. How to choose an analysis to handle missing data in longitudinal observational studies. Cambridge, UK: Medical Research Council.
- White, I. R. 2010. simsum: Analyses of simulation studies including Monte Carlo error. *Stata Journal*, 10, 369-385.
- Williams, J. S. 2011. Assessing the suitability of fractional polynomial methods in health services research: a perspective on the categorization epidemic. *Journal of Health Services Research and Policy*, 16, 147-152.
- Williams, R. 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. *Stata Journal*, 12, 308-331.
- Willmott, C. J. & Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79-82.
- World Health Organization, 2006, *Guidelines for the management of hypertension in patients with diabetes* Cairo. Available: <http://www.who.int/iris/handle/10665/119810>.
- World Health Organization 2016. Global report on diabetes. Geneva, Switzerland: World Health Organisation.
- Zhang, Y. T., Laraia, B. A., Mujahid, M. S., et al. 2015. Does Food Vendor Density Mediate the Association Between Neighborhood Deprivation and BMI?: A G-computation Mediation Analysis. *Epidemiology*, 26, 344-352.
- Zhao, L. P. & Kolonel, L. N. 1992. Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies. *American Journal of Epidemiology*, 136, 464-74.
- Zhou, S. & Shen, X. 2001. Spatially Adaptive Regression Splines and Accurate Knot Selection Schemes. *Journal of the American Statistical Association*, 96, 247-259.