



UNIVERSITY OF CAMBRIDGE

STATISTICAL MODELS FOR ESTIMATING THE INTAKE OF  
NUTRIENTS AND FOODS FROM COMPLEX SURVEY DATA

---

**This dissertation is submitted to the University of Cambridge for the degree of  
Doctor of Philosophy**

---

*Author:*

David Andrew PELL  
JESUS COLLEGE

*Supervisor:*

Dr. Ivonne SOLIS-TRAPALA

December 2018



## **Statistical models for estimating the intake of nutrients and foods from complex survey data** David Andrew Pell

**Background:** The consequences of poor nutrition are well known and of wide concern. Governments and public health agencies utilise food and diet surveillance to make decisions that lead to improvements in nutrition. Three important objectives of diet surveillance include to 1) assess the nutritional status of particular population groups; 2) identify population groups at high risk of deterioration in nutrient consumption, and 3) evaluate the allocation of resources to alleviate nutritional deficiencies. Iron deficiency is considered to be one of the most prevalent forms of malnutrition affecting both men and women, and people of all ages and socio-economic status; however, the magnitude of the problem has not been robustly quantified. National diet and nutrition surveys are important sources of information that is representative of the target population. These surveys often utilise complex sample designs for efficient data collection. Common designs include the use of sampling weightings, multistage sampling and stratification. Dietary data can be collected through food diaries that participants fill in for a period of 4-7 consecutive days. There are several challenges in the statistical analysis of dietary intake data collected using complex survey designs, which have not been fully addressed by current methods. Firstly, the shape of the distribution of intake can be highly skewed due to the presence of outlier observations and a large proportion of zero observations arising from the inability of the food diary to capture consumption within the period of observation. Secondly, dietary data is subject to variability arising from day-to-day individual variation in food consumption and measurement error, and this needs to be accounted for in the estimation procedure for correct inferences. Thirdly, the complex sample design needs to be incorporated into the estimation procedure to allow extrapolation of results into the target population. This thesis aims to develop novel statistical methods to address these challenges, motivated by the three objectives of diet surveillance described above and applied to the analysis of iron intake data from the UK National Diet and Nutrition Survey Rolling Programme (NDNS RP) and UK national prescription data of iron deficiency medication.

**Methods:** 1) To assess the nutritional status of particular population groups a two-part model with generalised gamma distribution was developed for the intake of foods that showed high frequencies of zero observations. The first component of the model

was specified to estimate the probability of consumption and the second to estimate the mean amount consumed given a positive consumption. The first component used mixed-effects logistic regression and the second a generalised gamma mixed-effects regression model. The use of a generalised gamma distribution for modelling intake is an important improvement over existing methods, as it includes many distributions with different shapes and its domain takes non-negative values. The two-part model accommodated the sources of data variation of dietary intake with a random intercept in each component, which could be correlated to allow a correlation between the probability of consuming and the amount consumed. This also improves existing approaches that assume a zero correlation. The utility of the proposed approach was demonstrated by modelling the mean consumption of iron intake from selected episodically consumed food groups using data from the NDNS RP in terms of sex, age and socio-economic status.

2) To identify population groups at high risk of deterioration in nutrient consumption, a linear quantile mixed-effects model was developed to model quantiles of the distribution of intake as a function of explanatory variables. The model utilises the asymmetric Laplace distribution which can accommodate many different distributional shapes, and likelihood-based estimation which is robust to model misspecification. This method is an important improvement over existing methods used in nutritional research as it explicitly models the quantiles in terms of explanatory variables using a novel quantile regression model with random effects. The proposed approach was illustrated by comparing the quantiles of iron intake with Lower Reference Nutrient Intakes (LRNI) recommendations using NDNS RP.

This thesis extended the estimation procedures of both the two-part model with generalised gamma distribution and the linear quantile mixed-effects model to incorporate the complex sample design in three steps: the likelihood function was multiplied by the sample weightings; bootstrap methods were used for the estimation of the variance of the parameters estimates to account for the correlation among observations taken from the same population sampling unit. Finally, the variance estimation of the model parameters was stratified by the survey strata. These procedures were implemented in SAS and R.

3) To evaluate the allocation of resources to alleviate nutritional deficiencies, a quantile linear mixed-effects model with the asymmetric Laplace distribution was used to analyse the distribution of expenditure on iron deficiency medication across health boards in the UK. Expenditure is likely to depend on the iron status of the region; therefore, for a fair comparison among health boards, iron status was estimated using the method developed in objective 2) and used in the specification of the median amount spent. Each health board is formed by a set of general practices (GPs), therefore, a random intercept was used to induce correlation between expenditure from two GPs from the same health board. Finally, the approaches in objectives 1) and 2) were compared with the traditional approach based on weighted linear regression modelling used in the NDNS RP reports. All analyses were implemented using SAS and R.

**Results:** The two-part model with generalised gamma distribution fitted to amount of iron consumed from selected episodically food groups using NDNS RP data, showed that females tended to have greater odds of consuming iron from foods but consumed smaller amounts. As age groups increased, consumption tended to increase relative to the reference group though odds of consumption varied. Iron consumption also appeared to be dependent on National Statistics Socio-Economic Classification (NSSEC) group with lower social groups consuming less, in general. The quantiles of iron intake estimated using the linear quantile mixed-effects model showed that more than 25% of females aged 11-50y are below the LRNI, of whom the 11-18y girls is the most severely affected group in the UK. Predictions of spending on iron medication in the UK based on the linear quantile mixed-effects model showed areas of higher iron intake resulted in lower spending on treating iron deficiency. In a geographical display of expenditure, Northern Ireland featured the lowest amount spent. Comparing the results from the methods proposed here showed that using the traditional approach based on weighted regression analysis could result in spurious associations.

**Discussion:** This thesis developed novel approaches to the analysis of dietary complex survey data to address three important objectives of diet surveillance, namely the mean estimation of food intake by population groups, identification of groups at high risk of nutrient deficiency and allocation of resources to alleviate nutrient deficiencies. To the best of my knowledge this work presents the first application and

extension of the two-part model with generalised gamma distribution, and the linear mixed-effects model to dietary complex survey data. The methods provided models of good fit to dietary data, accounted for the sources of data variability and extended the estimation procedures to incorporate the complex sample survey design. The use of a generalised gamma distribution for modelling intake is an important improvement over existing methods, as it includes many distributions with different shapes and its domain takes non-negative values. The two-part model accommodated the sources of data variation of dietary intake with a random intercept in each component, which could be correlated to allow a correlation between the probability of consuming and the amount consumed. This also improves existing approaches that assume a zero correlation. The linear quantile mixed-effects model utilises the asymmetric Laplace distribution which can also accommodate many different distributional shapes, and likelihood-based estimation is robust to model misspecification. This method is an important improvement over existing methods used in nutritional research as it explicitly models the quantiles in terms of explanatory variables using a novel quantile regression model with random effects. The computational implementation of these methods is also provided to make them readily available in SAS and R. The application of these models to UK national data confirmed the association of poorer diets and lower social class, identified the group of 11-50y females as a group at high risk of iron deficiency, and highlighted Northern Ireland as the region with the lowest expenditure on iron prescriptions.

David Pell

December 2018

# Contents

<b>1</b>	<b>Modelling nutrient data from complex sample national surveys</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Usual intake . . . . .	2
1.3	Measurement error . . . . .	3
1.3.1	Measurement error model . . . . .	7
1.4	Measuring dietary intake from multiple records . . . . .	9
1.5	Dietary assessment tools . . . . .	9
1.5.1	24-Hour recall . . . . .	10
1.5.2	Diet diaries . . . . .	11
1.5.3	Food frequency questionnaire (FFQ) . . . . .	12
1.5.4	Other methods . . . . .	12
1.5.5	Collection period . . . . .	13
1.6	Semi-continuous data . . . . .	14
1.7	Survey designs . . . . .	18
1.7.1	Simple random sampling . . . . .	18
1.7.2	Stratified simple random sampling . . . . .	20
1.7.3	Stratified balanced simple random sampling . . . . .	21
1.7.4	Stratified clustered simple random sampling . . . . .	21
1.7.5	Multistage sampling . . . . .	24
1.7.6	Postal surveys . . . . .	26
1.8	Statistical modelling of usual intake . . . . .	28
1.9	One-part model . . . . .	29
1.10	Two-part model . . . . .	29
1.11	Numerical integration . . . . .	31
1.12	Comparing methods of usual intake estimation . . . . .	32
1.12.1	Traditional approach . . . . .	32
1.12.2	Iowa State University (ISU) method . . . . .	33
1.12.3	National Cancer Institute (NCI) method . . . . .	36
1.12.4	Statistical Program to Assess Dietary Exposure (SPADE) . . . . .	37
1.13	Iron Intake . . . . .	40



1.14	Summary . . . . .	40
1.15	Aims and Objectives . . . . .	41
<b>2</b>	<b>National Diet and Nutrition Survey Rolling Programme</b>	<b>45</b>
2.1	The purpose of the NDNS RP . . . . .	46
2.2	Data Collected . . . . .	46
2.3	Data accessibility . . . . .	48
2.4	Sample design . . . . .	49
2.5	Weighting . . . . .	50
2.6	Impact of the weighting and the complex survey design . . . . .	52
2.7	Socio-economic factors: NSSEC . . . . .	55
2.8	Strengths and limitations of the NDNS RP . . . . .	58
<b>3</b>	<b>Two-part models of complex survey data using a generalised gamma distribution: Dietary Iron intake in the UK</b>	<b>60</b>
3.1	Introduction . . . . .	61
3.2	The generalised gamma distribution . . . . .	63
3.3	Two part model . . . . .	64
3.3.1	Extension to multistage sampling . . . . .	65
3.3.2	Standard error estimation of model parameters . . . . .	67
3.4	Application of the two-part model on iron consumption from selected food groups in the NDNS RP . . . . .	70
3.4.1	Model specification . . . . .	73
3.5	Standard error of parameter estimates . . . . .	77
3.6	Results . . . . .	77
3.6.1	Two-part model comparator analysis . . . . .	86
3.6.2	Number of bootstrap samples for SE estimation . . . . .	95
3.7	Discussion . . . . .	100
<b>4</b>	<b>Quantile regression of dietary intake in complex sample surveys</b>	<b>103</b>
4.1	Introduction . . . . .	103
4.2	The asymmetric Laplace distribution as working model in quantile regression with a random intercept . . . . .	105

4.3	Statistical inference taking the complex sample design into account . . . .	107
4.4	Pseudo likelihood estimation . . . . .	107
4.5	Model selection . . . . .	107
4.6	Estimation of standard errors . . . . .	108
4.7	Modelling of dietary iron consumption . . . . .	109
4.8	Results . . . . .	110
4.8.1	Quantile regression comparator analysis . . . . .	118
4.9	Simulation . . . . .	120
4.10	Discussion . . . . .	127
<b>5</b>	<b>Iron prescription costs across the UK</b>	<b>129</b>
5.1	Introduction . . . . .	129
5.2	Iron deficiency . . . . .	129
5.3	Iron deficiency consequences . . . . .	130
5.4	Global iron deficiency prevalence . . . . .	130
5.5	UK iron deficiency prevalence . . . . .	132
5.6	Iron deficiency treatment . . . . .	134
5.7	Methods . . . . .	137
5.7.1	The data . . . . .	137
5.7.2	Health boards . . . . .	137
5.7.3	Data sources . . . . .	138
5.7.4	Mapping health boards . . . . .	142
5.7.5	Number of patients by GP practice at the time of prescription . . . .	142
5.7.6	Index of Multiple Deprivation (IMD) . . . . .	143
5.7.7	Index of Multiple Deprivation by GP practice . . . . .	143
5.7.8	Prescription of iron medication . . . . .	144
5.7.9	Iron bioavailability . . . . .	144
5.7.10	Dietary Iron . . . . .	147
5.8	Statistical analysis . . . . .	149
5.9	Results . . . . .	150
5.10	Discussion . . . . .	156

<b>6 Discussion</b>	<b>160</b>
6.1 Summary of novel methods . . . . .	160
6.1.1 Software implementation . . . . .	163
6.2 Summary of findings . . . . .	163
6.2.1 Novel approach compared with the traditional approach: two-part model . . . . .	164
6.2.2 Novel approach compared with the traditional approach: Quantile regression . . . . .	165
6.2.3 Expenditure on iron prescriptions by health boards . . . . .	166
6.2.4 Implications of findings . . . . .	166
6.3 Strengths and limitations of this research . . . . .	167
6.4 Comparison to previous results . . . . .	169
6.5 Areas of future research . . . . .	171
<b>7 Conclusion</b>	<b>172</b>
<b>Appendix A Example of the advance letter sent to prospective NDNS partic- ipants</b>	<b>198</b>
<b>Appendix B Variables collected during the NDNS</b>	<b>200</b>
<b>Appendix C Example pages from the NDNS RP food diaries</b>	<b>204</b>
<b>Appendix D Two-part models of complex survey data using a generalised gamma distribution: supplementary tables</b>	<b>221</b>
<b>Appendix E Two-part models of complex survey data using a generalised gamma distribution: quadrature point comparison</b>	<b>238</b>
<b>Appendix F Two-part models of complex survey data using a generalised gamma distribution: data table</b>	<b>240</b>
<b>Appendix G Two-part models of complex survey data using a generalised gamma distribution: code</b>	<b>242</b>

<b>Appendix H</b>	<b>Quantile regression of dietary intake in complex samples: data table</b>	<b>250</b>
<b>Appendix I</b>	<b>Quantile regression of dietary intake in complex samples: code</b>	<b>252</b>
<b>Appendix J</b>	<b>Iron prescription costs across the UK: code</b>	<b>255</b>
<b>Appendix K</b>	<b>Iron prescription costs across the UK: data sources</b>	<b>272</b>
<b>Appendix L</b>	<b>Iron prescription costs across the UK: data table</b>	<b>274</b>
<b>Appendix M</b>	<b>Iron prescription costs across the UK: Estimated regression parameters for the amount spent by health boards in the UK</b>	<b>276</b>

## List of Tables

1	Number of days required to estimate true intake from diary records . . . .	10
2	An example of episodically consumed food reporting from Alcohol intake for 4947 participants aged 11+ in the National Diet and Nutrition Survey Years 1-4 (2008-2012). Adapted from Table 5.13 of the National Diet and Nutrition Survey Rolling Programme Report Steer et al. (2014) . . . .	35
3	Comparison of methods for estimating usual intake . . . . .	39
4	The distribution of UK addresses in the UK Postcode address file, with unweighted NDNS RP Y1-4 (2008-2012) sample, then adjusted percentage following the application of selection weights, by country, adapted from the Table B.1 from the National Diet and Nutrition Survey Y1-4 (2008-2012) (Tipping, 2014) . . . . .	51
5	Iron intakes from food sources in the UK from 6224 participants aged 4 years and older from the National Diet and Nutrition Rolling Programme Years 1-4 (2008-2012), adjusted and unadjusted for the NDNS RP weighting and complex survey design . . . . .	54
6a	Weighted demographic characteristics for females, males and all participants of the NDNS RP Years 1-4 (2008-2012) . . . . .	56
6b	Weighted demographic characteristics for NDNS RP Years 1-4 (2008-2012) . . . . .	57
7	Descriptive characteristics for selected foods using data from NDNS RP Years 1-4, (2008-2012). . . . .	79
8a	Estimated parameters of the two-part model for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	81
8b	Estimated parameters of the two-part model for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	82
8c	Estimated parameters of the two-part model for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	83
8d	Estimated parameters of the two-part model for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-12) . . . . .	84

8e	Estimated parameters of the two-part model for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	85
9a	Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	90
9b	Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	91
9c	Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	92
9d	Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	93
9e	Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012). . . . .	94
10a	Estimated regression parameters of the two-part model for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1 . . . . .	96
10b	Estimated regression parameters of the two-part model for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2 . . . . .	97
11a	Percentage difference between standard error estimates for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1 . . . . .	98

11b	Percentage difference between standard error estimates for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2 . . . . .	99
12	Weighted demographic characteristics of 6109 participants aged 65y and under in the NDNS Rolling Programme Y1-4 (2008-2012) . . . . .	111
13	Estimated regression parameters for 2.5th, 25th, 50th, 75th and 97.5th quantiles and 95% confidence intervals for dietary iron intake in the UK. The model used for quantile regression estimation was the linear mixed-effects quantile regression with ALD with SE estimated using bootstrap. . . . .	117
14	A comparison of regression parameters estimated for the 50 <sup>th</sup> quantile (median) and mean along with 95% confidence intervals for iron intake in National Diet and Nutrition Survey Rolling Programme (NDNS RP) Y1-4 (2008-2012) participants using linear mixed-effects quantile regression with ALD with SE estimated using bootstrap and a weighted linear mixed-effects model. . . . .	119
15	Simulation study scenarios used to evaluate the performance of the lqmm package adapted from (Geraci, 2014) . . . . .	121
16	Scenarios for the simulation study. . . . .	123
17	Coverage probability of 90% confidence intervals calculated from simulated data under six different scenarios. . . . .	126
18	Relative bias of estimated parameters from data simulated under six different scenarios. . . . .	126
19	Haemoglobin levels (g/dL) and status by age and sex from NDNS RP Years 1-4 (2008-2012) . . . . .	133
20	Dietary factors affecting the absorption of Iron . . . . .	136
21	Weighted dietary iron intake adjusted using an algorithm to adjust for concurrent food intake (g/dL). . . . .	145
22	English health boards financial assessment. . . . .	157
23	Variables collected available from the NDNS Rolling Programme Years 1-4, (2008-2012) . . . . .	200

24a	Estimated parameters of the two-part model for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1 . . . . .	222
24b	Estimated parameters of the two-part model for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2 . . . . .	223
24c	Estimated parameters of the two-part model for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1 . . . . .	224
24d	Estimated parameters of the two-part model for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2 . . . . .	225
24e	Estimated parameters of the two-part model for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1 . . . . .	226
24f	Estimated parameters of the two-part model for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2 . . . . .	227
24g	Estimated parameters of the two-part model for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1 . . . . .	228
24h	Estimated parameters of the two-part model for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2 . . . . .	229



25a	Percentage difference between standard error estimates for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1 . . . . .	230
25b	Percentage difference between standard error estimates for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2 . . . . .	231
25c	Percentage difference between standard error estimates for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1 . . . . .	232
25d	Percentage difference between standard error estimates for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2 . . . . .	233
25e	Percentage difference between standard error estimates for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1 . . . . .	234
25f	Percentage difference between standard error estimates for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2 . . . . .	235
25g	Percentage difference between standard error estimates for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1 . . . . .	236
25h	Percentage difference between standard error estimates for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 bootstrap samples: Part 2 . . . . .	237
26	Difference in time taken for model convergence using 5,10,15 and 20 quadrature points when estimating iron intake from vegetables of 50 bootstrap replicates using data from the NDNS RP Y1-4 (2008-2012) for 4156 participants aged 1.5 years and over. . . . .	238

27	Sample of NDNS RP Y1-4 (2008-2012) data used to estimate iron intake from the bread food group . . . . .	241
28	Sample of NDNS RP Y1-4 (2008-2012) data used to estimate quantiles of dietary iron intake . . . . .	251
29	Sample of data used to estimate health board spending on iron prescriptions . . . . .	275
30	Estimated regression parameters for Median with standard errors for amount spent by health boards in the UK . . . . .	277

## List of Figures

1	Mean intake of alcohol (g) by day of the week for all 3603 participants aged 18+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012). . . . .	4
2	Mean intake of vegetables (g) by day of the week for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012). . . . .	5
3	Mean intake of strawberries (g) by month for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012) . . . . .	6
4	Distribution of intake for energy (kcal) for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012). . . . .	15
5	Distribution of intake for omega-3 fatty acid (g) for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012). . . . .	16
6	Distribution of intakes for alcohol (g) for all 3603 participants aged 18+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012). . . . .	17
7	Simple random sampling representation adapted from Dodd (2011) . . . . .	19
8	Illustration of simple random sampling of the UK . . . . .	20
9	Stratified simple random sampling representation adapted from Dodd (2011) . . . . .	22
10	Stratified balanced simple random sampling representation adapted from Dodd (2011) . . . . .	23
11	Stratified clustered simple random sampling representation adapted from Dodd (2011) . . . . .	25
12	Illustration of multistage sampling using the National Health and Nutrition Examination Survey (NHANES) adapted from CDC and National Center for Health Statistics (2013) . . . . .	26

13	Intake of servings of fruit and vegetable, by one 24-hour recall (broken line), the mean of two non-consecutive 24-hour recalls (dotted line) and usual intake estimated by ISU model (solid line) for 14963 participants in the Continuing Survey of Food Intakes by Individuals (1994-1996) adapted from Guenther et al. (2006). . . . .	34
14	National Diet and Nutrition Survey Rolling Programme Study Design . . .	47
15	Mean daily vitamin D intake in 436 Females aged 65+ from NDNS RP Years 1-4 (2008-2014) by year of survey . . . . .	59
16	Generalised gamma distribution with varying parameter values displaying the standard gamma, Weibull, exponential and log-normal distributions.	62
17	Iron intake from vegetables by number of days of consumption using data from NDNS RP 2008-12 for 4156 participants aged 1.5 years and over . . . . .	66
18	Illustration of the process to create a survey adjusted single weight for sampled PSUs in the NDNS participants. . . . .	69
19	Partial histograms with kernel density estimates (dark line) for iron intake of selected food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over. . . . .	72
20	The percentage contribution to total iron intake by all food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 core participants aged 1.5 years and over . . . . .	74
21a	The percentage contribution to total iron intake by selected food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over. . . . .	75
21b	The proportion of zero consumption days by selected food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over. . . . .	76
22	Weighted histogram of mean daily iron intake for 6109 participants aged 65 and under in the NDNS Y1-4 (2008-12) (shaded area) and fitted asymmetric Laplace distribution (dashed line). . . . .	112
23	Estimated quantile iron intake with 95% confidence bands and observed median individual intake. . . . .	113

24	Estimated iron intake quantiles with 95% confidence bands by age groups with Lower Reference Nutrient Intake (LRNI) recommendations (broken lines). . . . .	114
25	Estimated iron intake quantiles with 95% confidence bands by age groups with LRNI recommendations (broken lines) that differ by sex. . . . .	115
26	Schematic illustration of a simulation study carried out to examine the coverage and relative bias of maximum pseudo likelihood estimation with bootstrapped variance estimates . . . . .	122
27	The global extent of the public health problem of iron deficiency anaemia in non-pregnant woman of child bearing age. . . . .	131
28	Timeline depicting the availability and selection of prescription data by country. . . . .	138
29	The distribution of UK GP Practices by postcode, with black borders indicating health board boundaries. GP surgeries indicated by a blue dot.	139
30	Sources of data for England, Northern Ireland, Scotland and Wales . . . .	140
31	Schematic diagram illustrating the merging of files containing iron prescriptions, GP addresses, Index of Multiple Deprivation ranking and registered patients . . . . .	141
32	Mean dietary Iron intake by UK government office region. Data taken from NDNS RP Y1-4 . . . . .	146
33	Iron (mg) intake as determined through an algorithm that considers nutrient interactions for individuals in the NDNS RP Y1-4 (2008-2012) and fitted asymmetric Laplace distribution (dashed line). . . . .	148
34	Total amount spent on Iron prescriptions by health boards in the UK from Oct 15 - Sept 16 overlain with an asymmetric Laplace distribution. . . . .	151
35	Quintiles of median amount spent from Oct 15 - Sept 16 on iron prescriptions in each health board across the UK adjusted for Index of Multiple Deprivation (IMD) and the number of registered patients per health board.	154
36	Quintiles of median amount spent from Oct 15 - Sept 16 on iron prescriptions in each health board across the UK adjusted for IMD, bioavailable iron intake and the number of registered patients per health board. . . . .	155

37	The letter sent in advance of an interviewer visit providing participant information on the NDNS RP years 1-4 (2008-2012). . . . .	199
38	The title page of the food diary used for dietary assessment of adults in the NDNS RP years 1-4 (2008-2012). . . . .	205
39	Participant instructions to be read before completing the NDNS RP years 1-4 (2008-2012) diet diary - part 1. . . . .	206
40	Participant instructions to be read before completing the NDNS RP years 1-4 (2008-2012) diet diary - part 2. . . . .	207
41	An example of a completed NDNS RP years 1-4 (2008-2012) diet diary - part 1. . . . .	208
42	An example of a completed NDNS RP years 1-4 (2008-2012) diet diary - part 2. . . . .	209
43	An example of a completed NDNS RP years 1-4 (2008-2012) diet diary - part 3. . . . .	210
44	Questions covering whether the day's intake is typical from a NDNS RP years 1-4 (2008-2012) diet diary - part 1. . . . .	211
45	Questions covering whether the day's intake is typical from a NDNS RP years 1-4 (2008-2012) diet diary - part 2. . . . .	212
46	An example of a completed recipe from a NDNS RP years 1-4 (2008-2012) diet diary. . . . .	213
47	Information covering the detail requested for commonly consumed foods from a NDNS RP years 1-4 (2008-2012) diet diary. . . . .	214
48	An example of the food atlas from a NDNS RP years 1-4 (2008-2012) diet diary. . . . .	215
49	General dietary intake questions from a NDNS RP years 1-4 (2008-2012) diet diary . . . . .	216
50	Front page from the NDNS RP years 1-4 (2008-2012) children's diet diary	217
51	An example of a completed NDNS RP years 1-4 (2008-2012) children's diet diary - part 1. . . . .	218
52	Front page from the NDNS RP years 1-4 (2008-2012) infant's diet diary	219
53	An example of a completed NDNS RP years 1-4 (2008-2012) infants's diet diary - part 1. . . . .	220

54	A boxplot showing the difference in estimated standard errors of iron intake from vegetables from four models using 5,10,15 and 20 quadrature points using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over. . . . .	239
----	---	-----

# Acronyms

**24HR** 24 Hour Recall.

**BNF** British National Formulary.

**BRR** Balanced Repeated Replication.

**CAPI** Computer Assisted Personal Interview.

**CCG** Clinical Commissioning Groups.

**CDF** Cumulative distribution function.

**DD** Diet Diaries.

**DLW** Doubly Labelled Water.

**DNFCS** Dutch National Food Consumption Survey.

**EWL** Elise Widdowson Laboratory.

**FFQ** Food Frequency Questionnaire.

**FPQ** Food Propensity Questionnaire.

**FSA** Food Standards Agency.

**GOR** Government Office Region.

**GP** General Practitioner.

**GBHS** Grouped Balanced Half Sample.

**HNR** Human Nutrition Research.

**IDA** Iron Deficiency Anaemia.

**IMD** Index of Multiple Deprivation.

**IoM** Institute of Medicine.



**ISU** Iowa State University.

**LRNI** Lower Reference Nutrient Intake.

**LSOA** Lower Layer Super Output Area.

**NatCen** National Centre for Social Research.

**NCI** National Cancer Institute.

**NDNS RP** National Diet and Nutrition Survey Rolling Programme.

**NHANES** National Health and Nutrition Examination Survey.

**NRC** National Research Council.

**NSSEC** National Statistics Socio-Economic Classification.

**OPEN** Observing Protein and Nutrition.

**PCT** Primary Care Trusts.

**PHE** Public Health England.

**PSU** Pseudo-likelihood ratio test.

**PSU** Primary Sampling Unit.

**QOF** Quality and Outcomes Framework.

**RNI** Reference Nutrient Intake.

**SACN** Scientific Advisory Committee on Nutrition.

**SPADE** Statistical Program to Assess Dietary Exposure.

**SRS** simple random sampling.

**SBSRS** stratified balanced simple random sampling.

**UCL** University College London.

## Glossary

**24 Hour Recall (24HR)** A dietary assessment method that collects food and drink intake for the previous 24 hours.

**British National Formulary (BNF)** A reference for information and advice on medication prescribing and pharmacology in the UK.

**Balanced Repeated Replication (BRR)** A resampling method used for variance estimation for complex survey designs. Requires 2 PSU per stratum.

**Computer Assisted Personal Interview (CAPI)** Method of recording supplementary questions covering household composition, employment status and dietary determinants.

**Clinical Commissioning Groups (CCG)** Groups responsible for commissioning health services in England.

**Cumulative distribution function (CDF)** A function with a value is the probability that a corresponding continuous random variable has a value less than or equal to the argument of the function.

**Diet Diaries (DD)** A prospective method of dietary assessment requiring participants to record all food and drink consumption in a diary for several consecutive days.

**Doubly Labelled Water (DLW)** A biomarker used to validate energy intake.

**Dutch National Food Consumption Survey (DNFCS)** The Dutch equivalent of the NDNS.

**Elsie Widdowson Laboratory (EWL)** Medical Research Council funded research nutrition unit that superseded Human Nutrition Research.

**Food Frequency Questionnaire (FFQ)** A long-term method of dietary assessment that requires participants to complete a questionnaire regarding food consumption typically over the previous 6-12 months.

**Food Propensity Questionnaire (FPQ)** Similarly to FFQ collects food consumption over the previous 6-12 months.

**Food Standards Agency (FSA)** Regulatory body for UK food standards.

**Government Office Region (GOR)** Regions of the UK used by government and the NDNS RP.

**General Practitioner (GP)** A medical doctor, usually based in the community and is generally the first contact for patients. They are mainly responsible for the prescription of medication for chronic conditions including anaemia.

**Grouped Balanced Half Sample (GBHS)** A method of combining PSUs within each stratum into two groups allowing for BRR to be carried out on data that is not sampled with 2 PSU per stratum.

**Human Nutrition Research (HNR)** Medical Research Council funded research nutrition unit.

**Iron Deficiency Anaemia (IDA)** Anaemia due to deficiency of Iron, symptoms include shortness of breath and lethargy, severe anaemia can cause death.

**Index of Multiple Deprivation (IMD)** An index of deprivation in the UK based on 7 criteria including health deprivation and disability.

**Institute of Medicine (IoM)** US agency produced method of estimating usual intake.

**Iowa State University (ISU)** US University that produced a method of estimating usual intake.

**Lower Reference Nutrient Intake (LRNI)** Threshold value set by SACN used to characterise the risk of nutrient deficiency in a population.

**Lower Layer Super Output Area (LSOA)** A small geographic area that contains approximately 1500 people. Index of Multiple Deprivation scores are assigned at the LSOA level.

**National Centre for Social Research (NatCen)** UK organisation that runs many national surveys and is responsible for NDNS data collection.

**National Cancer Institute (NCI)** US agency that produced methods for usual intake estimation.

**National Diet and Nutrition Survey Rolling Programme (NDNS RP)** National survey to assess food and drink intakes in the UK.

**National Health and Nutrition Examination Survey (NHANES)** National survey to assess food and drink intakes in the US.

**National Research Council (NRC)** US nongovernmental agency responsible for a method of usual intake estimation.

**National Statistics Socio-Economic Classification (NSSEC)** Similar to social classes, NSSEC is used to classify and rank class according to occupation.

**Observing Protein and Energy Nutrition (OPEN)** US Study attempting to quantify dietary assessment measurement error through the use of biomarkers.

**Primary Care Trusts (PCT)** Prior to the creation of Clinical Commissioning Groups in England, PCTs were responsible for responsible for commissioning primary, community and secondary health services.

**Public Health England (PHE)** An English government agency responsible for public health in England that commissions the NDNS.

**Pseudo-likelihood ratio test (PLRT)** Used to compare the goodness of fit of two models, one of which (the null model) is a special case of the other (the alternative model).

**Primary Sampling Unit (PSU)** Sampling unit selected in the first stage of a multi-stage sample containing a set of sampling individuals of interest..

**Quality and Outcomes Framework (QOF)** A programme collecting GP practice achievement results.

**Reference Nutrient Intake (RNI)** Threshold value issued by SACN appropriate for the majority of the population.

**Scientific Advisory Committee on Nutrition (SACN)** UK Nutrition advisory group.

**Statistical Program to Assess Dietary Exposure (SPADE)** Computational implementation of methods of estimation of usual intake in R.

**Simple random sampling (SRS)** Common sampling technique that does not contain any multistage sampling of complex sampling.

**Stratified balanced simple random sampling (SBSRS)** A sampling design where the population is first grouped in to strata with equal numbers of individuals sampled using simple random sampling.

**University College London (UCL)** London university that collaborated on the NDNS.

# 1 Modelling nutrient data from complex sample national surveys

## 1.1 Introduction

Diet plays an important role in many health and disease related conditions including cancer, diabetes and heart disease (Lopez et al., 2006). Governments throughout the world have recognised the importance of diet and health and have commissioned surveys to collect dietary intakes for their populations.

National surveys of dietary intake allow the evaluation of nutrient intakes of a population to inform health policies, such as the provision of healthy eating advice and interventions to improve nutrition status. Dietary surveillance should be capable of providing specific information relating to specific groups, for example the proportion of adults consuming at least 5 portions of fruit and vegetable per day has been shown to fluctuate over time from 27% to 31% (Roberts et al., 2018) and across income groups from 24% in the lowest income group to 38% in the highest (Bates et al., 2014a).

The impact diet has on disease, notwithstanding acute allergy and poisoning, is caused by long-term exposure thus a measure of usual intake is required rather than a single instance of intake. Yet collecting accurate information on usual dietary intake that is free from measurement error requires dealing with some sizeable challenges as it is difficult to collect a representative sample of people who are willing to accurately report their diet for long enough to reflect their usual intake (Bingham, 1987). It is therefore important to reduce the burden upon survey participants when collecting dietary information to facilitate the capture of as much dietary intake detail as possible but also to develop statistical methods that make efficient use of the available data and minimise the bias of dietary intake estimates.

The following challenges to estimating usual intake from a dietary assessment measure have been identified in the literature (Tooze et al., 2006):

- Accounting for days without consumption of a particular food or nutrient during the period of observation

- Allowing for consumption-day amount data that are generally positively skewed and have extreme values in the upper tail of the intake distribution
- Distinguishing within-person variability, which consist of day-to-day variation in intake and random reporting errors, from between-person variation
- Allowing for the correlation between the probability of consuming a food and the consumption-day amount
- Relating covariate information (e.g., sex, age or socio-economic status) to usual intake

A further challenge is accounting for the complex sampling design necessary to carry out surveys over a large geographic area, within a single statistical analysis method that meets the above requirements.

## 1.2 Usual intake

Due to the natural variability that occurs in diets it is important to collect more than one day of intake to derive a measure of usual intake. Diets often vary because of factors such as the day of the week and seasonality as foods come in and out of season (Shahar et al., 2001; Hoare et al., 2004). This leads to skewed distributions of intake as illustrated using data from the National Diet and Nutrition Survey Rolling Programme Y1-4 (2008-2012), **Figures 1, 2 and 3** show how alcohol and vegetable intakes vary over the week and the extent to which strawberry intake differs across the months of the year. In Figure 1 it can be seen that mean daily alcohol intake in adults is approximately 7g on Mondays steadily increasing throughout the week to approximately 18g at the end of the working week on Friday then increasing yet further to around 24g on Saturdays (Figure 1). Figure 2 shows that mean vegetable intake across all ages is higher on a Sunday at around 170g dropping throughout the week to a low of around 160g on Saturdays likely influenced by the typical British Sunday lunch. Similarly Figure 3 shows that the mean monthly intake of strawberries is lowest at the start of the year then peaks in the summer months, reflective of the growing season of

soft fruit, though this should be considered by surveys monitoring fruit intake as intakes would differ depending on the month it was carried out.

An individual's usual dietary intake has been defined as the individual's long-run average dietary intake (Carrquiry, 2003). By determining usual dietary intakes for a population it is possible to determine adherence to dietary recommendations and get an indication of health status, toxicity risk and the effectiveness of public health policies (Riley, 2010). This can be done for a sample of the population with a single record collection, if it is representative of the population, seasons and days of the week (Biro et al., 2002), although because of the variation in intakes the variance of the sample will be inflated (Carrquiry, 2003; Mackerras and Rutishauser, 2005). In longitudinal studies examining the impact of diet on disease it is common to measure food intake and then to follow individuals over time to draw associations between what has been consumed and the development of disease; for example: (Key et al., 1996) and (Anweiler et al., 2012). The assumption underlying usual intake states that diet remains constant over an individual's life however, this may not be the case. Willett et al. (1988) found that diets did vary over 4 years in a group of 150 women, but the degree of variance depended upon the nutrient. They reported correlation coefficients between the initial dietary assessment and follow up four years later, that ranged from 0.28 for iron intake with supplements to 0.61 for total carbohydrate consumption.

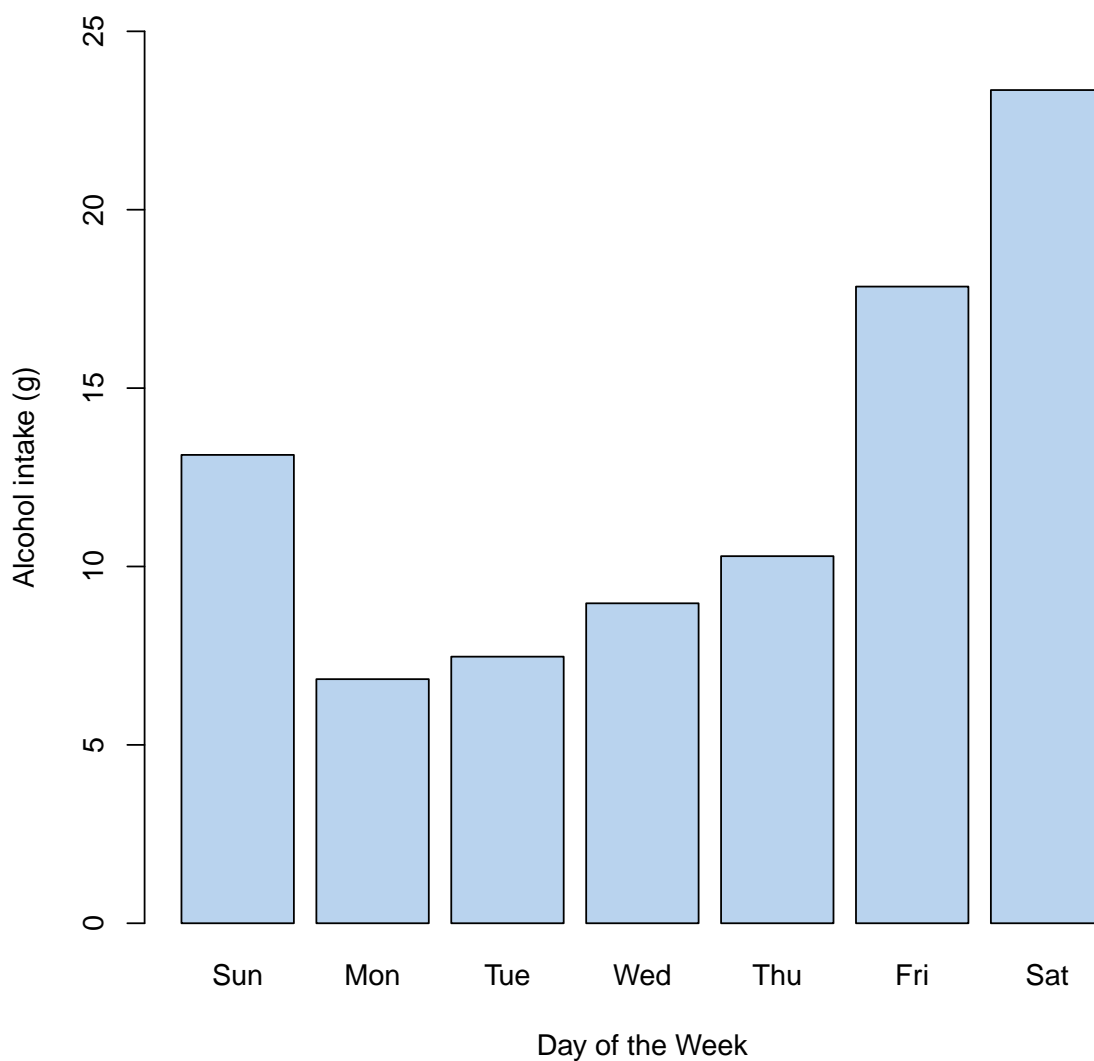
### **1.3 Measurement error**

The limitations to the measurement of usual intake arise because currently there is no accurate method to collect this information without requiring the participant to knowingly provide the data. Once the participant is conscious that their food is being recorded they may deviate from their usual diet either through deliberately not recording a food that has been eaten (intentional under-reporting), by a deviation from usual intake - typically to one that contains more perceived healthy foods and fewer perceived unhealthy foods (intentional alteration of diet) or they may forget to record the foods consumed (unintentional under-reporting) (Macdiarmid and Blundell, 1998). It is possible to examine the extent to which this occurs through studies that observe the partici-



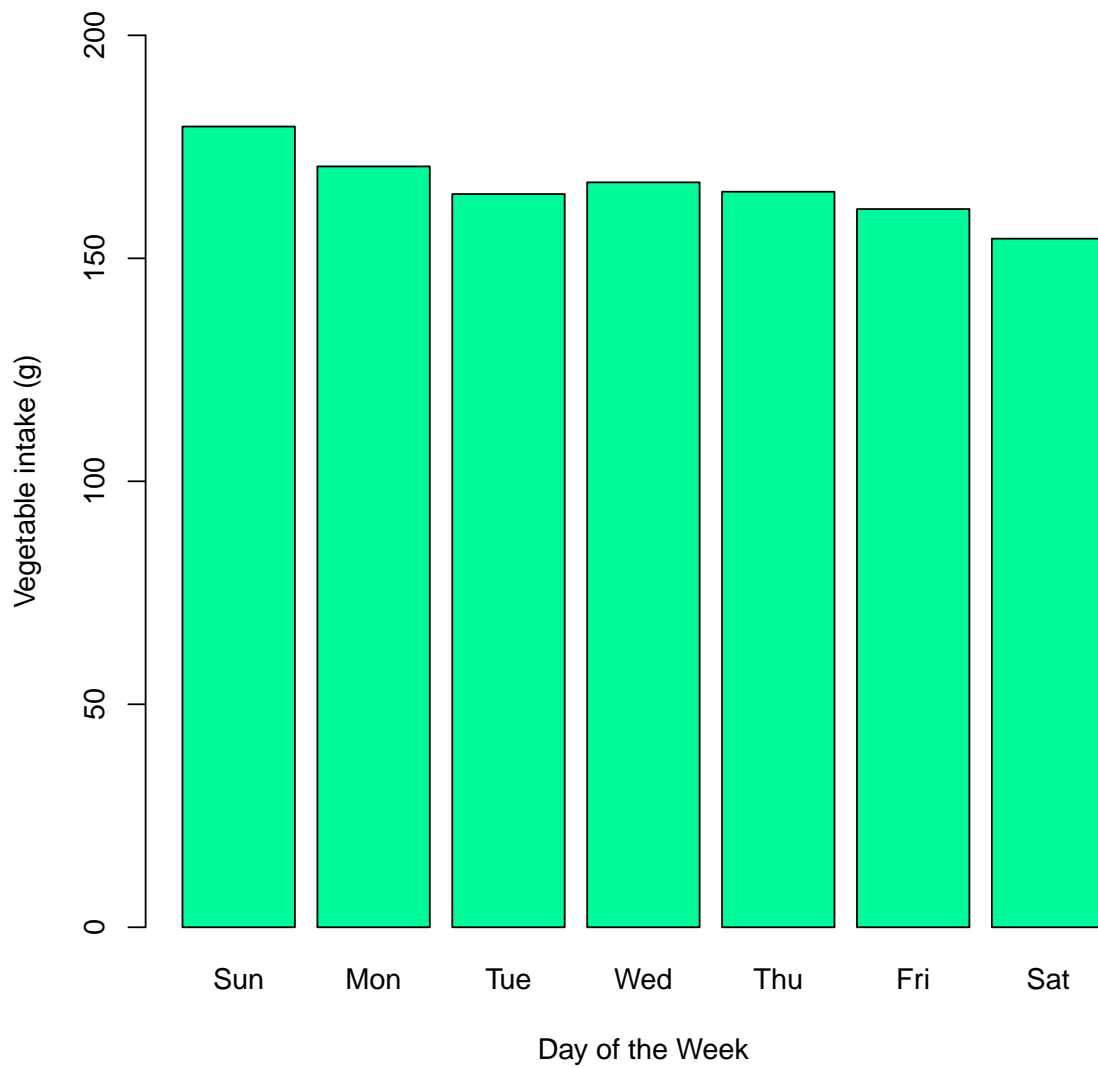
**Figure 1**

Mean intake of alcohol (g) by day of the week for all 3603 participants aged 18+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012).



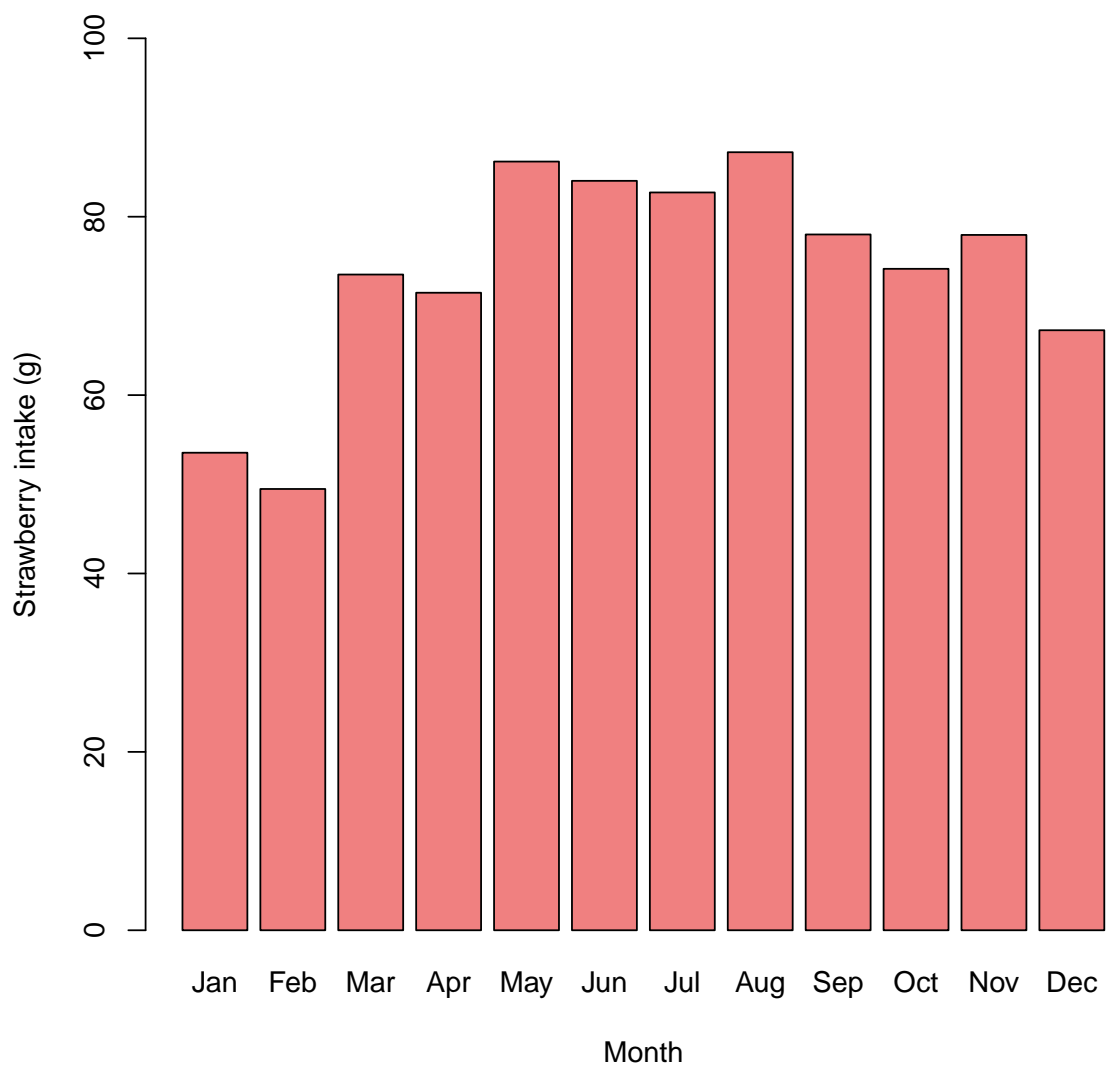
**Figure 2**

Mean intake of vegetables (g) by day of the week for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012).



**Figure 3**

Mean intake of strawberries (g) by month for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012)



part's meals but these are more expensive than surveys and do not represent a typical free living situation. Alternatively, there are a small number of nutritional biomarkers available for validation of intakes such as Doubly Labelled Water (DLW) which can be used to objectively measure total energy expenditure. Based on the assumption that energy expenditure equates to energy intake in weight stable individuals, DLW is an unbiased measure of energy intake that requires subjects to drink a small quantity of water labelled with a stable isotope, then by collecting excreted urine, energy balance can be determined (Buchowski, 2014; Lennox et al., 2014). This, however, is prohibitively expensive to be used for all subjects in large national surveys (Rennie et al., 2007) and therefore typically, a subset of the survey is selected and given DLW. This was the case in the National Diet and Nutrition Survey Rolling Programme (NDNS RP) in 2008/09 and 2010/11 where approximately 10% of participants received DLW and as a result it was found that energy (kcal) was under-reported by 27% on average across age and sex groups (Lennox et al., 2014). These findings are similar to previous work that has shown that study participants under-report energy intake by, on average, 30%; with the degree of under-reporting much higher in teenage and adult females than in males and children (Black and Cole, 2001; Rennie et al., 2007).

### 1.3.1 Measurement error model

The discrepancy that arises between the diets people report and their true intake can be described using the classical measurement error model (Keogh and White, 2014). In the following model  $X_i$  indicates a vector relating to true intake for individual  $i$  and the reported intake for individual  $i$  is a vector denoted by  $W_{i1}$  on day 1,  $W_{i2}$  on day 2 and so on. It is assumed that the error on each day of reported intake will be identically distributed. Using a logistic regression model with a binary outcome  $Y_i$ , the association between exposure and outcome is given by:

$$\log \left\{ \frac{\Pr(Y_i = 1 | X_i, Z_i)}{1 - \Pr(Y_i = 1 | X_i, Z_i)} \right\} = \alpha + \beta' X_i + \gamma' Z_i \quad (1)$$

where  $Z_i$  is a vector containing error free covariates and  $\beta$  and  $\gamma$  are vectors of regression parameters. In dietary assessment it is not uncommon for the dietary intake from a single diary, measured with error, represented by  $W_{i1}$  to replace  $X_i$  the unbiased

measure of intake:

$$\log \left\{ \frac{\Pr(Y_i = 1 | \mathbf{W}_{i1}, \mathbf{Z}_i)}{1 - \Pr(Y_i = 1 | \mathbf{W}_{i1}, \mathbf{Z}_i)} \right\} = \alpha + \boldsymbol{\beta}^* \mathbf{W}_{i1} + \boldsymbol{\gamma}' \mathbf{Z}_i \quad (2)$$

This will mean that the estimator for  $\boldsymbol{\beta}^*$  will also be biased. The classical measurement error model compartmentalises  $\mathbf{W}_{ij}$  into :

$$\mathbf{W}_{ij} = \mathbf{X}_i + \epsilon_{ij} \quad (3)$$

that is the recorded intake for individual  $i$  on day  $j$  is the true intake plus the error term  $\epsilon_{ij}$  which has mean 0 and constant variance  $\sigma_\epsilon^2$ . The error is random and can be placed into one of two categories: within-person or between-person variance. Within-person variance arises due to the inconsistency of an individual's diet over time whereas between-person variance reflects the degree to which an individual varies from the sample mean and large between-person variances would reflect a heterogeneous sample. It has been shown that the error can depend upon true exposure and further that  $\epsilon_{ij}$  can be divided into a systematic part if extra information relating to true intake is known, therefore:

$$\mathbf{W}_{ij} = \psi + \theta \mathbf{X}_i + \epsilon_{ij} \quad (4)$$

where  $\psi$  indicates a constant shift which may occur, for example, from a food that has an incorrect nutrient value attached to it or from an inaccurate measurement tool used to record the amount of food consumed. For example, scales that are not calibrated or a participant that records a tablespoon as a dessertspoon.  $\theta \neq 1$  is an error that is dependent upon the true intake and as such impacts upon the slope, for example, a participant who inaccurately reports smaller portion sizes of a food perceived to be unhealthy but accurately reports portion sizes for foods that are perceived to be healthy. If either of these conditions occur then simply increasing the number of records will not remove the error. Commonly, the reported energy intake amount is lower than energy expenditure determined by DLW suggesting that the participant has not recorded all of the energy consumed during the recording period. However matching values for energy intake and expenditure does not indicate an unbiased energy value, as it may be the case that the individual has deviated from their usual intake by consuming less energy but that they truthfully reported everything that was consumed, i.e. unintentional under-reporting.

## 1.4 Measuring dietary intake from multiple records

To characterise the diet of the UK, thresholds have been determined by the Scientific Advisory Committee on Nutrition (SACN) (Scientific Advisory Committee on Nutrition (SACN), 1991) that are sufficient to meet the nutrient intake requirements. These thresholds provide values that are sufficient for the majority (97.5%) of the population (Reference Nutrient Intake (RNI)) and a lower level at which the risk of deficiency is increased (Lower Reference Nutrient Intake (LRNI)), thought to be sufficient for 2.5% of the population. To classify an individual as being below the LRNI for a nutrient, a reliable measure of their usual intake is required; however, measuring usual intake can be problematic as the tools used to collect intake are subject to substantial within-person variation. This can be due to changes in the diet which can vary considerably from day to day with some foods and nutrients consumed almost every day whilst others consumed less often, perhaps once or twice per week or less. Surveys of dietary intake should aim to capture a minimum of two non-consecutive days of intake as this provides an indication of the extent to which the day-to-day variations in intake occur. The number of days required to capture usual intake depends on the food or nutrient but is thought to vary greatly (Nelson et al., 1989). For example, the number of collection days required to adequately capture energy in female children is 10 days whereas 16 days are necessary to record usual iron intake (Nelson et al., 1989) (see **Table 1**). However, because dietary survey response rates have been shown to be linked to participant burden, recruiting a representative sample of participants becomes difficult as the number of days of dietary record increases.

## 1.5 Dietary assessment tools

The methods used to capture diet are referred to as dietary assessment methods and may be categorised into two groups; either short-term methods such as the 24 Hour Recall (24HR) or Diet Diaries (DD) or longer-term methods e.g. Food Frequency Questionnaire (FFQ). These can be further classified into prospective methods that collect dietary intake at the time of consumption, for example, DD or retrospective either in the short-term using 24HR or longer term, typically 6 months to one year with an FFQ.

**Table 1**

Number of days required to estimate true intake from diary records

Nutrient	All		Males		Females	
	1-4y	5-17y	18y+	5-17y	18y+	
Energy	7	9	4	10	6	
Protein	5	6	6	15	8	
Fat	7	8	6	12	7	
Carbohydrate	6	10	4	9	6	
Iron	6	10	10	16	9	
Carotene	20	40	18	72	38	
Vitamin C	3	9	12	12	7	
Vitamin E	10	13	8	16	16	
Calcium	4	4	5	12	8	

The number of days of intake to be collected for reported intake to meet true intake ( $r \geq 0.9$ ) for selected nutrients. Intakes for infants and young children are often reported together due to low variation between the sexes. Adapted from Nelson et al. (1989)

Retrospective methods rely on the participant's ability to remember their consumption, usually through a short interview, making them less burdensome than the diary method. In contrast, the DD allows participants to record all the food and drink they consume over the recording period including as much detail as required and because of this they are a more precise method of dietary assessment and may be preferred where multiple days of intake are required. By taking repeated measures to account for the day-to-day variation, the daily records are correlated within individuals, which needs to be considered in any analysis to obtain valid estimates of intake.

### 1.5.1 24-Hour recall

Used in many national dietary surveys in Europe (Biro et al., 2002) and the US (Conway et al., 2003), the 24HR collects information from participants about food and drink consumed in the previous 24 hour period. It is typically carried out as an interview by a trained nutritionist or dietitian who methodically goes through a series of questions refining the record after each round (or pass) of questions, known as the multi pass

method (Moshfegh et al., 2008). Participants are recruited into the dietary study then contacted by the interviewer, ideally, at a time unknown to the participant (Conway et al., 2003) to reduce the possibility of the participant modifying their diet beforehand. The interview needs coding into a digital format so it can be linked to nutritional information though online and mobile methods have been developed reducing the time processing time (Subar et al., 2016).

### **1.5.2 Diet diaries**

Diet diaries require the participant to record all food and drink at the time of consumption, rather than relying on memory. The participant is provided with a diary that typically contains questions regarding eating habits and contextual questions to prompt an accurate record (Bingham, 1987). Diaries used in the NDNS RP have sections prompting for the time of consumption along with socio-contextual questions. The NDNS RP diaries require a trained interviewer to explain to the participant what information is needed for adequate completion of the diary and then to check that the diary has been filled in correctly and answer any questions regarding completion that may have arisen during the recording process (Nelson et al., 1989). Diet diaries have a high participant burden and as such are typically used to collect intake data for a few days only, this makes diet diaries unsuitable for collecting long term average intakes. The NDNS RP diet diary collects four consecutive days of dietary intake with estimated portion sizes to reduce participant burden, as opposed to the previous NDNS surveys that collected seven days of intake with weighed portion sizes. The majority of studies using NDNS RP data take the dietary assessment values to represent true intake and do not consider measurement error in the analysis and results (Adams and White, 2015; McGeoghegan et al., 2015; Murakami and Livingstone, 2016; Syrad et al., 2016; Ziauddeen et al., 2017; Hobbs et al., 2018), highlighting the importance of developing accessible methods capable of dealing with measurement error in diet diaries.



### **1.5.3 Food frequency questionnaire (FFQ)**

An FFQ requires participants to record the frequency with which they consumed foods on average in the past, usually 6 months to one year. Participants select one category that best describes their frequency of consumption, typically ranging from *more than once per day* to *not in the previous 12 months*. The list of foods can encompass the whole diet or, where the FFQ excels, a select list of foods that are rich in a particular nutrient of interest. For example calcium is present in significant amounts in a limited number of foods and so an FFQ aimed at collecting calcium intake would contain markedly fewer items than an FFQ examining the whole diet (Taylor and Goulding, 1998). Conversely, attempting to get a complete measure of dietary intake would need to cover all possible foods, which is likely to cause participants to lose interest and result in less accurate intakes (Cade et al., 2004). As usual intake covers the long term average of an individual an FFQ would seem an appropriate tool to use, however, the main limitation with FFQs is that they collect nutrient intake data which are weakly correlated with objective measures of intake that include energy intake using DLW and protein intake which are both determined from urine excretion (Schatzkin et al., 2003). These are findings from the Observing Protein and Energy Nutrition (OPEN) study that compared repeated FFQs and 24HRs with intake biomarkers (Subar et al., 2003).

### **1.5.4 Other methods**

There are other methods of recording diet, used to lesser degrees, which include the duplicate diet; where the participant collects two portions of each food and drink item they intend to consume - the first is consumed and the second is given to the investigator to be analysed for nutrient composition. Though the cost of analysis make this method unsuitable for national surveys it does provide the most accurate food composition data reflecting, as close as possible, the participant's intake (Abdulla et al., 1981). There are recently developed methods that use advances in technology including cameras and mobile phones to unobtrusively observe intake such as: FIRSST (Rockett et al., 2003), TADA (Mariappan et al., 2009), FIVR (Weiss et al., 2010), DDRS (Shang et al., 2011) and the eButton (Chen et al., 2013). These methods may prove

better in time at collecting usual dietary intakes and solving some of the above challenges, but as the UK national dietary survey, the NDNS RP, currently uses diet diaries, developing statistical tools which are able to provide better estimates of usual intakes from data already collected is required.

### 1.5.5 Collection period

When conducting dietary surveys, there remains a trade off between twin goals of collecting reliable, precise dietary intakes and recruiting sufficient numbers of people to make the sample representative of the population. The number of people willing to take part in a study falls as the burden imposed by recording a greater number of days of intake increases (Bingham, 1987). It is important to determine the number of days required to obtain reliable measures of the dietary intake. Nelson et al. (1989) examined the number of days required to produce correlation coefficients from 0.75 to 0.95 between observed intake and true intake using diet diaries. They examined data from 6 studies and found that the number of days required for  $r = 0.95$  depended upon age and sex but also the nutrient being examined (see **Table 1**). Examining the number of 24HRs necessary to provide an accurate indication of usual intake in overweight and obese individuals, Jackson et al. (2008) proposed that eight days was sufficient to produce a correlation coefficient of 0.9 between measured intake and true intake for individual macronutrients, based on 50 individuals who each completed ten 24HRs. Furthermore the authors considered the error that arose between intakes on weekends versus weekdays, though because the participants were reluctant to be interviewed at the weekend, Fridays were omitted from weekdays and weekends were represented by Sundays alone. Similarly Ma et al. (2009), found that one 24HR provided an under-estimation of intake and that three recalls were needed to accurately estimate energy intake. They further reported that increasing the number of recalls did improve the accuracy of energy intakes. It is also worth noting that DLW is a suitable validation tool for energy, a habitually consumed food component, only. The prevailing view in the US in dietary assessment is that a minimum of two 24 hour recalls should be carried out on non-consecutive days and ideally in conjunction with a FFQ to minimize the within-individual variation and to indicate consumption frequency over a longer period (Dwyer

et al., 2003). Conversely in the UK, diet diaries are preferred with many major surveys, (Price et al., 1995; Bingham, 1997; Bingham et al., 1997; Stallone et al., 1997; Public Health England, 2014), opting for between four to seven diary days.

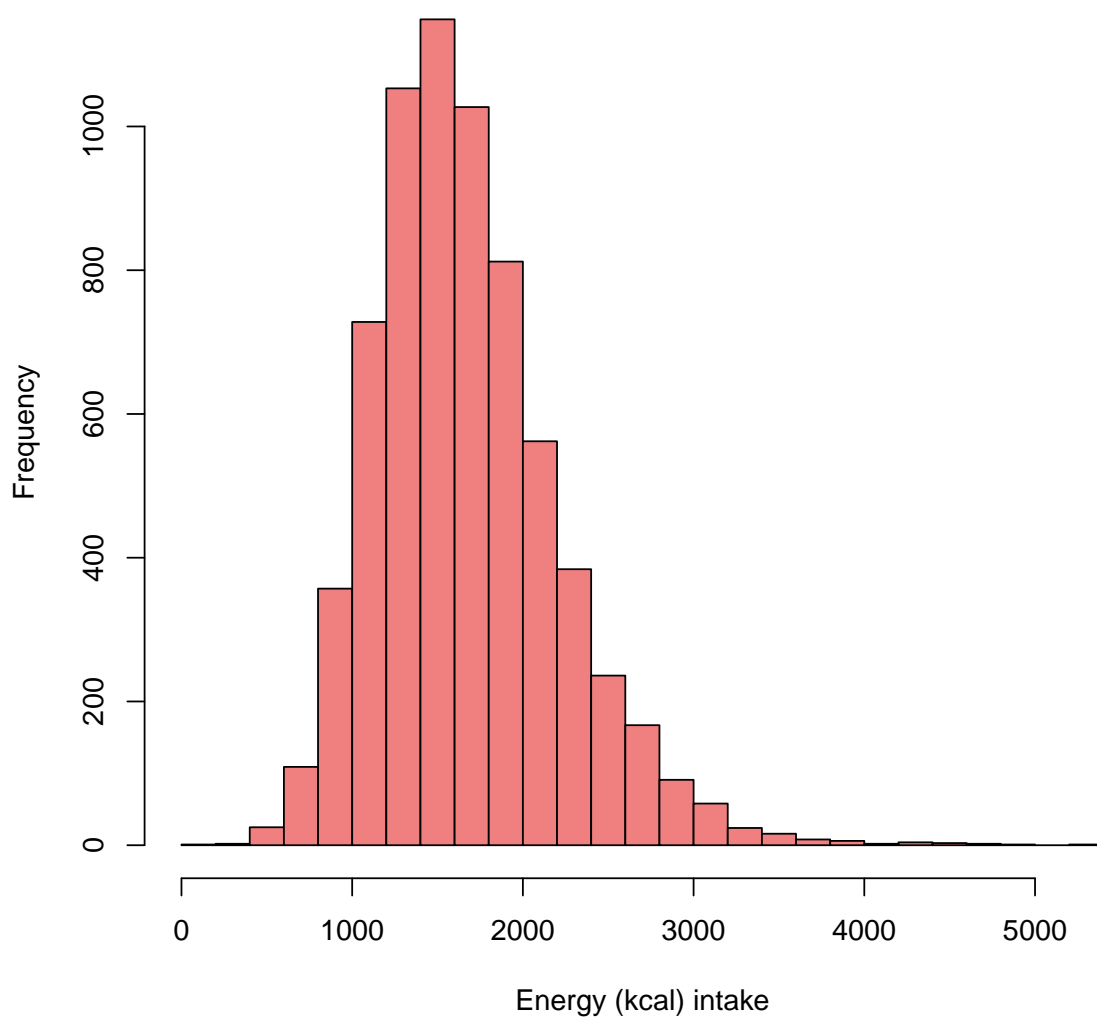
## 1.6 Semi-continuous data

The consumption of foods and nutrients can be categorised as being either habitually or episodically consumed. Habitual consumption can be defined as foods or nutrients being consumed by the majority of people on the majority of days and episodic consumption would indicate the converse intake pattern, intakes on a minority of days by a minority of people. As nutrients are generally distributed throughout foods, nutrients are usually consumed habitually whereas foods are usually consumed episodically as individuals tend to vary their diet from one day to the next. Episodically consumed foods have a semi-continuous distribution that has a large number of observations at zero to indicate non-consumption and the remainder following a continuous distribution (Olsen and Schafer, 2001). **Figures 4,5 and 6** show energy, omega-3 and alcohol intakes as an example of frequency distributions, with all participants consuming energy on all days, some participants consuming omega-3 on some days, and few participants consuming alcohol on a few days.

The classification into habitually consumed foods is not clear-cut and may be done according to background consumption knowledge with one study classifying total vegetable consumption as an episodically consumed food group although only 3% of participants did not consume it (Carroll, 2014). Foods and nutrients may be episodically consumed for one of two possible reasons: either the participant does not consume the food and is therefore a never-consumer with zero probability of consumption or, the participant does consume the food and does have a probability of consumption greater than zero, however too few days of intake have been collected to capture their consumption. With extra information, such as an FFQ, participants can be correctly assigned to the never- or non-consumer category and modelled appropriately (Haubrock et al., 2011).

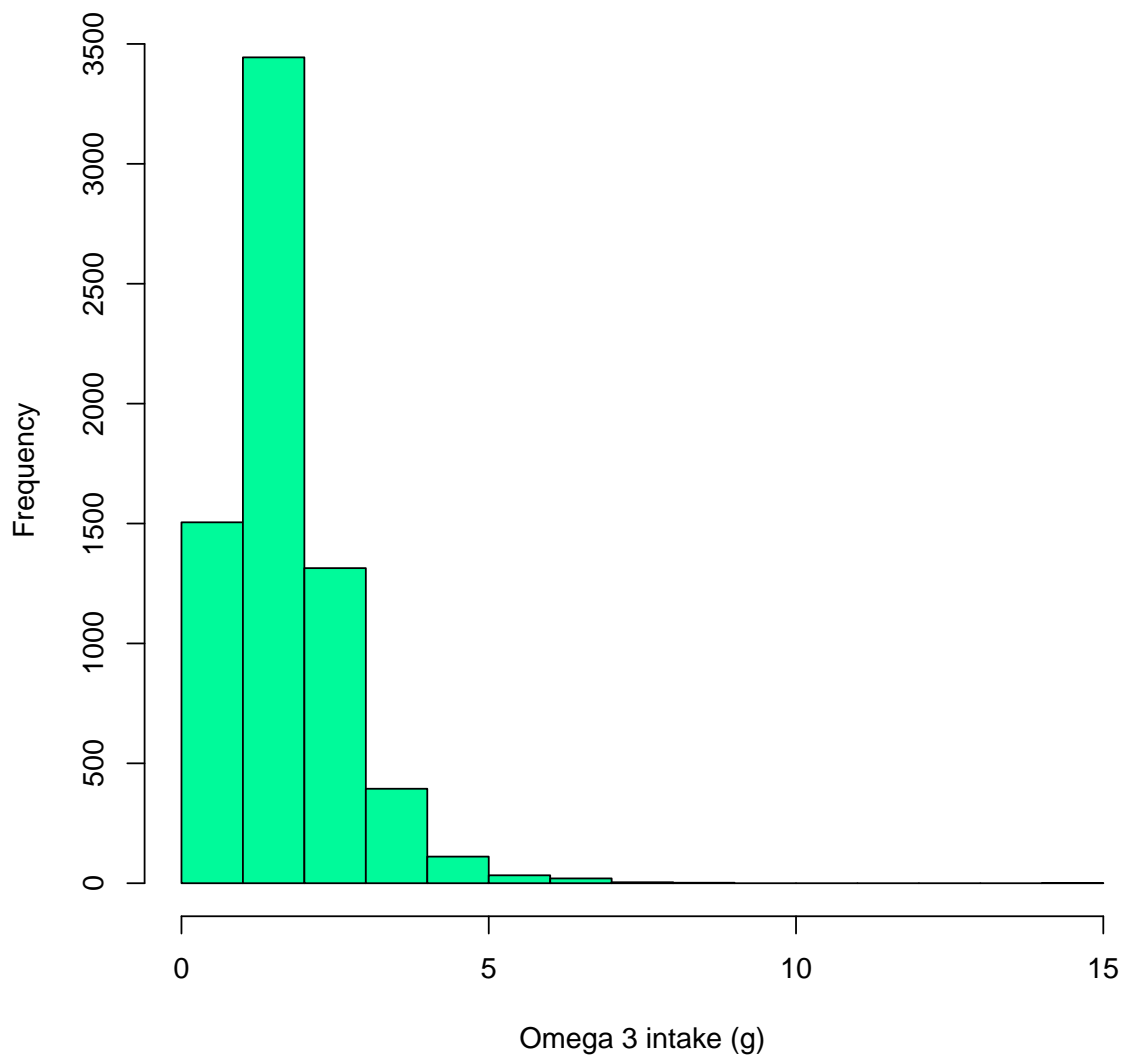
**Figure 4**

Distribution of intake for energy (kcal) for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012).



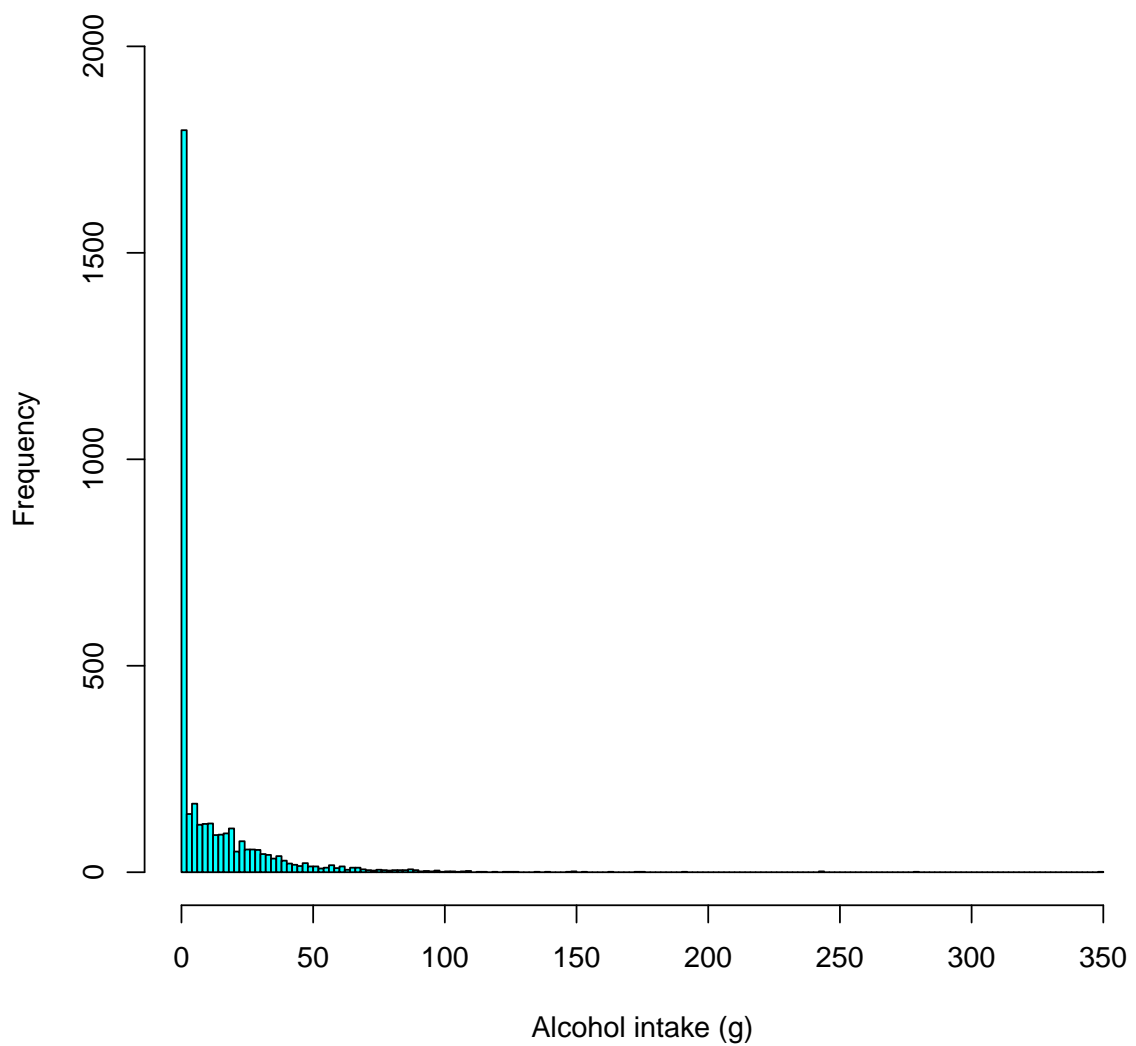
**Figure 5**

Distribution of intake for omega-3 fatty acid (g) for all 6828 participants aged 1.5+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012).



**Figure 6**

Distribution of intakes for alcohol (g) for all 3603 participants aged 18+y from the National Diet and Nutrition Survey Rolling Programme (NDNS RP) Years 1-4 (2008-2012).



## 1.7 Survey designs

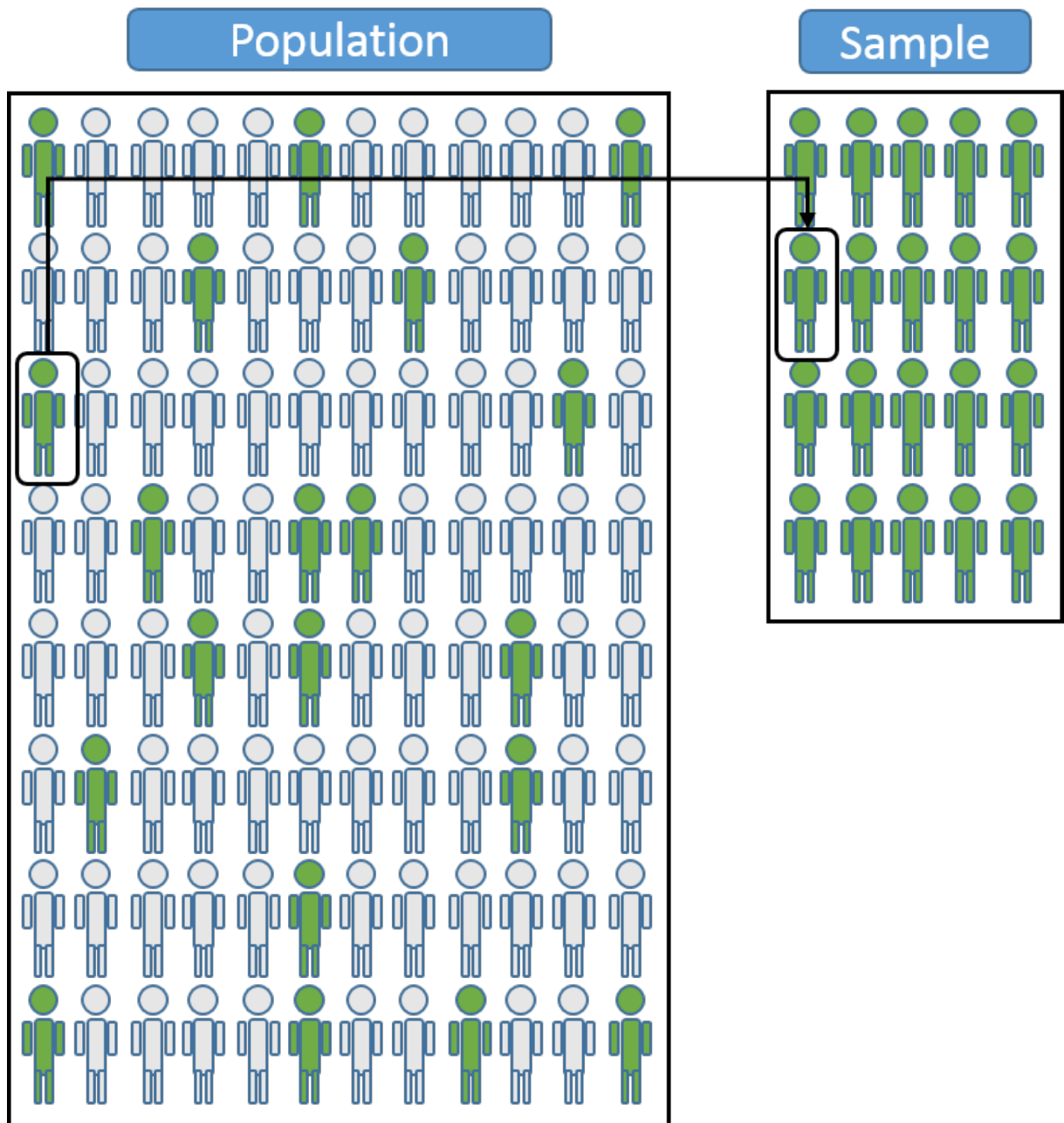
### 1.7.1 Simple random sampling

When attempting to examine the diets of a population, the ideal scenario would be to take the entire population, survey them and then examine the results. This is usually impractical, however, and so a sample of participants are chosen. It is important that the selected sample is still representative of the population, otherwise any inference may not be applicable back to the population, and so to ensure that the sample reflects the population individuals are selected at random to mitigate any bias that may occur in selecting certain groups. For example, to determine the average height of a population it is important to sample equal numbers of males and females as a greater number of males would suggest a taller than expected population, as males are, on average, taller than females. This is referred to as simple random sampling (SRS) and is illustrated in **Figure 7**, where the members of the population (left) are selected at random (indicated in green) to become part of the sample (right) (Dodd, 2011).

Carrying out simple random sampling in national surveys taking place over a large geographic area can become expensive due to the travel and time costs. A hypothetical example of the distances travelled when sampling in the UK are shown in **Figure 8**, where the grey dots indicate a random selection of participants dispersed throughout the UK. Visiting these participants would require travelling large distances and if repeated visits over a number of days are required, the time and cost of travelling by the interviewer would become prohibitive. Due to these reasons a complex sample design is often preferred. A further draw back to simple random sampling is due to the precision of estimates for population subgroups as these may be reduced if there are not enough subgroup members sampled to be representative.

**Figure 7**

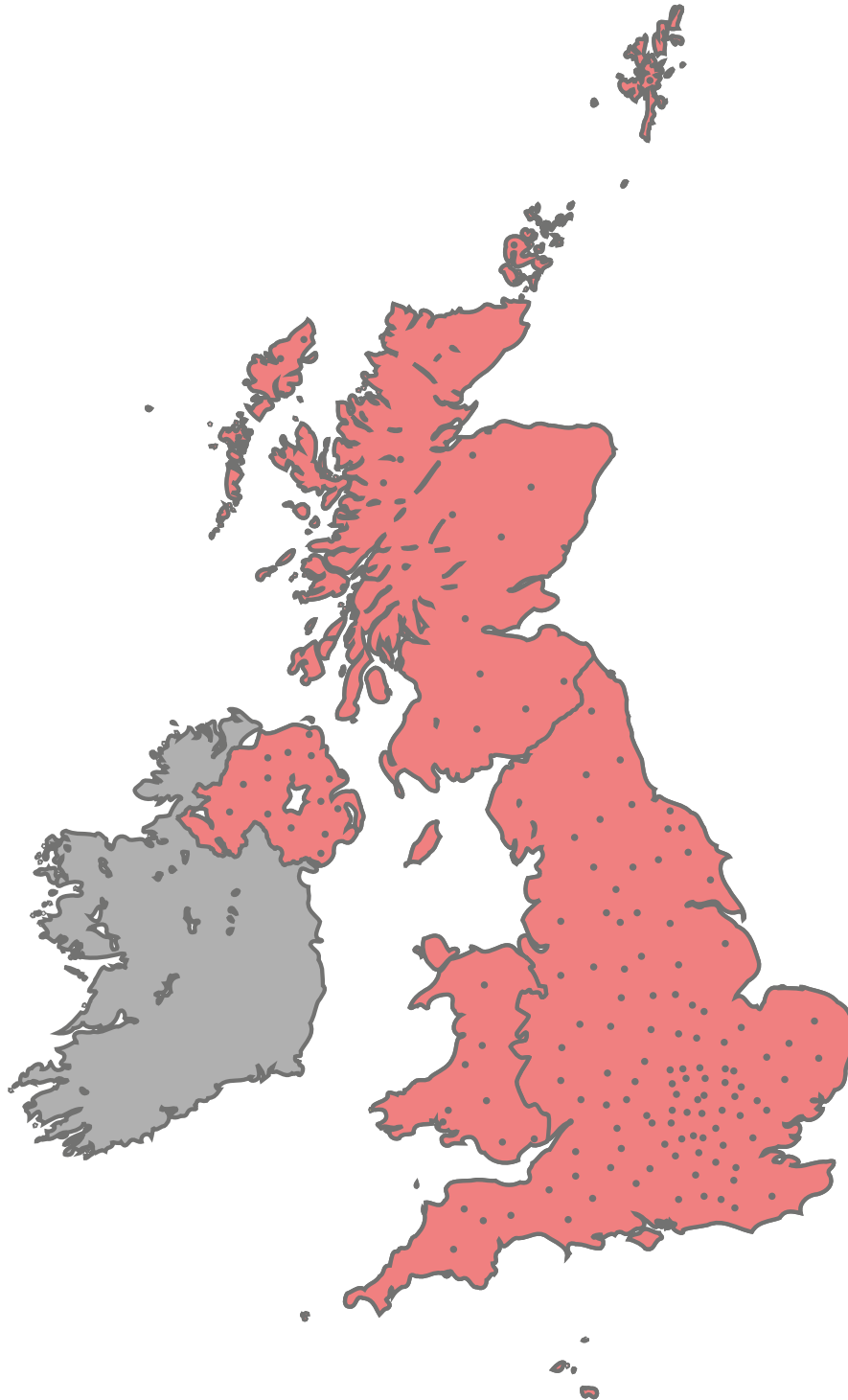
Simple random sampling representation adapted from Dodd (2011)





**Figure 8**

Illustration of simple random sampling of the UK



### **1.7.2 Stratified simple random sampling**

More efficient approaches to survey sampling include first dividing the population into groups, or strata, and then performing simple random sampling within each strata level.

This is illustrated in **Figure 9**, here the population is divided into four strata and selected individuals (coloured green, blue, orange and yellow to indicate different strata) become part of the sample.

The choice of strata is usually based on population demographic characteristics or factors thought to be related to the measurement of interest e.g. dietary intake. Examples of stratification factors include age, income and, in the case of UK national surveys: English regions and the devolved countries; Northern Ireland, Scotland and Wales. The selection of individuals with similar characteristics allows the comparison of like with like within each stratum which leads to more precise estimates of the population parameters.

### **1.7.3 Stratified balanced simple random sampling**

Sampling with stratification has the advantage that there is a greater probability of including individuals in minority demographic groups in the sample, however this can end up with strata sizes that are potentially too small to calculate the variance of estimates where a single individual has been selected. An alternative approach is to ensure that each strata has the same sample size, this is known as stratified balanced simple random sampling (SBSRS). This approach is preferred when having similar levels of precision between subgroups is more important than having a sample representative of the population. SBSRS is illustrated in **Figure 10** where five members of each strata are sampled per population strata level.

### **1.7.4 Stratified clustered simple random sampling**

Whilst stratification is effective at ensuring minority groups within the population are included in the sample, it does not address the logistical issues of carrying out surveys over large areas, this challenge can be met through clustering. This involves sampling clusters of individuals within a small geographic area leading to a much reduced journey time for the interviewer. In practice this often means that the interviewer will have a list of houses that are all in the same street and be able to walk from door to door.

**Figure 9**

Stratified simple random sampling representation adapted from Dodd (2011)

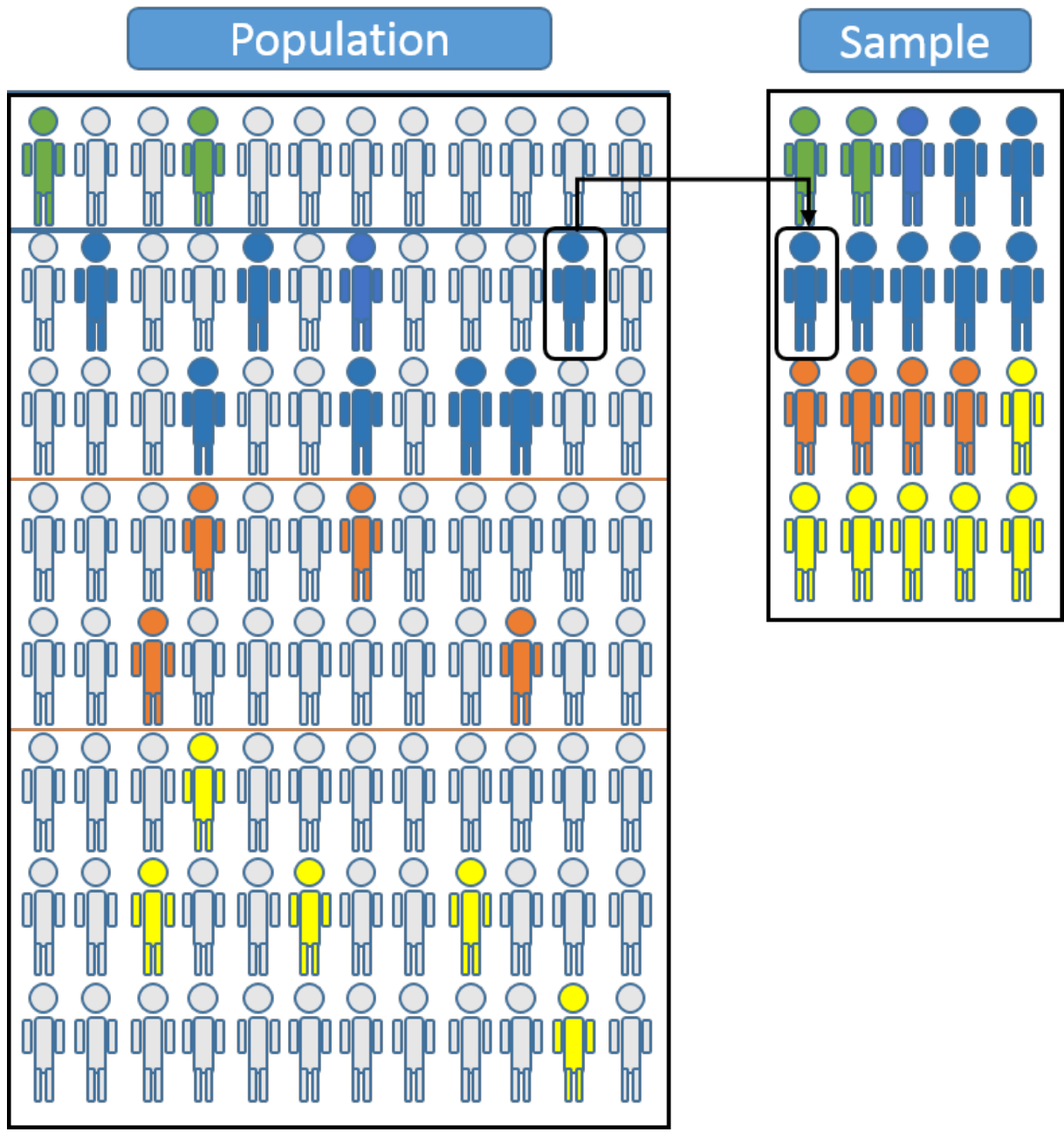
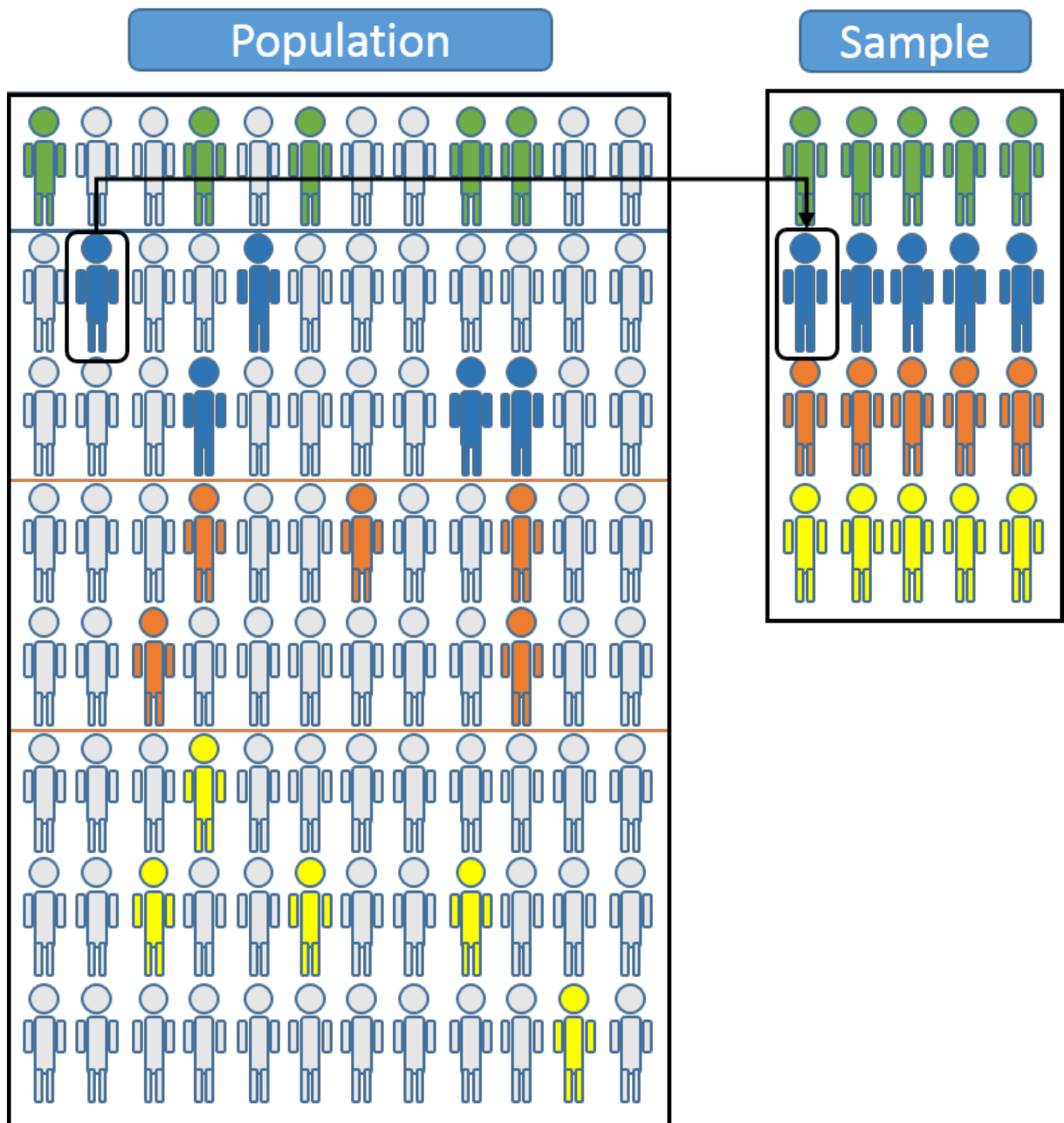


Figure 10

Stratified balanced simple random sampling representation adapted from Dodd (2011)



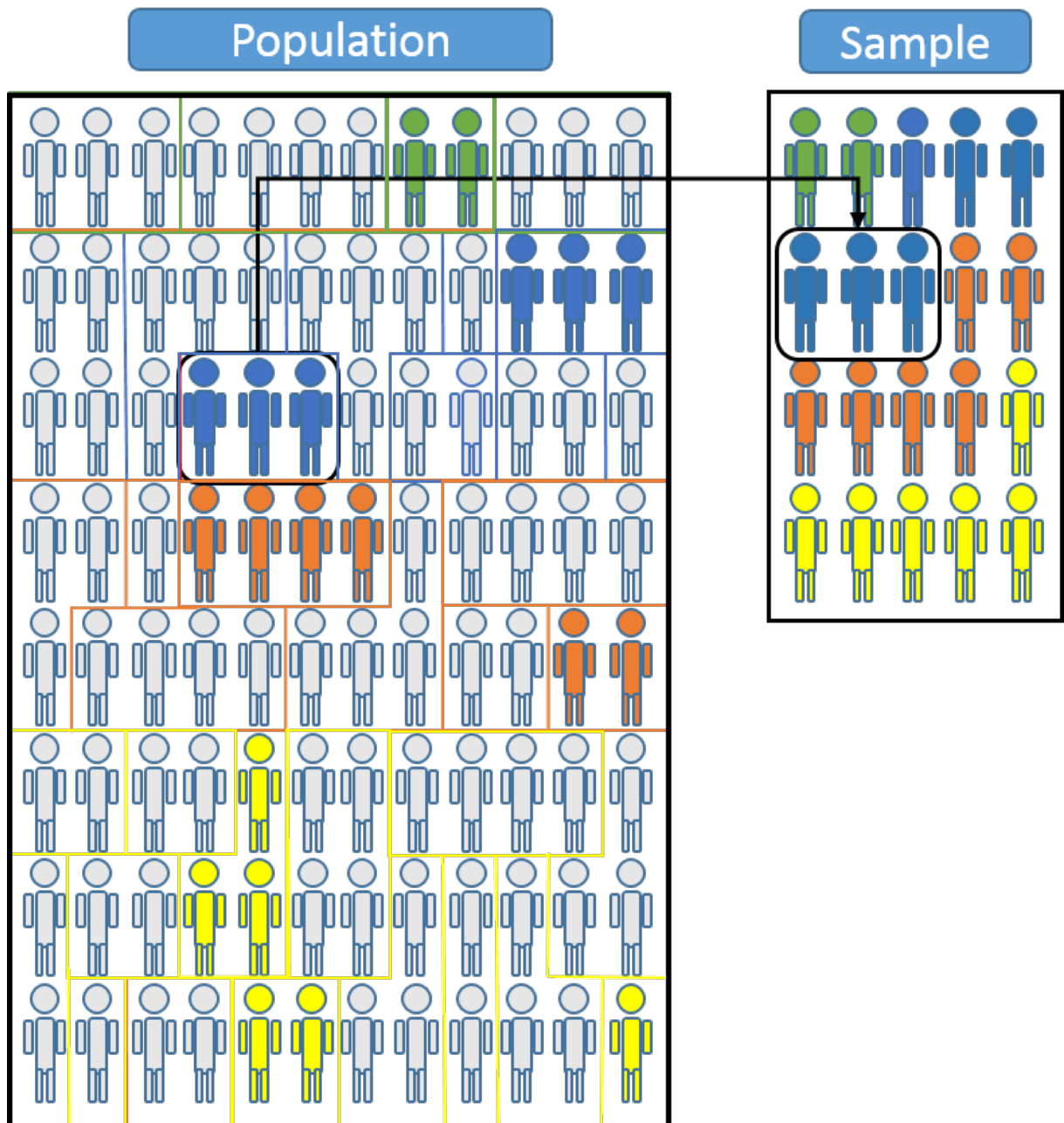
Clustering can have an impact upon the precision of the estimates as participants sampled in the same area are likely to be exposed to the same dietary determinants than participants from other areas. These may include access to fruit and vegetables, cost of foods and transport provision in accessing shops. An example of stratified clustered design is illustrated in **Figure 11**, where strata are created as in Figure 9 then within these strata clusters of individuals are selected to be in the sample.

### **1.7.5 Multistage sampling**

The selection of clusters can be extended to further levels, known as multistage sampling, such as is used in the National Health and Nutrition Examination Survey (NHANES) (**Figure 12**). The first stage selects Primary Sampling Units (PSU) which is generally at the county level, then the next level is to select segments within the Primary Sampling Unit (PSU) that are approximately the size of city blocks. The third stage selects households within the city block and then finally individuals are selected from within the household.

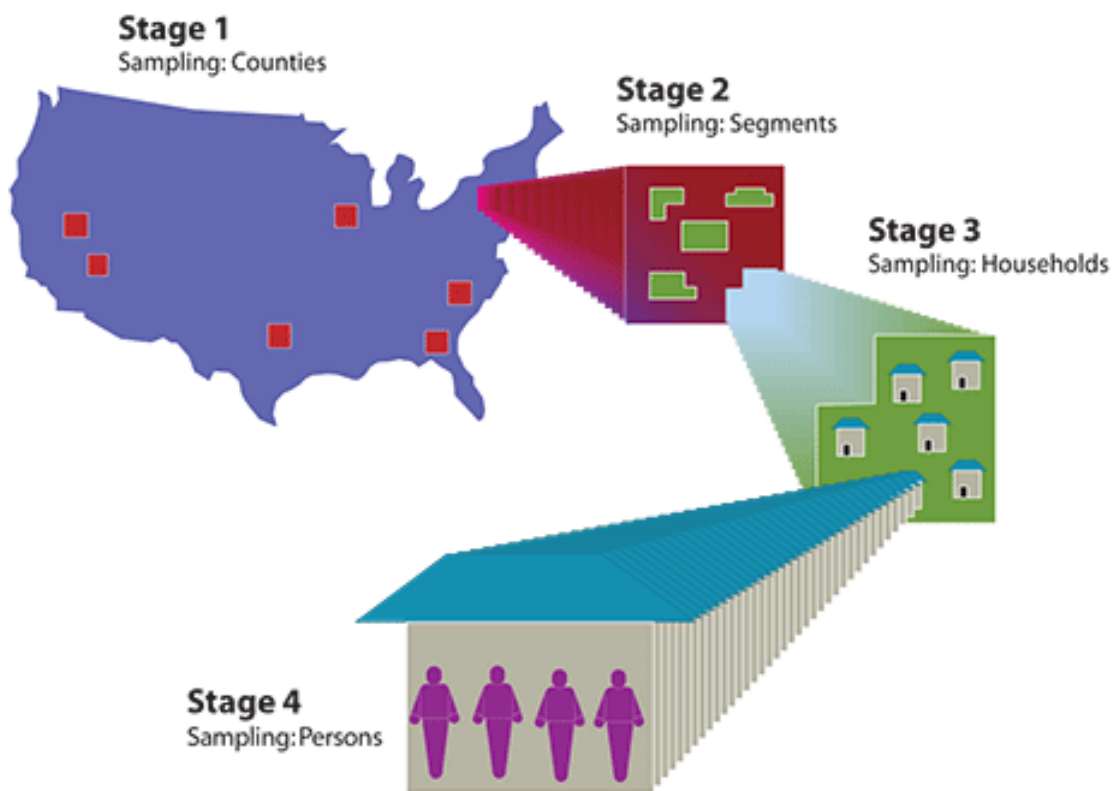
**Figure 11**

Stratified clustered simple random sampling representation adapted from Dodd (2011)



**Figure 12**

Illustration of multistage sampling using the National Health and Nutrition Examination Survey (NHANES) adapted from CDC and National Center for Health Statistics (2013)



### 1.7.6 Postal surveys

A different approach to overcome the logistic challenges of collecting participant information over a large geographic area is to post the questionnaire to respondents. A postal survey has the advantage that it does not require an interviewer to visit participants to recruit them into the survey, and, whereas the participants may not be at home when the interviewer calls, a posted survey will be seen when they return home. Postal surveys also have the advantage that participants feel less pressured to respond and so can complete the survey at a suitable time. The downside is that individuals are more likely to take part in the survey if asked in person and consequently response rates for postal surveys can be low. Postal surveys can be useful when recipients have an interest in the topic, and when the questionnaire is well designed high response rates have been observed (Kazzazi et al., 2018).

As mentioned above the variance of estimates are affected by clustering and stratification and as a result methods of analysis based on a SRS assumption may no longer be valid as the groups containing participants are no longer randomly selected, therefore statistical methods that can account for the complex sampling are required.



## 1.8 Statistical modelling of usual intake

The assessment of dietary intake in a country is typically undertaken through national dietary surveys. However, due to the daily variation seen in food intake, dietary data are prone to measurement error and therefore repeated measures are collected. The two sources of variability arising from between- and within-individual fluctuations need to be considered in the method of statistical analysis. Moreover, estimating the intake of episodically consumed foods and nutrients leads to modelling challenges. This is caused by dietary data containing records where the food or nutrient of interest is not consumed. When modelling all food intake there will be a number of zero observations together with a positive continuous distribution that is typically skewed to the right where it was consumed. In this situation, standard methods of analysis based on the normality assumption are not adequate, because a logarithmic transformation will not be sufficient to obtain a symmetric distribution. Similarly, restricting the analysis to non-zero observations would be suboptimal as knowledge of whether consumption took place would be ignored. Distributions of this type are referred to as semi-continuous distributions and can be analysed using a two-part model (Cragg, 1971; Manning, 1981; Duan et al., 1983) which deals with semi-continuous data in two parts. The first part indicates whether the food has been consumed  $P(Y > 0)$  and the second part models the amount eaten, given that it has been consumed  $Y|Y > 0$ . This allows instances of zero intake to be considered as genuine zeros instead of considering them as values below detection such as in the Tobit (Tobin, 1958) and Heckman selection models (Heckman, 1976, 1977). A model that combines both a two-part and a tobit approach has been suggested (Moulton and Halsey, 1995), which would be applicable where data contain true zeros in addition to values that are below the limit of detection. Here the interest is in continuous variables though for count data alternate methods have been proposed e.g. zero-inflated Negative Binomial and Poisson models. Similar to the two-part models, the zero-inflated models partition the population into consumers and non-consumers using a two component mixture that contains a degenerate distribution centred at 0 for non consumers and a distribution such as a Poisson or Negative Binomial is used for the count of consumption. The proportion of zeros from the non consumers is the mixing probability of the two component mixture

distribution (Tang et al., 2015). The advantage of the zero-inflated Negative Binomial distribution is that it contains a parameter to model large variances whereas the zero-inflated Poisson regression model does not and can therefore lead to biased estimates (Tang et al., 2015).

## 1.9 One-part model

The one-part model forms the basis of the measurement error model in its simplest case. It models the intake  $y_{ij}$  of foods or nutrients that are regularly consumed by most people on most days, examples include energy, iron and water. The one-part model is specified as follows:

$$y_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + u_i + \epsilon_{ij} \quad (5)$$

where  $\mathbf{X}'_{ij}$  is a  $1 \times q$  vector of covariates for individual  $i$ ;  $i = 1, \dots, n$  collected at observation  $j$ ;  $j = 1, \dots, n_i$ ,  $\boldsymbol{\beta}$  is a  $q \times 1$  vector of regression coefficients that includes an intercept and regression coefficients of  $q - 1$  ( $q = 1, \dots, n$ ) covariates,  $u_i$  is a random intercept to account for the correlation between repeated observations of intake and  $\epsilon_{ij}$  is an error term. This model has the following assumptions:

$$\epsilon_{ij} \sim N(0, \sigma)$$

$$u_i \sim N(0, \sigma_u)$$

$$\epsilon_{ij} \perp u_i$$

that is the error terms  $\epsilon_{ij}$  and  $u_i$  are independent and are normally distributed with variances  $\sigma$  and  $\sigma_u$  respectively.

## 1.10 Two-part model

The two-part model (Olsen and Schafer, 2001; Toozé et al., 2002) is suitable for the analysis of intake of episodically consumed foods or nutrients. It accounts for the large number of zeros observed in the data by introducing a logistic regression model to determine the probability of consumption then a linear regression part which models

the amount consumed. Let  $Y_{ij}$  denote a semi-continuous response for subjects  $i = 1, \dots, n$  at day  $j = 1, \dots, n_i$ . This response can be represented by two variables: an indicator of consumption

$$Z_{ij} = \begin{cases} 1, & \text{if } Y_{ij} > 0 \\ 0, & \text{if } Y_{ij} = 0 \end{cases}$$

and the amount consumed  $Y_{ij}$ , given that it is greater than 0, which may be transformed, eg  $\log(Y_{ij})$ , to make it approximately normally distributed. For the binary part of the model we assume that  $Z_{ij}$  follows a logistic regression model with a random intercept to account for the correlation between repeated observations

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \mathbf{X}'_{ij}\boldsymbol{\beta} + u_i, \quad (6)$$

where  $\pi_{ij} = \Pr(Z_{ij} = 1|u_i)$ ,  $\mathbf{X}'_{ij}$  is a  $1 \times q$  vector of covariates,  $\boldsymbol{\beta}$  is a  $q \times 1$  vector of regression coefficients and  $u_i$  is a random effect. The contribution of participant  $i$  to the log likelihood from the logistic regression part given the random intercept  $u_i$  is

$$\ell_{Z_i} = \sum_{j=1}^{n_i} Z_{ij} \log(\pi_{ij}) + (1 - Z_{ij}) \log(1 - \pi_{ij}) \quad (7)$$

The amount consumed given that this was greater than 0, follows a linear mixed-effects model:

$$Y_{ij}|(v_i, Z_{ij} = 1) = \mathbf{X}^{*'}_{ij}\boldsymbol{\gamma} + v_i + \psi_{ij}$$

where  $\mathbf{X}^{*'}_{ij}$  is a  $1 \times p$  vector of explanatory variables,  $\boldsymbol{\gamma}$  is a  $p \times 1$  vector of regression coefficients, and  $v_i$  is a random intercept. The error term  $\epsilon_{ij}$  is assumed to be distributed as  $N(0, \sigma_\epsilon^2)$ . The contribution of participant  $i$  to the log likelihood from the linear regression part is:

$$\ell_{Y_i} = -n_i^* \log(\sigma_\epsilon) - \sum_{j=1}^{n_i^*} \frac{1}{2\sigma_\epsilon^2} (y_{ij} - \mathbf{x}^{*'}_{ij}\boldsymbol{\gamma} + v_i)^2 \quad (8)$$

where  $n_i^*$  is the number of positive intakes for individual  $i$ . Importantly the two random intercepts above are assumed to be jointly normal and possibly correlated:

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N\left(0, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}\right)$$

The likelihood function for the model defined by (6) and (7) can be expressed as:

$$L \propto \prod_{i=1}^N \int \int \exp\{\ell_{Z_i}\} \exp\{\ell_{Y_i}\} f(u_i, v_i; \boldsymbol{\Sigma}) du_i dv_i \quad (9)$$

where  $\ell_{Z_i}$  and  $\ell_{Y_i}$  are the logistic regression and linear regression log likelihood contributions and  $f(u_i, v_i; \Sigma)$  denotes the joint normal distribution for the random effects. The parameters of interest are the regression coefficients  $\beta$  and  $\gamma$ . It is natural to assume that the amount consumed is related to the probability of consuming. The parameter  $\sigma_{u,v}$  correlates the two parts of the model. Alternatively if  $\sigma_{u,v} = 0$  the two models can be estimated separately. This would imply that consuming or not consuming the food in one day does not influence the amount consumed. If this is not a plausible assumption, failure to take into account the correlation may lead to biased estimates of the parameters (Su et al., 2009). The choice of covariates for the logistic and the normal parts may coincide, although the linear part is based on those cases where the consumption is greater than zero.

## 1.11 Numerical integration

Marginal maximum likelihood estimation is commonly used to estimate mixed-effects models including those presented in this thesis. However maximum likelihood estimation of random effects is often intractable and therefore numerical integration is carried out. Numerical integration is used to estimate the log likelihood then numerical derivatives are used to maximise it. Approaches include Laplace approximation and ordinary and adaptive quadrature methods such as Gaussian quadrature. Laplace approximation (also referred to as Laplace's method) has been used to estimate integrals in the mixed-effects model setting, where it estimates the mode of the integrand, with respect to the random effects (Rizopoulos et al., 2009). The second approach, quadrature methods, approximates the entire distribution of the integral using a weighted sum of the predefined abscissas for the random effect. The level of precision can be related to the number of quadrature points chosen, though the increasing the number of quadrature points will lengthen the duration of the approximation Liu et al. (2010). The choice of weights and abscissas is dependent upon the shape of the integral being evaluated with standard Gauss-Hermite weights and abscissas (Golub and Welsch, 1969) used for estimation under a normality assumption and Gauss-Laguerre weights and abscis-

sas preferred when the random effects are assumed to have an asymmetric Laplace distribution (Geraci and Bottai, 2014).

## **1.12 Comparing methods of usual intake estimation**

There have been a number of different methods developed for estimating usual intake (Food and Nutrition Board et al., 1986; Slob, 1993; Wallace et al., 1994; Buck et al., 1995; Nusser et al., 1996a; Hoffmann et al., 2002; Slob, 2006; Tooze et al., 2006; Waijers et al., 2006), differing in statistical models, assumptions and data transformations used and there have been approaches to working with data in its original scale (Nusser et al., 1990). However none of the published methods satisfactorily meet all of the challenges required in modelling the NDNS RP data. The first method to attempt to estimate usual intake distributions was detailed in a report by the National Research Council (NRC) in the US (Food and Nutrition Board et al., 1986). This method was then modified by the US Institute of Medicine (IoM) (Subcommittee on Interpretation and Uses of Dietary Reference Intakes and the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes, 2003) by including a power or log transformation to transform the nutrient data to normality. The NRC and IoM methods have a number of limitations that render them unsuitable for application with the NDNS RP, as their capability at incorporating complex survey design is unclear: they do not allow for the inclusion of covariates and can only estimate habitual consumption. Attempting to address these limitations a number of methods have been suggested as summarised in **Table 3**, of these, three methods will be explored in further detail below, but firstly the traditional method currently employed in the NDNS RP and its limitations will be discussed.

### **1.12.1 Traditional approach**

The traditional method of estimating usual intake is known as the within-person mean method and this involves taking the average intake for an individual then averaging the intakes of all individuals within the group. Estimating usual intake for foods that are consumed episodically can be prohibitive both in terms of expense and participant

burden, particularly for a national survey of hundreds or thousands of individuals, as the number of days of measured intake required increases. Therefore the intake reported from a small number of days contains greater within-person variance than usual intake which gives biased estimates of the population distribution.

As an illustration, **Figure 13** displays the probability density of fruit and vegetable servings based on a single day's intake, the average of intake measured by two 24-hour recalls and intake estimated using the Iowa State University (ISU) (Nusser et al., 1996a,b) method (discussed below) using Continuing Survey of Food Intakes by Individuals (CSFII 1994-1996) data (Tippett and Cleveland, 2001). The proportion of individuals who eat less than 5 servings per day is approximately the same at 40% for all three distributions. However, examining the tails of each distribution shows that the proportion of individuals consuming less than one serving has greater variance with a much larger proportion of individuals estimated to be consuming less than one portion using one day of intake (lined area), falling with the average of two 24-hour recalls (hatched area) and falling further still with the ISU method (filled area) (Guenther et al., 2006).

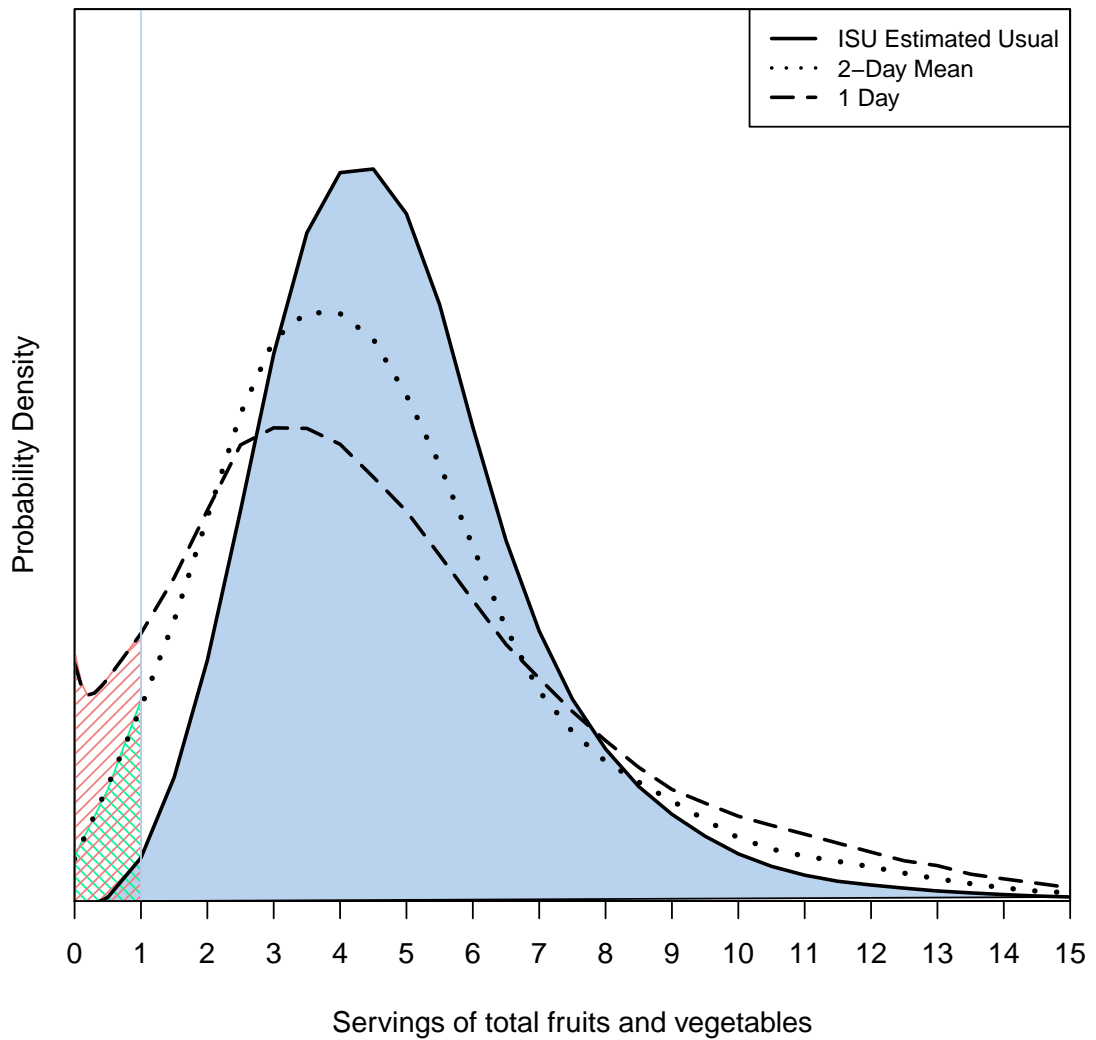
The traditional approach is currently used in the NDNS RP to estimate habitually consumed foods where four days of intake are collected per individual and the mean taken. The individual means are then averaged to give a single value per group defined by age and/or sex. This gives estimates that do not account for the variation in intake that occurs within the individual. These are then reported in the tables published annual reports. In the Year 1-4 NDNS RP report episodically consumed foods and nutrients were not analysed differently to habitually consumed foods and nutrients considered, with the exception of Alcohol. Here descriptive statistics including means and quantiles of alcohol intake were reported both for consumers and non-consumers combined and for consumers only broken down by age and sex groups (**Table 2**).

### **1.12.2 Iowa State University (ISU) method**

The ISU method was developed in 1996 for habitually (Nusser et al., 1996b) and episodically (Nusser et al., 1996a) consumed foods and nutrients. The ISU method was the first method developed to attempt to model both the frequency of consumption and

**Figure 13**

Intake of servings of fruit and vegetable, by one 24-hour recall (broken line), the mean of two non-consecutive 24-hour recalls (dotted line) and usual intake estimated by ISU model (solid line) for 14963 participants in the Continuing Survey of Food Intakes by Individuals (1994-1996) adapted from Guenther et al. (2006).



**Table 2**

An example of episodically consumed food reporting from Alcohol intake for 4947 participants aged 11+ in the National Diet and Nutrition Survey Years 1-4 (2008-2012). Adapted from Table 5.13 of the National Diet and Nutrition Survey Rolling Programme Report Steer et al. (2014)

Alcohol intake	Sex and age group (years)								
	Males			Females			Total		
	11-18	19-64	65+	11-18	19-64	65+	11-18	19-64	65+
Total (including non-consumers)									
Alcohol (g)									
Mean	2.5	18.5	12.9	1.8	10.1	4.9	2.2	14.3	8.4
Median	0.0	8.5	4.8	0.0	1.7	0.0	0.0	4.7	0.1
sd	11.0	28.3	18.0	8.5	15.5	8.4	9.9	23.2	14.0
Upper 2.5 percentile	35.0	87.6	61.3	22.3	56.1	26.2	27.6	70.4	54.0
Lower 2.5 percentile	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total energy (%)									
Mean	0.8	5.6	4.5	0.7	4.1	2.3	0.7	4.9	3.3
Median	0.0	2.9	1.6	0.0	0.7	0.0	0.0	1.8	0.1
sd	3.3	7.5	6.0	3.1	5.9	4.1	3.2	6.8	5.1
Upper 2.5 percentile	12.3	23.7	21.9	8.6	19.1	13.4	10.7	22.2	18.7
Lower 2.5 percentile	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Bases (unweighted)									
	744	1126	317	753	1571	436	1497	2697	753
Consumers only									
Alcohol (g)									
Mean	19.3	29.2	21.5	14.5	19.2	11.1	17.0	24.7	16.5
Median	10.1	20.8	16.0	7.5	14.8	9.1	8.0	18.1	12.3
sd	24.6	30.8	18.8	19.9	16.8	9.4	22.5	25.9	15.9
Upper 2.5 percentile	98.5	100.6	76.0	89.9	67.4	33.6	89.9	86.0	61.3
Lower 2.5 percentile	0.1	2.4	0.5	0.3	0.9	0.1	0.1	1.0	0.1
Total energy (%)									
Mean	5.9	8.9	7.5	5.6	7.8	5.3	5.8	8.4	6.4
Median	3.3	7.0	5.6	3.1	6.4	3.8	3.3	6.7	4.7
sd	7.2	7.7	6.2	7.0	6.2	4.7	7.0	7.1	5.6
Upper 2.5 percentile	26.7	29.0	23.3	31.9	22.9	14.4	27.3	26.6	21.9
Lower 2.5 percentile	0.0	0.5	0.2	0.1	0.3	0.1	0.0	0.5	0.1
Per cent consumers									
	13	63	60	13	53	44	13	58	51
Bases (unweighted)									
	91	697	174	87	788	177	178	1485	351



the amount consumed with a step to integrate out the random effects to estimate usual intake distribution. This method cannot cope with covariates, however it can make adjustments for nuisance effects such as seasonal variations, week day versus weekend and start day. The effects should firstly be examined by linear regression, to determine whether they do contain significant variance and should therefore be adjusted for. The ISU method typically has greater uncertainty, as indicated by the standard deviation of the bias, than the NCI and SPADE methods described below (Souverein et al., 2011).

### **1.12.3 National Cancer Institute (NCI) method**

The National Cancer Institute (NCI) method (Tooze et al., 2002, 2006, 2010) was developed in 2006 and is currently used to estimate usual intake in multiple surveys including the NHANES report as well as secondary analysis by a number of authors that include usual intake estimation of added sugar in adolescents (Zhang et al., 2015), fish and omega-3 intake in adults (Papanikolaou et al., 2014) and intakes of female breast cancer survivors (Milliron et al., 2014), making it the predominate method used in the US. It has also been used to estimate usual intakes in other populations including the Bavarian food consumption survey (Wawro et al., 2017), the Brazilian national dietary survey (Sousa and Costa, 2015) and for beverage intake using Australian national nutrition and physical activity survey (Sui et al., 2016). The NCI method uses a two-part model where the first part models the probability of consumption using logistic regression then, in the second part, the data are transformed to normality or approximate normality using a Box-Cox transformation, conditional upon intake occurring. The data are then modelled on the transformed scale using a linear mixed-effects model that is capable of separating the within-person and between person variation. Then pseudo person intakes for each member of the sample are simulated using the estimated mean and the between person variance of the linear mixed-effects model. The final step involves back transforming the data to the original scale using the original Box-Cox transformation parameter and the within-person variation from the linear mixed-effects model in the second step. The Box-Cox transformation is effective at transforming to normality when the data show moderate levels of skewness. However, where large amounts of non-consumption is observed along with high degrees of

skewness in the positive consumption part of the distribution the box-cox transformation provides inconsistent predictions (Duan et al., 1983; Herrick et al., 2018). Furthermore it has been shown to provide biased estimates when within-person greatly exceeds the between person variance (Souverein et al., 2011). The NCI method can be used to estimate habitually consumed foods and nutrients using the second part of the two-part model only. The NCI method can be fitted using SAS macros that are available from the `riskfactor.cancer.gov` website and can include information on never consumers from a questionnaire such as an FFQ by including covariates indicating values of whether consumption took place.

#### **1.12.4 Statistical Program to Assess Dietary Exposure (SPADE)**

The Statistical Program to Assess Dietary Exposure (SPADE) software (Dekkers et al., 2014) is written in R, and was developed to be used in analysing the Dutch National Food Consumption Survey (DNFCS) data. The package offers three analysis options: habitually consumed foods and nutrients, episodically consumed foods and nutrients, and food or nutrient intake from food sources with supplements. Estimates are determined following similar steps to those in the NCI method. A two-part model is used for foods that are consumed episodically, that calculates the probability of consumption first then models the amount consumed. When no long term measure of consumption is included, the method assumes that participants who have zero intake over the two 24HRs are consumers, but that two 24HRs were not sufficient to record an intake and that if the number of observations increased, consumption would eventually be observed. Data are transformed to normality using a Box-Cox transformation, then a linear mixed-effects regression model estimates the mean and the within- and between-person variances. Data are then back transformed using Gaussian quadrature using the within-person variance from the mixed-effects model along with the initial Box-Cox transformation parameter. The method does not correlate the probability of consumption with the amount consumed. It is generally assumed that when an individual consumes a food they will consume larger amounts of it for a number of reasons, either as they have easy access or because they like the food. However, the difference in results from SPADE to other methods that do include this correlation have shown

that the results from SPADE are comparable to the ISU and NCI methods for single nutrients (Souverein et al., 2011). A three-part model includes dietary supplements in the model alongside foods. As supplements are typically consumed in standard portions every day and thus with little variation, the assumptions that are included in the method are less speculative. Covariates cannot be modelled using SPADE other than sex, although the method is implemented as a function of age. SPADE is based on one positive intake, if the intake value for an individual is zero throughout their records then they are removed from the dataset. If an individual has more than one positive intake then one of their intakes is randomly selected (Dekkers et al., 2014). **Table 3** provides a summary of methods used to estimate usual intake.

**Table 3**

Comparison of methods for estimating usual intake

	<b>NRC/IOM</b>	<b>ISU</b>	<b>NCI</b>	<b>SPADE</b>
Software	SAS	Stand-alone software (SIDE)	SAS Macros	R
Cost	Requires SAS	Suggested donation (\$300-\$500)	Requires SAS	Nil
Covariates	No	<i>A priori</i> adjustments	Yes	Age only
Episodical Foods	No	Yes	Yes	Yes
Multistage sample variance	No	Yes	BRR	Bootstrapping
Sample weighting	No	No	Yes	Yes
Quantiles	No	Age	Yes	Yes
Frequency and amount correlation	No	No	Yes	No

Adapted from Souverein et al. (2011).

NRC / IOM - National Research Council / Institute of Medicine

ISU - Iowa State University

NCI - National Cancer Institute

SPADE - Statistical Program to Assess Dietary Expose

The current method for estimating intake for the NDNS RP, the within-person mean method, has the advantage that it is easy to implement however a mean is not an adequate summary statistic for skewed distributions as the bias arising from not separating the within-person and between person variation and the high frequency of zero observations when estimating mean intake of episodically consumed foods are not taken into account. Alternative approaches such as the two methods of usual intake estimation used in the Dutch and US national dietary surveys show similar estimates as determined by simulation studies (Souverein et al., 2011; Laureano et al., 2016) but may fail to cope with data that is strongly skewed as the methods are restricted to the use of the Box-Cox transformation. Furthermore, in the case of SPADE when measuring episodically consumed foods fail to consider the correlation between the probability of consumption and the amount consumed.

### **1.13 Iron Intake**

In this thesis I shall use iron intake and expenditure on iron medication to illustrate the statistical methods developed, due to its public health significance both in the UK and across the world. Furthermore, iron is a useful nutrient to examine because the intake distributions observed for iron as a nutrient and for foods that contain iron provide a good illustration of the challenging distributions arising in food consumption. Iron is an habitually consumed nutrient: consumed by the majority of individuals on the majority of days. It has an important role in the body as an oxygen transporter in haemoglobin and myoglobin, in many enzymic reactions and is important as a transporter for electrons within cells (Scientific Advisory Committee on Nutrition (SACN), 1991). So estimating iron intake with reduced bias is important in distinguishing clinical iron deficiency from statistical artefact.

### **1.14 Summary**

In this chapter I have introduced the statistical issues present when estimating usual intake of foods and nutrients that have been collected as part of a survey employing a

complex survey design. These include the challenge that diets are not consistent within individuals from one day to the next (Figures 1 and 2) and that they vary by season (Figure 3) and are prone to measurement error, which was introduced in Section 1.3. The methods used to collect dietary data were introduced in Section 1.5 explaining why it is difficult to use dietary assessment to capture true usual intake due to the number of days required to capture true intake, and that the burden this places on participants means that it is unlikely that a representative sample from the population would be collected. The challenge of skewed data seen in episodically consumed foods and nutrients was introduced in Section 1.6 and illustrated in Figures 4, 5 and 6. The challenges of selecting participants to take part in a national survey, ensuring that the sample is representative of the population but that has sufficient members from minority subgroups for analysis to be possible along with the logistical constraints of collecting data by visiting disparate addresses were presented in Section 1.7. Then existing methods of estimating usual intake of dietary components were presented and contrasted in Section 1.12, where the traditional approach currently used in the NDNS RP was shown to lack the flexibility to cope with many of the challenges mentioned including: skewed distributions; separating within- and between-person variation; and correlating the probability of intake with the amount consumed. Methods used in other national dietary surveys that were more able to meet the challenges were described in Section 4.1 but have been shown to struggle to produce reliable estimates where data are considerably skewed and, in the case of SPADE, the correlation between intake probability and intake amount is assumed to be zero.

## **1.15 Aims and Objectives**

This thesis aims to meet the statistical challenges described in this chapter, by developing statistical approaches to the estimation of dietary intake collected as part of a national survey whilst accounting for a complex survey design and measurement error that also address the limitations of the NCI and SPADE methods. Specifically the objectives of the thesis are:

1. To develop a novel method for estimating the mean intake of episodically consumed foods and nutrients, collected using a complex survey design. The method should allow the specification of the mean in terms of explanatory variables and account for the within- and between-person variation.

This will be achieved through developing a novel two-part model for semi-continuous data with random effects, that has improved flexibility in modelling skewed distributions of non-negative data by using a generalised gamma distribution to model intake of episodically consumed foods. This ensures a good model fit to the data and removes the requirement for a Box-Cox transformation. Furthermore the two parts of the proposed model have a joint correlation structure allowing for each individual's probability of consumption to be correlated with their intake amount. The estimation procedure will be extended to incorporate the complex sample design. This is presented in **Chapter 3**.

2. To develop methods for estimating quantiles of intake from habitually consumed foods and nutrients using data collected under a complex survey design. The method should allow the specification of quantiles in terms of explanatory variables and account for within- and between-person variation.

This will be achieved through the development of a semi-parametric approach to quantile regression that is based on the asymmetric Laplace distribution which can deal with skewed distributions and non-negative observations. The method is extended to account for the complex survey design in the estimation of model parameters (**Chapter 4**). This allows improved flexibility in estimating specified quantiles of intake for example at the Lower Reference Nutrient Intake (Section 1.4)

3. To model expenditure on iron prescriptions in the UK across health boards using the quantile regression model introduced in Chapter 4, whilst incorporating estimated dietary intake as a covariate in the model. This information is important for a fair comparison of expenditure among health boards, as expenditure will depend on iron bioavailability status.

This will be achieved by analysing national electronic records of expenditure on iron prescriptions by health boards in the UK. The quantile regression model will account for

clustering at the health board level. The covariates will include estimated bioavailable iron intake based on age and sex make up of registered patients. Estimated quintiles of expenditure by region will be graphically presented in **Chapter 5**.



The remainder of the thesis is structured as follows: **Chapter 2** introduces the National Diet and Nutrition Survey Rolling Programme detailing the purpose of the survey, how the data are collected including the survey design and weighting and describing the explanatory variables used throughout the thesis. Chapter 3 presents methods for analysing the mean iron content of selected foods which demonstrate a semi-continuous distribution, using a two-part model with a generalised gamma distribution, and extends the estimation procedure to incorporate the complex survey design. Chapter 4 describes methods for the modelling of quantiles of iron intake using linear mixed-effects quantile regression and extends the model estimation procedure to incorporate the complex sample design. Chapter 5 utilizes the quantile regression model introduced in Chapter 4 and methods of estimation of dietary intake to explore differences in the amount spent on iron prescription between UK health boards and **Chapter 6** is a discussion. Included as appendices is an example of the letter inviting participants to take part in the NDNS RP (**Appendix A**), a table listing the measures collected by the NDNS RP (**Appendix B**) and an example of the food diaries used to collect dietary data in the NDNS RP (**Appendix C**). **Appendix D** contains supplementary tables showing the impact of various numbers of bootstrap resamples upon standard error estimation for the methods presented in Chapter 3. **Appendices F, H** and **L** provide examples of the structure of the data used in Chapters 3, 4 and 5. R and SAS scripts that include the code used in Chapters 3, 4 and 5 are presented in **Appendices G, I** and **J**. The sources of data used in Chapter 5 are illustrated in **Appendix K** along with the regression coefficients table for the analysis performed in Chapter 5 in **Appendix M**.

## 2 National Diet and Nutrition Survey Rolling Programme

The NDNS RP (Public Health England, 2014) is a cross sectional national survey set up with the aim of capturing food and nutrient intakes in the UK. It is jointly funded by Public Health England (PHE) and the Food Standards Agency (FSA) and carried out by the Department of Health, Elsie Widdowson Laboratory (EWL) (formerly Human Nutrition Research (HNR)), the National Centre for Social Research (NatCen) and University College London (UCL). It is jointly funded by the Department of Health in England and the food standards agencies for Scotland, Wales and Northern Ireland.

National dietary surveys in the UK started with the Dietary and Nutritional Survey of British adults 1986-87 (Gregory et al., 1990), subsequently the NDNS programme began with separate surveys carried out for different age groups starting with children aged 1.5-4.5y in 1992-3 (Gregory et al., 1995); People aged 65+y (1994-5) (Finch, 1998); Young People aged 4-18y (1997) (Smithers et al., 2000) and a second survey for adults aged 19-64y with fieldwork carried out in 2000-01 (Henderson et al., 2004). A further survey was carried out on low income families – the Low Income Diet and Nutrition Survey for people aged 4+y (2003-05) (Nelson et al., 2007), and in addition a national dietary survey was carried out in 2011 capturing the diets of infants aged 4-18 months (Stephen et al., 2013).

The NDNS RP moved to a rolling programme format in 2008 and has continued running in yearly cycles to date. Moving to a rolling programme was carried out as it was felt the survey would be better at capturing temporal changes allowing trends over time to be examined, it would be more responsive to policy needs and it would be able to collect additional data at short notice. Approximately 1000 non-institutionalised, non-pregnant non-breastfeeding participants aged 1.5 years and older are recruited each year, with some over-sampling of children and participants in the UK devolved countries. Participants are asked to record all food and drink in an estimated DD for four consecutive days, with those who complete a minimum of three days included. Along with food and nutrient information the NDNS RP collects blood and urine samples, sun exposure measures and asks many other questions that include physical activity,

medicine use and food security, see **Appendix B** for the complete list of collected variables.

## 2.1 The purpose of the NDNS RP

The NDNS RP is used as a surveillance tool by governmental agencies including the FSA and PHE to monitor the diet and nutritional status of the UK population for factors including intakes that deviate from reference nutrient intake and groups that are higher or lower than national averages. Furthermore it is used to develop policies and monitor their effectiveness, for example the Healthy Lives, Healthy People white paper (Department of Health, 2010) used the NDNS RP figures on salt intake and the number of adults consuming five portions of fruit and vegetables per day in its evidence base. Similarly the SACN used NDNS RP to highlight high levels of carbohydrate intake in their latest report on Carbohydrates and Health (SACN (Scientific Advisory Committee on Nutrition), 2015). Exposure to chemicals in food is also monitored using NDNS RP data by the FSA.

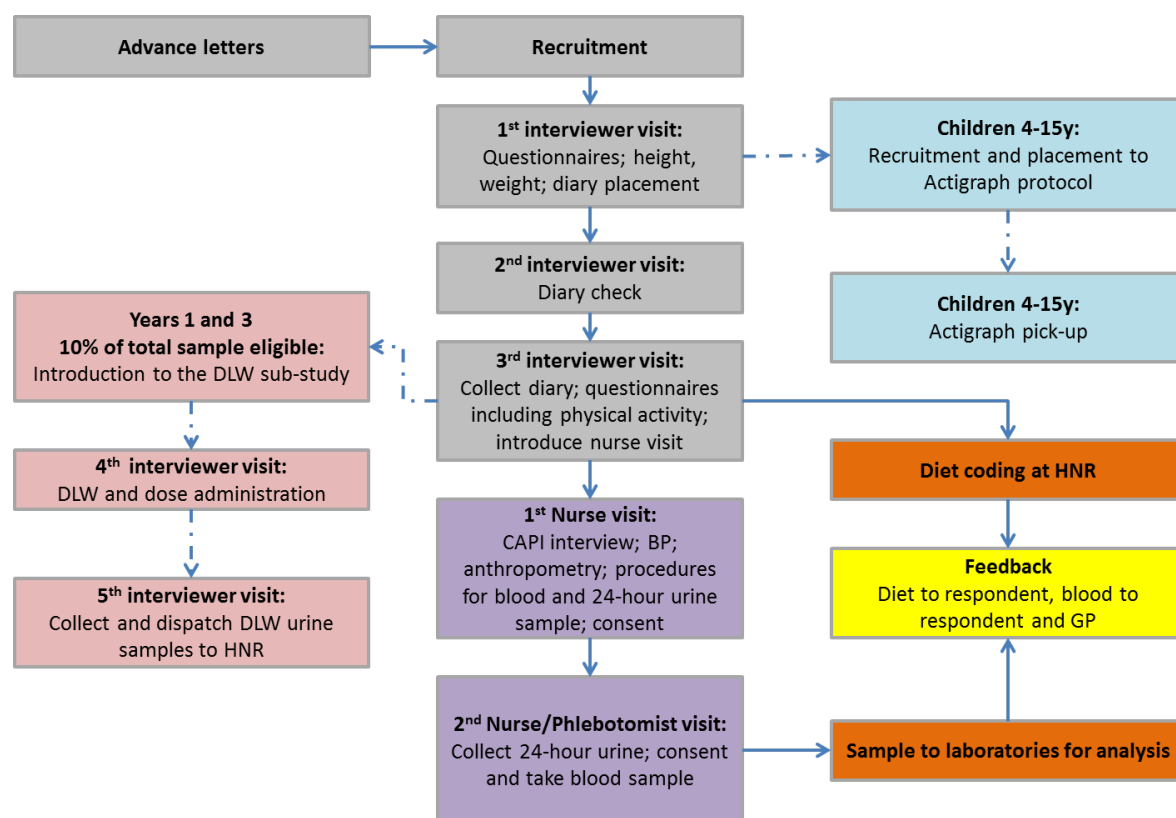
## 2.2 Data Collected

**Figure 14** presents the design of the study data collection. The first step is to send letters to selected addresses describing the survey and informing them that an interviewer will be visiting their house to invite household members to take part (see **Appendix A**). The interviewer then arrives at the address to recruit the participants and ask them to complete a diet diary recording all food and drink consumed over 4 consecutive days, with a random start date given to ensure an equal distribution of days are covered. During the first visit anthropometric data, including height and weight, are collected along with the Computer Assisted Personal Interview (CAPI) questionnaire. Furthermore if the participant is aged 4-15y and is willing to participate they are recruited into the physical activity arm and given an ActiGraph monitor to record their physical activity levels, this requires a further visit to collect the ActiGraph monitor.

Also at the first visit, participants are given instructions on how to fill in the diary. The interviewer returns on the second day of the diary to check for completeness and answer any questions the participant may have regarding completion. After the four days have been completed, the interviewer returns for their 3rd visit to collect the diary, again checking the diary and collecting any food packaging to be used to help identify the nutrient content of any unusual foods. The diary pages are divided into categories prompting participants to record a thorough description of their diet including, the date and time of consumption, the brand of food and the estimated amount of food and drink consumed, along with some socio-contextual questions, such as who the food was eaten with and whether the television was on (see **Appendix C** for example diary pages).

**Figure 14**

National Diet and Nutrition Survey Rolling Programme Study Design



Once complete, the diaries are sent to MRC EWL for processing, where the data are manually coded and entered into the DINO database (Fitt et al., 2014) linked to the NDNS RP Nutrient Databank (Smithers, 1993) which contains food composition data, primarily based on McCance & Widdowson 6<sup>th</sup> edition (Food Standards Agency, 2002),

augmented with manufacturers' data and standardised portion sizes (Food Standards Agency, 1988).

Two further arms to the study are introduced on the third interviewer visit. A sub-sample of participants are recruited to take part in the DLW study to objectively measure energy intake (10% sub-sample in Years 1 and 3 only). Those that agree to take part receive two further visits by the interviewer; in the first of these the DLW is administered and at the second urine samples are collected and sent for analysis. Also on the third interviewer visit, participants are invited to receive a visit by a nurse who will ask health related questions and take a blood sample if the participant is willing. Blood samples are sent for analysis and passed on to the respondent's General Practitioner (GP). The analysis for blood samples includes a full blood count, lipid profile to assess cardiovascular disease risk and HbA1c to test for diabetes. The data used throughout this thesis are from the NDNS RP years 1-4 collected in 2008 to 2012.

### **2.3 Data accessibility**

Once processing, analysis and reporting have been completed, the data are uploaded to the UK data archive <http://www.data-archive.ac.uk>, hosted at the university of Essex. From here interested researchers are able to create an account and download the data for use. The files available include previous versions of the NDNS and the latest version of the Rolling Programme data available in multiple data formats. The files available include a product level list of every dietary record for every individual with nutrient and food information included and an aggregated version of this information containing the within-person means and therefore 1 record per individual. Also included are the questions and participant answers collected during the CAPI interview along with information relating to average daily nutrient intakes per participant, average daily food intakes per participant, and associated variables listing demographic, blood, sodium, physical activity and health measures. The UK NDNS RP Nutrient Databank containing the entire list of foods available in the DINO database is provided with a separate file for each year of the Rolling Programme. Further information on the types of variables available for analysis is given in Appendix B.

## 2.4 Sample design

The NDNS RP uses a complex survey design that includes stratification, clustering and sampling weightings. The weightings provide an adjustment for selection probability at the address, household and individual level and also adjust for non-response. To select an individual a number of steps are carried out. Firstly a sample of 799 postcode sectors were drawn from the postcode address file, referred to as strata and from within these sectors, 21,573 addresses were sampled known as clusters. Dwelling units within each address are chosen, and then catering units within the dwelling units are selected until finally individuals within the catering units are reached. Selection at the dwelling and catering units will only be relevant where these exist. Selection at the dwelling unit level would be required where the household space is not behind a door that only one household can use. This would apply where more than one household is sharing a bathroom or kitchen. A household is defined as either an individual or group of individuals who have the address as their main or only residence and either eat at least one meal together per day or share a living room. Catering units are defined as people living together who buy and eat meals together (Tipping, 2014).

The postcode sectors and addresses were randomly sampled from the postcode address file. This is a list of postcodes used for postal delivery within the UK limited to addresses that receive less than 25 mail items per day, to exclude business addresses, and as such is comprehensive in its coverage of the UK. This sample is the PSU. Where a PSU contains less than 500 addresses other PSUs are grouped together. The sampling frame is then split by country giving strata, then the PSUs in England are sorted by Government Office Region (GOR) and in all countries by the Index of Multiple Deprivation (IMD) and population density and from this list a systematic random sample of PSUs is chosen. Each PSU is randomly chosen by dividing the number of cases by the number required in the sample  $I = \frac{\text{number of cases}}{\text{number required in sample}}$ . A random number ( $R$ ) between 0 and  $I$  is generated and this number indicates the position of the first PSU to be chosen from the list. The next value is found by  $R + 1 \cdot I$  rounded up to the next integer, then subsequent values are  $R + 2 \cdot I, R + 3 \cdot I, \dots, R + N \cdot I$  and is continued until a complete sample frame is compiled. (Tipping, 2014).

## 2.5 Weighting

Participants in the NDNS RP are given a weight to adjust for their probability of selection into the survey and for non-response of selected individuals who did not take part. Separate sets of weights are given for the all 6828 participants, then further weights are created for participants in the various extra arms of the study who agreed to provide further health information (Nurse visit), provided blood samples, for individuals aged 16+y who completed the recent physical activity questionnaire, for participants who provided 24-hour urine samples and for children aged 4-15y who agreed to wear an ActiGraph activity monitor. These weights are important as they adjust the sample of participants to reflect the population they were sampled from, i.e. the UK, ensuring that the findings are reflective of the population as a whole. The weights are calculated in two stages first the weights are adjusted for the probability of selection. The initial sample was adjusted to take into consideration various factors that impacted upon the probability of selection, these were due to the pilot study, an increased sample in Northern Ireland, Scotland and Wales and an increase in the recruitment of adults. The pilot study, known as the Run In was carried out in February and March 2008 before the start of the survey proper to test methods and the data were subsequently included into the first year of the survey meaning that there are 14 months in Year 1 that required adjustment to ensure that no months were over represented. Due to the small population size in the devolved countries, relative to England, an increase in the number of residents was commissioned in Northern Ireland and Scotland in Year 1 and then Northern Ireland, Scotland and Wales in Years 2 to 4. This was to allow meaningful numbers of participants to be sampled allowing for country specific reports to be produced. Furthermore in the final quarter of Year 4 an additional number of addresses were sampled in Scotland. Referred to as country boosts, this oversampling has an impact upon the sample as the probability of being selected in England is lower than expected and therefore the contribution of participants from the devolved countries is weighted down, relative to participants in England. The final adjustment made was as a result of an increase in the recruitment of adults into the survey in Year 4. Prior to the increased recruitment addresses were designated as main: where both an adult and child could be surveyed and child addresses: where only a child was selected. In Year

4, adults could be interviewed at all addresses, meaning that, in Year 4, the probability of adults being sampled increased in comparison to previous years.

The probability of selection for each address was determined for each country by dividing the total number of selected addresses by the total number of addresses. Then, after adjustment for the Run In, and increased selection probabilities for adults, the weights for each country were combined denoted  $w_0$ . **Table 4** highlights the impact of weighting in correcting the sample to be representative of the population. For example, the percentage of English addresses in the UK postcode address file is 83% compared to Northern Ireland which has 2.8% of the addresses in the UK yet, 50.8% and 12.1% of the NDNS RP are from England and Northern Ireland respectively. Therefore weighting is important as it inflates the English sample and deflates the Northern Irish sample to reflect the probability of selection.

**Table 4**

The distribution of UK addresses in the UK Postcode address file, with unweighted NDNS RP Y1-4 (2008-2012) sample, then adjusted percentage following the application of selection weights, by country, adapted from the Table B.1 from the National Diet and Nutrition Survey Y1-4 (2008-2012) (Tipping, 2014)

	Selected sample of addresses		
	Postcode address file	Unweighted	Weighted by selection weight
	%	%	%
Government Office Region			
England	83.0	50.8	82.8
Wales	5.0	9.9	5.0
Scotland	9.2	27.0	9.2
Northern Ireland	2.8	12.1	2.8
Base (unweighted)	27,147,524	21,573	21,573



Further adjustment to the weights occurred where the address contained more than one dwelling unit ( $w_1$ ) or catering unit ( $w_2$ ) existed within an address to ensure that accommodation of this nature was not under represented in the sample.

Individuals within catering units were selected based on the type of address. Within each sample point were 27 addresses in Years 1-4, 9 addresses were designated as main addresses meaning that an adult and a child could be randomly selected to take part in the survey and the remaining 18 addresses within the sample point contained child addresses where children (aged <19y) were selected. Each individual is given a selection weight ( $w_3$ ) to reflect the household size otherwise individuals within smaller catering units will be over-represented in comparison to those in large catering units. The selection weight is the inverse of the individual selection probability, which for adults equates to the number of eligible adults within the catering units (i.e. excluding pregnant and breastfeeding women) and for children this is the number of eligible children in the household (i.e. aged 1.5-18y).

A final set of weights giving the probability of selection is given by

$$W_{sel} = w_0 * w_1 * w_2 * w_3$$

Where  $w_{sel}$  are the selection weights,  $w_0$  is for each country adjusted for Run In and differences in address type,  $w_1$  adjusts for multiple dwelling units,  $w_2$  adjusts for multiple catering units and  $w_3$  adjusts for selection probability at the catering unit level.

These selection weights are then used to create final weights for each member of the NDNS RP through calibration methods, whereby the selection weight is adjusted until it reflects age, sex and government office region groups in the UK. Further adjustment means that a single set of weights can be used to analyse the UK overall and for separate analysis by country, by children and adults and by males and females (Tipping, 2014).

## **2.6 Impact of the weighting and the complex survey design**

The impact that the weighting and survey design has upon point estimates is shown in **Table 5** where it can be seen that the majority of estimates that consider the weighting

and survey design aspects differ from those estimates that do not take the weighting and survey design into account, for example the adjusted upper 2.5 percentile for iron intake in Men aged 19-64y is 21.2mg per day whereas the unadjusted intake for 20.8mg per day.

**Table 5**

Iron intakes from food sources in the UK from 6224 participants aged 4 years and older from the National Diet and Nutrition Rolling Programme Years 1-4 (2008-2012), adjusted and unadjusted for the NDNS RP weighting and complex survey design

Average daily intake of iron from food sources only, by sex and age

		Sex and age group (years)							
		Boys		Men		Girls		Women	
		4-10	11-18	19-64	65+	4-10	11-18	19-64	65+
Adjusted for weighting and complex survey design									
	Mean	9.0	10.7	11.7	11.1	8.4	8.4	9.6	9.4
54	Median	8.8	10.5	11.5	10.8	8.2	8.1	9.5	9.1
	sd	2.5	3.4	4.0	3.7	2.4	2.7	3.0	2.7
	Upper 2.5 percentile	14.6	18.7	21.2	19.3	14.4	14.1	15.9	15.3
	Lower 2.5 percentile	4.8	4.8	5.3	5.2	4.4	3.6	4.1	4.9
Unadjusted for weighting and complex survey design									
	Mean	8.8	10.7	11.7	10.8	8.1	8.4	9.5	9.2
	Median	8.6	10.3	11.4	10.5	8	8.2	9.4	8.9
	sd	2.6	3.5	4.0	3.7	2.2	2.7	3.1	2.7
	Upper 2.5 percentile	14.9	18.6	20.8	19.4	12.7	14.2	16.0	14.9
	Lower 2.5 percentile	4.6	5.2	5.2	5.1	4.3	3.7	4.1	4.6
	Participants (unweighted)	665	744	1126	317	612	753	1571	436

## 2.7 Socio-economic factors: NSSEC

Diets have been demonstrated to vary according to income (Nelson et al., 2007; Stevens and Nelson, 2011), therefore an indicator of socio-economic status was included throughout the thesis in models estimating dietary intake. The indicator recorded in the NDNS RP is the National Statistics Socio-Economic Classification (NSSEC). The NSSEC is a classification constructed to measure employment relations and conditions of occupations (Office for National Statistics, 2010) (**Tables 6a** and **b**). The NSSEC classes indicate a social class gradient with those in higher categories expected to have greater income and material advantage over those in lower classes (Rose et al., 2005). The “Never worked” category comprises those who have never worked and the long-term unemployed. Students, those in occupations not stated, inadequately described or not classifiable for other reasons make up the “Other” category. Retired participants remain in the same NSSEC category they were in prior to retirement. Detailed demographic characteristics are listed in Tables 6a and b.

**Table 6a**

Weighted demographic characteristics for females, males and all participants of the NDNS RP Years 1-4 (2008-2012)

		Females		Males		Total	
		Count	Percent	Count	Percent	Count	Percent
Age Groups	1.5 - 3y	66	1.6	63	1.5	130	3.1
	4-10y	169	4.1	161	3.9	330	8.0
	11-18y	211	5.1	200	4.8	411	9.9
	19-64y	1287	31.0	1294	31.2	2582	62.2
	65+	305	7.4	390	9.4	695	16.8
NS-SEC	Higher managerial & professional occupations	360	8.7	266	6.4	626	15.1
	Lower managerial & professional occupations	541	13.0	580	14.0	1121	27.0
	Intermediate occupations	149	3.6	193	4.6	341	8.2
	Small employers & own account workers	203	4.9	266	6.4	469	11.3
	Lower supervisory & technical occupations	217	5.2	200	4.8	418	10.1
	Semi-routine occupations	266	6.4	281	6.8	548	13.2
	Routine occupations	241	5.8	214	5.2	455	11.0
	Never worked & long-term unemployed	33	0.8	59	1.4	92	2.2
	Other	28	0.7	50	1.2	78	1.9
Ethnicity	White	1799	43.4	1872	45.1	3672	88.5
	Mixed ethnic group	29	0.7	41	1.0	69	1.7
	Black or Black British	67	1.6	70	1.7	136	3.3
	Asian or Asian British	103	2.5	96	2.3	199	4.8
	Any other group	41	1.0	31	0.7	72	1.7
Employment status	Employed	1044	25.2	862	20.8	1907	46.0
	Full time student	390	9.4	388	9.4	779	18.8
	Not working	554	13.3	811	19.6	1365	32.9

**Table 6b**

Weighted demographic characteristics for NDNS RP Years 1-4 (2008-2012)

		Females		Males		Total	
		Count	Percent	Count	Percent	Count	Percent
UK region	North east	94	2.3	80	1.9	173	4.2
	North west	228	5.5	235	5.7	462	11.2
	Yorkshire & the Humber	145	3.5	207	5.0	352	8.5
	East midlands	148	3.6	151	3.6	299	7.2
	West midlands	177	4.3	188	4.5	365	8.8
	East of England	218	5.3	168	4.0	386	9.3
	London	254	6.1	266	6.4	520	12.6
	South east	267	6.4	301	7.3	568	13.7
	South west	175	4.2	177	4.3	352	8.5
	Wales	96	2.3	106	2.6	202	4.9
	Scotland	177	4.3	171	4.1	348	8.4
Northern Ireland	60	1.4	59	1.4	119	2.9	
Highest qualification	Degree or equivalent	406	9.8	367	8.8	772	18.6
	Higher education, below degree level	145	3.5	191	4.6	336	8.1
	A levels or equivalent	274	6.6	235	5.7	509	12.3
	GCSE grades A-C or equivalent	282	6.8	317	7.7	599	14.4
	GCSE grades D-G or equivalent	57	1.4	51	1.2	108	2.6
	Foreign or other qualifications	77	1.9	69	1.7	147	3.5
	No qualifications	308	7.4	396	9.6	704	17.0
	Still in full-time education	116	2.8	126	3.0	242	5.8
Total	2038	49.2	2108	50.8	4148	100	

## 2.8 Strengths and limitations of the NDNS RP

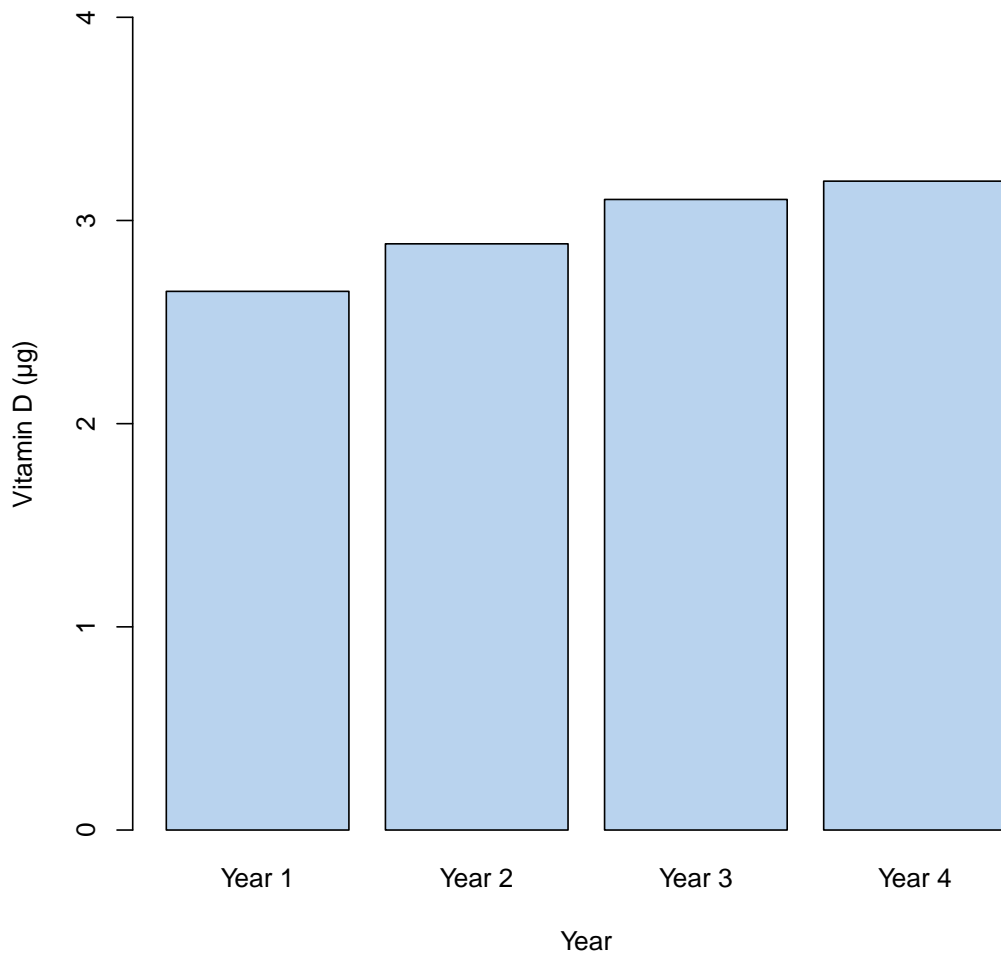
The NDNS RP is sampled such that over a yearly cycle participants are representative of UK citizens in terms of sex, age, location, and ethnicity. Furthermore daily and seasonal variations are captured. This makes it the foremost dataset for examining the dietary intakes of the UK and this is reflected in its widespread use in food and nutrition policy making, for example monitoring sugar reduction policies and dietary recommendations in the change4life initiative (Tedstone et al., 2014). Because of the rolling programme design that collects a complete cross-sectional sample each year it is possible to examine trends over time making this an important public health monitoring tool, see for example **Figure 15** showing an increase in average vitamin D intakes in Females aged 65 and over.

The NDNS RP Nutrient Databank (Smithers, 1993) is monitored and continually revised to match the foods recorded in the diary with those available and can be updated to reflect reformulations carried out by manufacturers. This is particularly important when foods become fortified as this often causes a dramatic change in the micronutrient content of the food and subsequently the overall diet. Whilst the Nutrient Databank is continually updated and therefore contains the most accurate possible data it also means that care should be taken when interpreting dietary trends to examine whether these represent genuine temporal differences in intakes rather than simply reflecting updated food composition.

Limitations to the NDNS RP are mentioned in Section 1.5 and include the possibility that a participant recording their dietary information may alter their diet or misreport actual consumption. This is a limitation inherent in all surveys using diet diaries and dietary assessment methods in general. Measures are included in the NDNS RP to mitigate this through the use of interviewers prompting for missing foods, participants being encouraged to collect food packaging, using a food atlas to prompt for portion sizes and DLW.

**Figure 15**

Mean daily vitamin D intake in 436 Females aged 65+ from NDNS RP Years 1-4 (2008-2014)  
by year of survey





### 3 Two-part models of complex survey data using a generalised gamma distribution: Dietary Iron intake in the UK

The aim of this chapter is to present methods of estimating usual mean intakes foods with an application to iron intake from five food groups that are the greatest contributors to iron in the diet. The approach comprises a two-part model capable of estimating mean intakes of episodically consumed foods using both parts of the model or, using the second part only, estimating mean intakes of habitually consumed foods. The current method of estimating intake from episodically consumed foods and nutrients in the NDNS RP is detailed in Section 1.12.1, highlighting that the only case where intake from episodically consumed food is considered is in alcohol, where descriptive statistics for all adults (regardless of alcohol consumption) and alcohol consumers only are reported. In the remainder of the NDNS RP analysis no distinction is made between intake from episodically and habitually consumed foods and nutrients with the within-person mean method used throughout. As discussed previously this method is likely to bias estimates as no consideration is made for the within- and between-person variance, nor the high frequency of zero observations. Methods such as the NCI and SPADE methods have been developed to address these issues but fail to handle data that are highly skewed and, in the case of the SPADE approach, the correlation between the probability of consumption and the amount consumed. To address this, novel methods are presented extending a two-part model with a generalised gamma distribution to analyse dietary components that exhibit skewed distributions collected using a complex sample design whilst including correlated random effects between both the probability of consumption and the amount consumed. The results from an application of these methods are compared to those from survey weighted linear regression.

This chapter is structured as follows: **Section 3.2** introduces the generalised gamma (GG) distribution. In **Section 3.3** the two-part model using a GG distribution and its extension to incorporate the complex sample design are introduced. **Section 3.4** presents an application of the two-part model estimating iron intake from foods that

contain iron using NDNS RP survey data with random effects in both model components. A comparison between the methods presented here and results from the currently used survey weighted regression analysis is given in **Section 3.6.1** and **Section 3.7** provides a discussion.

### 3.1 Introduction

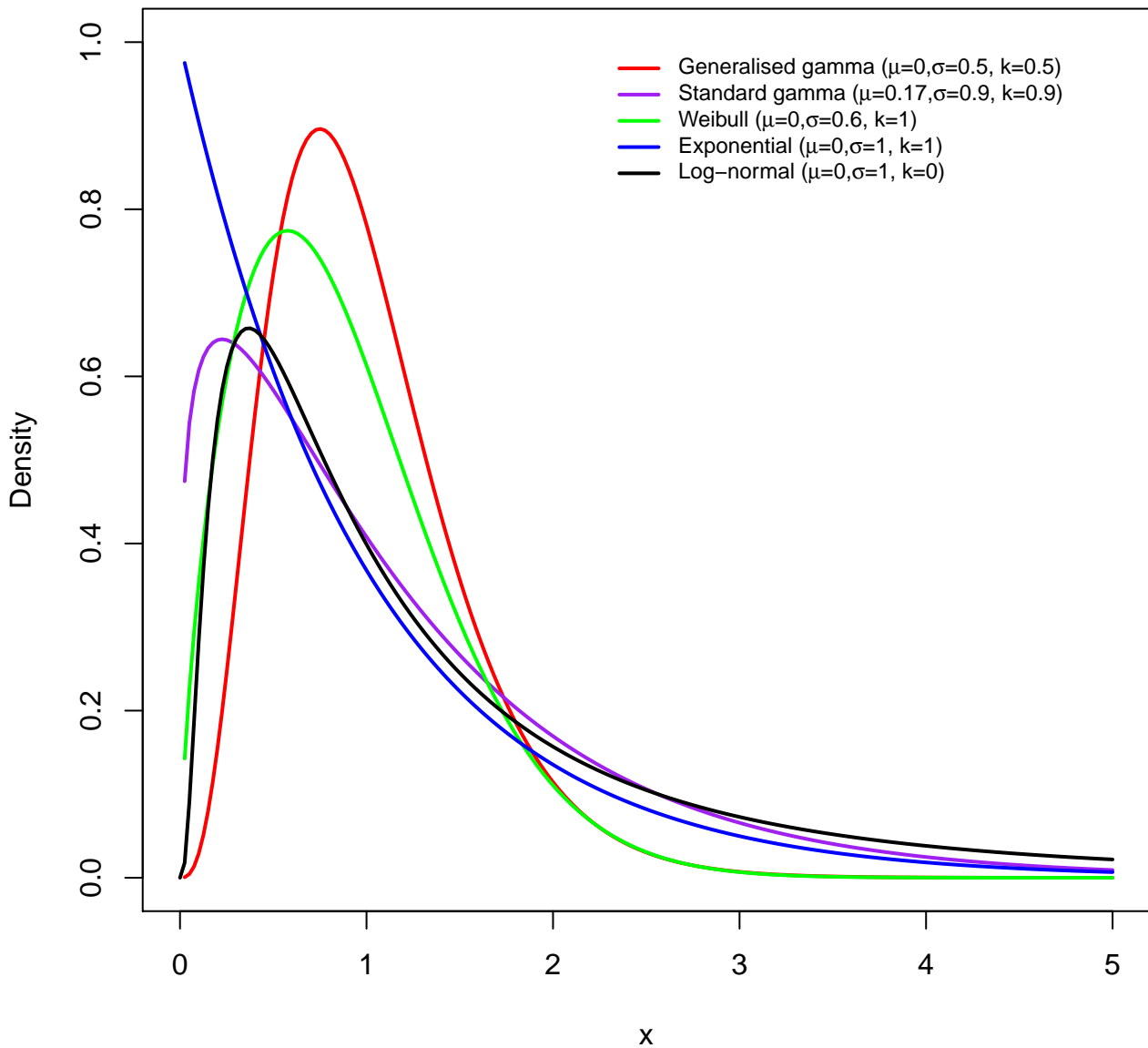
Dietary intake can only take non-negative values, and its distribution is often non-symmetric with a high frequency of zero observations, as shown in Figures 5 and 6, and to deal with this various methods have been proposed. These include transforming the data to normality using a Box-Cox transformation (Box and Cox, 1964) then carrying out ordinary linear regression followed by a transformation back to the original scale (Tooze et al., 2002). However this approach cannot cope with heteroscedastic data and therefore may lead to biased inference. Suggested alternative approaches to a two-part model are to either

- Include the zero observations and model all data
- Exclude the zero observations and model the positive part alone
- Change the zero observations into a positive value by adding a small constant.

However the suitability of the methods depends upon the data generating process and in this case the zeros represent a genuine process indicating a participants lack of consumption during the period of observation, rather than a true non-consumer. The GG distribution (Stacy, 1962; Manning et al., 2005) has been used as an attractive alternative to the use of a normal distribution to model non-negative data (Liu et al., 2008). The GG distribution offers flexibility on the shape of the distribution and contains the standard gamma, log-normal, exponential and Weibull distributions as special cases (**Figure 16**).

**Figure 16**

Generalised gamma distribution with varying parameter values displaying the standard gamma, Weibull, exponential and log-normal distributions.



The repeated measures, representing each day of intake, lead to observations that are correlated at the individual level and this has been dealt with through the introduction of random effects (Olsen and Schafer, 2001; Tooze et al., 2002). Here random effects are specified for each part of the model allowing a correlation structure to be imposed on both parts.

The motivating example for this chapter is an examination of the iron intake of commonly consumed foods that are episodically consumed and hence display a semi-continuous distribution, using data collected under a complex sampling plan. In individuals with iron deficiency, advice is given to increase the intake of foods containing iron. One approach is to include foods containing high amounts of iron such as liver, meat and beans in the diet. However this approach is likely to require a greater dietary change than simply increasing the intake of currently consumed foods that contain moderate sources of iron and may therefore be less successful. This is because food choices are often made on the basis of preference, taste and social context and dietary modifications in this way may have limited success. Dietary advice should be tailored to the recipient as diet has been shown to vary by age and sex (Steer et al., 2014). Therefore a modelling strategy of dietary intake that includes explanatory variables such as age and sex, allows the greatest sources of iron intake in groups with low levels of iron intake to be quantified. Furthermore providing these estimates based on a nationally representative sample taken from the NDNS RP brings additional challenges in incorporating the survey weighting, clustering and strata. The resulting estimates can be used to offer guidance to increase the consumption of foods containing moderate sources of iron that are currently eaten and will be therefore more likely to succeed, rather than advocating the increased consumption of foods with low consumption patterns and therefore unlikely to be consumed.

### 3.2 The generalised gamma distribution

The following is the definition of the GG distribution for a positive random variable  $Y$ . The probability density function following the parametrisation given by Manning et al. (2005) depends on the parameters  $k$ ,  $\mu$  and  $\sigma$  and is given by:

$$f(y; k, \mu, \sigma) = \frac{\gamma^\gamma}{\sigma y \sqrt{\gamma} \Gamma(\gamma)} \exp[z\sqrt{\gamma} - u] \quad y \geq 0 \quad (10)$$

where  $\Gamma(\cdot)$  denotes the standard gamma function  $\gamma = |k|^{-2} > 0$ , and  $z = \text{sign}(k)(\log(y) - \mu)/\sigma$  is dependent on the sign of the shape parameter  $k$ , the location parameter  $\mu > 0$ ,

the scale parameter  $\sigma > 0$  and  $u = \gamma \exp(|k|z)$ . The expectation of  $Y$  is given by:

$$E(Y) = \exp \left[ \mu + \left( \frac{\sigma}{k} \right) \log(k^2) + \log \left( \Gamma \left\{ \left( \frac{1}{k^2} \right) + \left( \frac{\sigma}{k} \right) \right\} \right) - \log \left( \Gamma \left\{ \frac{1}{k^2} \right\} \right) \right] \quad (11)$$

and the variance can be written as:

$$\text{Var}(Y) = \left\{ \exp(\mu) \cdot k^{2\sigma/k} \right\}^2 \left\{ \frac{\Gamma(1/k^2 + 2\sigma/k)}{\Gamma(1/k^2)} - \left[ \frac{\Gamma(1/k^2 + 2\sigma/k)}{\Gamma(1/k^2)} \right]^{-2} \right\} \quad (12)$$

### 3.3 Two part model

The two-part model for longitudinal semi-continuous data introduced in Section 1.10 is re-formulated to substitute the linear mixed-effects model used to model the continuous component by a GG distribution with random effects. This comprises a logistic mixed-effects model in the first part and a generalised gamma mixed-effects model for the second part where the GG distribution follows that proposed by Liu et al. (2010). The semi-continuous observation for participant  $i$  where  $i = 1, 2, \dots, N$  and on day  $j$ ,  $j = 1, 2, \dots, n_i$  is represented by two random variables an indicator of consumption  $Z_{ij}$  and the amount consumed  $Y_{ij}$  given that the indicator is equal to 1

$$Z_{ij} = \begin{cases} 1, & \text{if } Y_{ij} > 0 \\ 0, & \text{if } Y_{ij} = 0 \end{cases}$$

The indicator variable  $Y_{ij}$  is modelled using a logistic regression model that includes a random effect to account for the correlation between repeated observations, as given in Expression 6 in Section 1.10, is

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \mathbf{X}'_{ij} \boldsymbol{\beta} + u_i, \quad (13)$$

where  $\pi_{ij} = \Pr(Z_{ij} = 1 | u_i)$ ,  $\mathbf{X}'_{ij}$  is a  $1 \times q$  vector of explanatory variables,  $\boldsymbol{\beta}$  is a  $q \times 1$  vector of regression coefficients and  $u_i$  is a random intercept. The contribution of participant  $i$  to the log likelihood from the logistic regression part given the random effect  $u_i$ , as given in Expression 7 in Section 1.10, is

$$\ell_{Y_i} = \sum_{j=1}^{n_i} Z_{ij} \log(\pi_{ij}) + (1 - Z_{ij}) \log(1 - \pi_{ij}). \quad (14)$$

The positive consumption amount  $Y_{ij}|(Y_{ij} > 0, v_i)$  is modelled using the GG distribution of Section 3.2 with the location parameter modelled as:

$$\mu_{ij} = \mathbf{X}_{ij}^{*'} \boldsymbol{\gamma} + v_i$$

where  $\mathbf{X}_{ij}^{*'}$  is a  $1 \times p$  vector of explanatory variables,  $\boldsymbol{\gamma}$  is a  $p \times 1$  vector of regression coefficients, and  $v_i$  is a random intercept. The contribution to the likelihood from the second part is given by

$$\ell_{Y_i} = \sum_{j=1}^{n_i^*} \left[ (\gamma - 0.5) \log(\gamma) - \log(\sigma) - \log(z_{ij}) - \log(\Gamma(\gamma)) + y_{ij} \sqrt{\gamma} - \gamma \exp(|k|y_{ij}) \right] \quad (15)$$

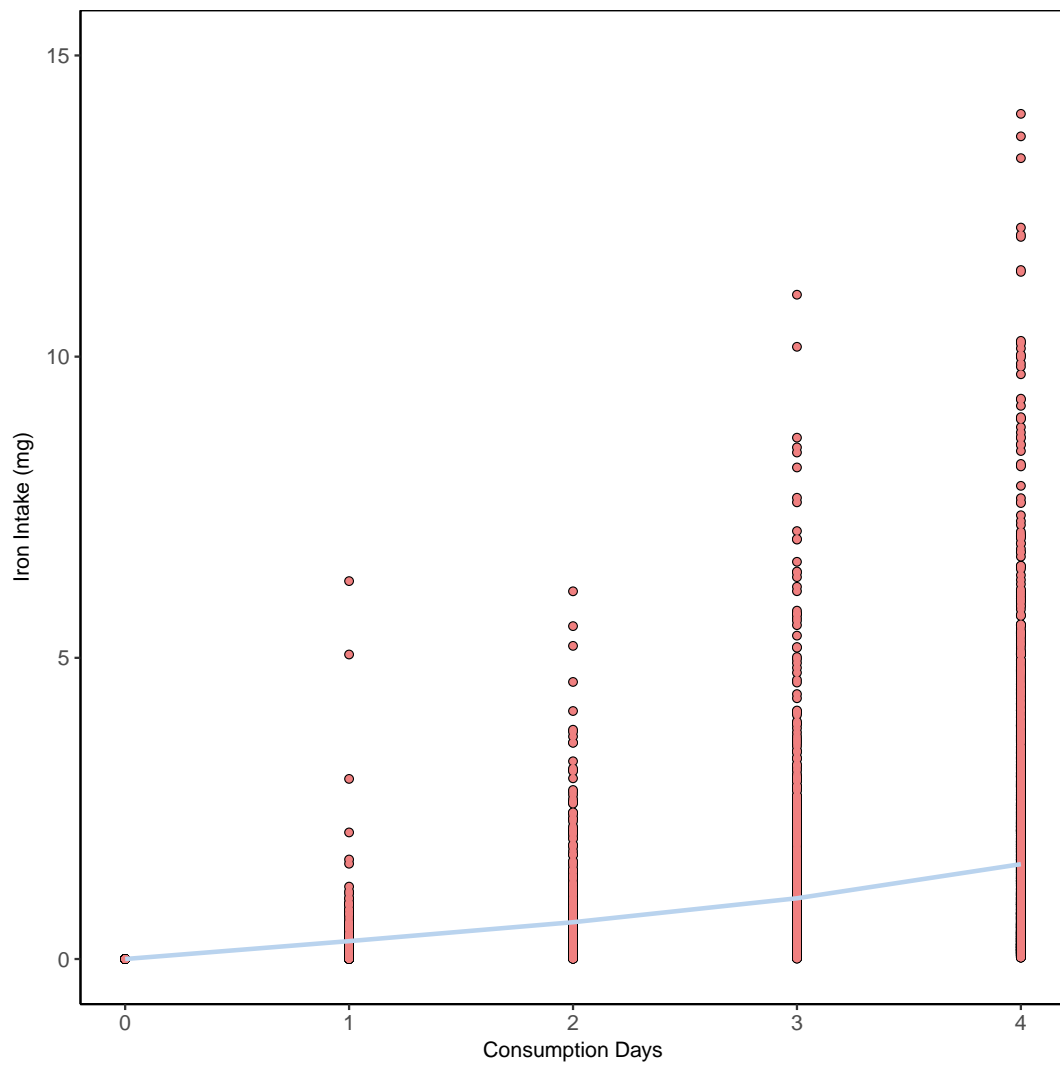
where  $n_i^*$  is the number of positive intakes for individual  $i$  and  $z_{ij} = \text{sign}(k) (\log(y_{ij}) - \mu_{ij}) / \sigma$ . To account for the cross equation correlation the two random effects,  $u_i$  and  $v_i$ , have a bivariate normal distribution with means zero and covariance matrix  $\Sigma$  where  $\Sigma$  is a positive definite covariance matrix and this will capture the positive correlation between days of consumption and the amount of food consumed as shown empirically in **Figure 17**. Allowing the random effects from both parts to be correlated leads to a better model fit (Neelon et al., 2016) and prevents possible biased inferences (Su et al., 2009). Various approaches to estimating the intractable integrals necessary to integrate out the random effects have been proposed, including sixth order Laplace approximation (Olsen and Schafer, 2001) and adaptive Gaussian quadrature (Tooze et al., 2002). The model can be extended to account for heteroscedasticity through modelling of the scale parameter  $\sigma$  as a function of explanatory variables  $\mathbf{X}_{ij}$  where  $\sigma_{ij}^2 = \exp(\mathbf{X}_{ij}' \boldsymbol{\delta})$ , and  $\boldsymbol{\delta}$  is a vector of regression coefficients.

### 3.3.1 Extension to multistage sampling

The target population comprises  $L$  strata denoted by  $l, l = 1, \dots, L$  which are divided into  $N_l$  primary sampling units (PSU) denoted by  $k, k = 1, \dots, K$  then using simple random sampling  $N_{kl}$  individuals are sampled from PSU  $k$  in stratum  $l$ . The dietary records are denoted by  $y_{ijkl}, j = 1, \dots, N_{ikl}$  clustered within individual  $i, i = 1, \dots, N_{kl}$  in PSU  $k$  in stratum  $l$ . To adjust the characteristics of the sampled individuals back to those of the target population and to correct for selection bias and individual non-response, weights  $w_{ikl}$  are created for the  $i$ th individual within the  $k$ th PSU in the  $l$ th stratum. The survey

**Figure 17**

Iron intake from vegetables by number of days of consumption using data from NDNS RP 2008-12 for 4156 participants aged 1.5 years and over



weights are included using a pseudo likelihood approach by multiplying the likelihood function by the weighting at the individual level.

### 3.3.2 Standard error estimation of model parameters

Variance estimation that does not consider the multilevel sampling will typically lead to biased estimates (Rabe-Hesketh and Skrondal, 2006); thus strata and PSU should be considered and this can be done through the bootstrap technique. The covariance matrix of the maximum pseudo likelihood estimate can be estimated through bootstrap:  $B$  bootstrap replicates are generated by random sampling from each PSU in each stratum, from each replicate the pseudo likelihood parameter estimate  $\hat{\theta}_b^{(p)}$  ( $b = 1, \dots, B$ ) is obtained. The bootstrap estimate of the covariance matrix  $\hat{\theta}^{(p)}$  is given by

$$\text{cov}(\hat{\theta}^{(p)}) = \frac{A}{B} \sum_{b=1}^B (\hat{\theta}_b^{(p)} - \hat{\theta}^{(p)*})(\hat{\theta}_b^{(p)} - \hat{\theta}^{(p)*})^T$$

where  $\hat{\theta}^{(p)*} = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^{(p)*}$  and  $A$  is a scaling factor defined by  $A = \frac{\bar{M}_l}{\bar{M}_{l-1}}$ , where  $\bar{M}_l$  is the average number of PSUs per stratum. The weights for each bootstrap replicate need to be adjusted according to the sampling method (Canty and Davison, 1999). For example, if  $N_l$  PSUs are sampled in stratum  $l$  then the adjusted weights are  $w_{ikl}^b = w_{ikl} k_{kl}^b$ , where  $w_{ikl}$  is the original weight for the  $i$ th individual in the  $k$ th PSU of the  $l$ th stratum, and  $k_{kl}^b$  is the number of repeated samples from the  $k$ th PSU of the  $l$ th stratum, in the  $b$ th bootstrap replicate.

For implementation in SAS, the NLMIXED procedure (Littell et al., 2006) is used. Point estimates are calculated using a single modelling fitting with all participants and weighted using the individual survey weighting. However the NLMIXED procedure does not allow arguments specifying a complex survey design to be included and is only capable of taking a single set of frequency weights. Variance estimates need to adjust for the complex survey design and therefore a set of individual weights that adjust for the complex survey design and simultaneously the individual survey weighting were created. These were calculated by sampling at the PSU level with replacement to get a data frame of PSUs, the same number as in the original NDNS RP data. A count of the number of times each PSU was sampled is kept, e.g. if PSU is included 3 times



then the bootstrap weight column will contain 3. This bootstrap weight is then multiplied by the survey weight for the same individual and this newly created weight enters the pseudo likelihood and the model is fitted. The process is repeated with a new sample of PSUs drawn from the NDNS RP and the count of the number of times they are sampled kept and multiplied by the corresponding survey weight for each individual and is repeated  $B$  times and the average of the estimates resulting from the model fitted to each bootstrapped sample is taken. This is illustrated in **Figure 18** where step 1. shows the data with ID numbers, an iron value, the primary sampling unit for each individual and their individual survey weight. Step 2. shows the list of sampled PSUs, in step 3. a count of the PSUs is taken, this is referred to as the bootstrap weight and finally the individual's survey weight is multiplied by the bootstrap weight to create the new weight which will enter the model.

**Figure 18**

Illustration of the process to create a survey adjusted single weight for sampled PSUs in the NDNS participants.

1. NDNS Data				2. Sampled PSUs	3. Count of Sampled PSUs		4. Sampled data with created survey adjusted weighting					
ID number	Iron value	Primary Sampling Unit	Survey weight	Primary Sampling Unit	Primary Sampling Unit	Bootstrap weight	ID number	Iron value	Primary Sampling Unit	Survey weight	Bootstrap weight	New weighting
1	4.05	10	1.74890	1	1	3	1	4.05	1	1.74890	3	5.2467
1	4.20	10	1.74890	1	2	1	1	4.20	1	1.74890	3	5.2467
1	1.02	10	1.74890	1	3	0	1	1.02	1	1.74890	3	5.2467
1	4.92	10	1.74890	2	4	1	1	4.92	1	1.74890	3	5.2467
2	3.88	10	1.48298	4	5	0	2	3.88	1	1.48298	3	4.44894
2	4.05	10	1.48298	6	6	2	2	4.05	1	1.48298	3	4.44894
2	1.34	10	1.48298	6	7	0	2	1.34	1	1.48298	3	4.44894
2	4.87	10	1.48298	10	8	0	2	4.87	1	1.48298	3	4.44894
3	1.62	10	0.82921	10	9	0	3	1.62	1	0.82921	3	2.48763
3	4.63	10	0.82921	10	10	3	3	4.63	1	0.82921	3	2.48763
3	4.88	10	0.82921	11	11	1	3	4.88	1	0.82921	3	2.48763
3	2.23	10	0.82921	13	12	0	3	2.23	1	0.82921	3	2.48763
...	...	...	...	...	...	...	...	...	...	...	...	...
6828	1.28	300	1.10031	295	297	0	6828	1.28	300	1.10031	0	0
6828	1.58	300	1.10031	298	298	2	6828	1.58	300	1.10031	0	0
6828	1.16	300	1.10031	298	299	1	6828	1.16	300	1.10031	0	0
6828	2.31	300	1.10031	299	300	0	6828	2.31	300	1.10031	0	0

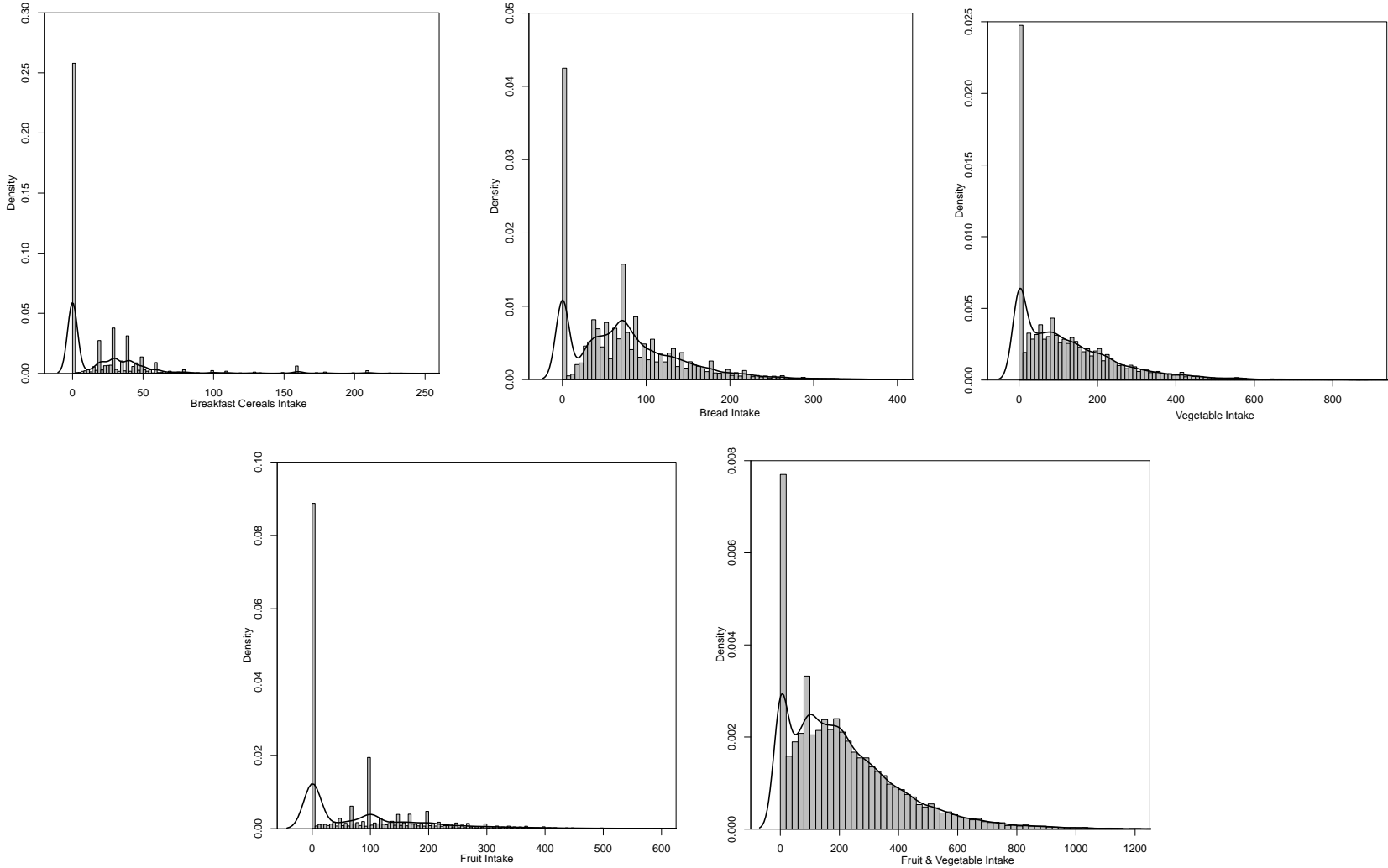
### 3.4 Application of the two-part model on iron consumption from selected food groups in the NDNS RP

Iron is habitually consumed by almost all individuals on all days and its intake can thus be modelled relatively simply as its distribution is unlikely to have a peak at zero. Yet to make effective recommendations that fit with national food intake guidelines to increase iron consumption it is important to examine foods that are sources of iron rather than the nutrient itself. Foods, however, are more likely to be consumed episodically and therefore will intake exhibit a semi-continuous distribution requiring a two-part model approach. To illustrate the methods the NDNS RP Years 1-4 presented in Section 2 (Public Health England, 2014) is considered. Participants are asked to record everything they consume over a four day period with estimated portion sizes which are then entered into a database (Fitt et al., 2014) using appropriate food codes. The NDNS RP contains approximately 3700 unique food codes which are grouped into 58 food groups (excluding dietary supplements and artificial sweeteners). The contribution of iron intake from each food group is displayed in **Figure 20** and from here five food groups were chosen to illustrate the flexibility of the two-part model presented. These are: 'Breakfast cereals' which is a combination of 'high fibre breakfast cereals' and 'other breakfast cereals'; 'Bread' which combines the food groups 'white bread', 'wholemeal bread', 'brown, granary and wheatgerm bread' and 'other bread'; 'Vegetables' which combines the food groups 'Vegetables not raw' and 'salad and other raw vegetables', 'Chips, fried and roast potatoes and potato products' and 'Other potatoes, potato salads and dishes'; 'Fruit' which is composed of the 'Fruit' group and 'Fruit and Vegetables' which is a combination of the 'Fruit' and 'Vegetables' food groups (Figure 21b). The food groups breakfast cereals, bread and pasta rice and other cereals are all significant contributors to dietary iron intake due to the mandatory fortification of flour with iron in the UK (Food Standards Agency, 2008). These particular food groups were also chosen to fit with dietary recommendations (NHS Choices, 2011), at the expense of other foods that are good sources of iron for example meat, as recommending an increase in the consumption of meat would be at odds with the eatwell plate which suggests individuals should consume small amounts of red and processed meat. In addition the combined fruit and vegetables group was chosen to align with the five a day campaign

(NHS Choices, 2016). Histograms with a small number of extreme outliers removed for ease of interpretation (partial histograms) for the 5 food groups are presented in **Figure 19** also displayed are the percentage of zeros, to demonstrate the flexibility of the methods in coping with varying distributions from >50% zeros where a zero indicates a diary record where the food has not been consumed.

**Figure 19**

Partial histograms with kernel density estimates (dark line) for iron intake of selected food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over.



### 3.4.1 Model specification

Increases in iron requirements occur during periods of growth and to replace menstrual losses; consequently age and sex are important factors to consider and are included in the specification of the location parameter of the GG regression model and the logistic regression model. In addition socio-economic class (NSSEC) is included as there appears to be an impact upon iron status that is inversely proportional to class (Thane et al., 2000; Bates et al., 1999). The percentage contribution to total iron of all foods is displayed in **Figure 20** and then for the selected foods in **Figure 21a**. **Figure 21b** displays the number of zeros present in each food group describing a range of cases from breakfast cereals that had approximately 50% zeros to the fruit and vegetable category that had less than 20% zeros. The two parts of the model are specified as follows, firstly the mixed-effects logistic regression model (part 1):

$$\log\left(\frac{\pi_{ijkl}}{1 - \pi_{ijkl}}\right) = \alpha_0 + \alpha_1 \text{Sex}_{ikl} + \alpha_2 \text{AgeGroup}_{ikl} + \alpha_3 \text{NSSEC}_{ikl} + u_i, \quad (16)$$

and for the GG mixed-effects regression model (part 2):

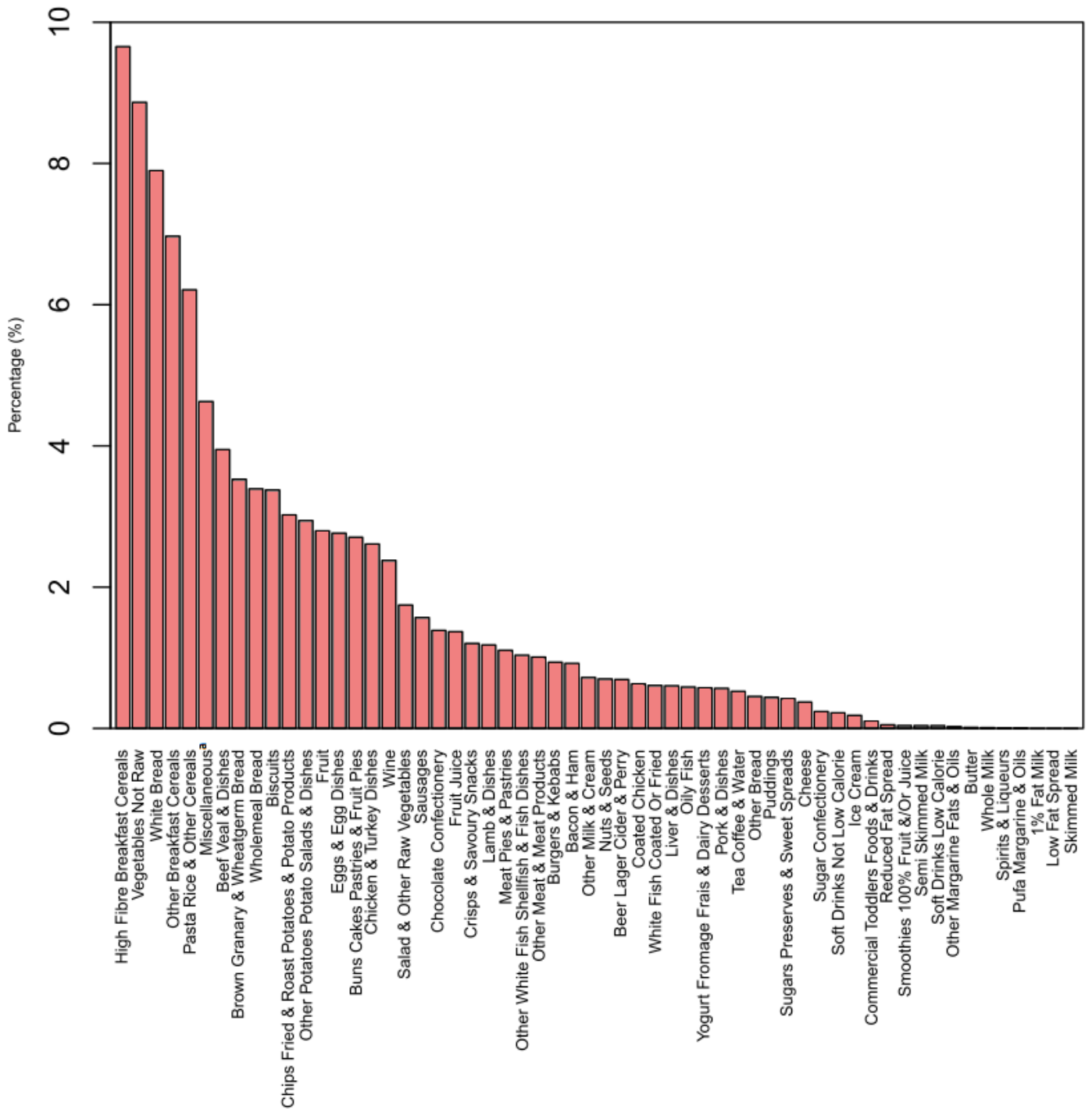
$$\mu_{ijkl} = \beta_0 + \beta_1 \text{Sex}_{ikl} + \beta_2 \text{AgeGroup}_{ikl} + \beta_3 \text{NSSEC}_{ikl} + v_i, \quad (17)$$

where  $\pi_{ijkl}$  is the probability of consumption and  $\mu_{ijkl}$  is the location parameter of the GG for the model of food intake for individual  $i$  on day  $j$  in PSU  $k$  within strata  $j$  given that intake occurred and  $u_i$  and  $v_i$  are random intercepts.

Maximum likelihood estimation was carried out using adaptive Gaussian quadrature in the SAS NLMIXED procedure (Littell et al., 2006) using 5 quadrature points following Liu et al. (2010) who found that increasing the number of points to 10 had little impact upon results but did take significantly longer for convergence to be reached. To confirm the findings, a study was carried out here by comparing standard errors for iron in take from vegetables estimated with 5, 10, 15 and 20 quadrature points in terms of accuracy and the time taken for each model to converge. In total 200 models were run, 50 with 5 quadrature points, 50 with 10 quadrature points, 50 with 15 quadrature points and 50 with 20 quadrature points. An error was returned by the software when overall

**Figure 20**

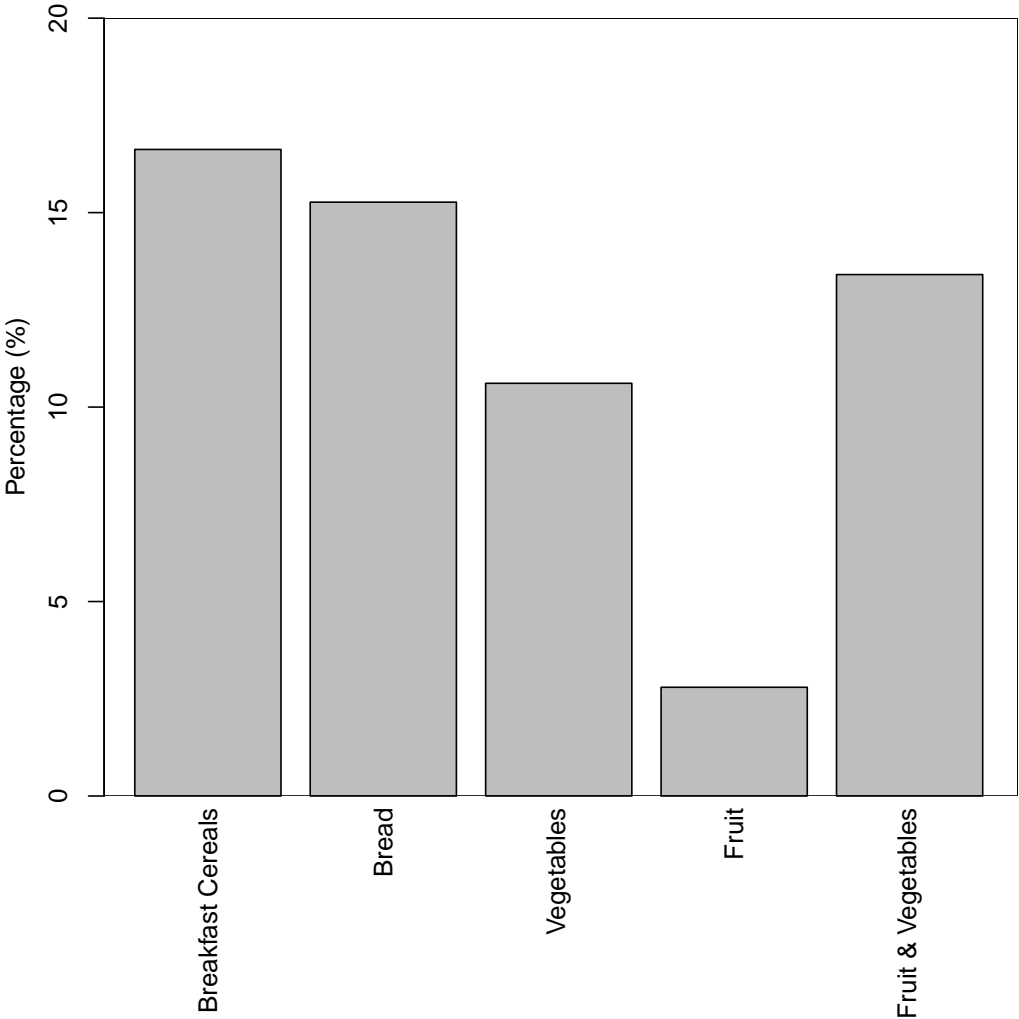
The percentage contribution to total iron intake by all food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 core participants aged 1.5 years and over



<sup>a</sup>The Miscellaneous food group includes iron fortified nutrition powders and drinks along with foods in the food groups: dry weight beverages; soup, manufactured/retail and homemade; savoury sauces, pickles, gravies and condiments; and commercial toddler foods.

**Figure 21a**

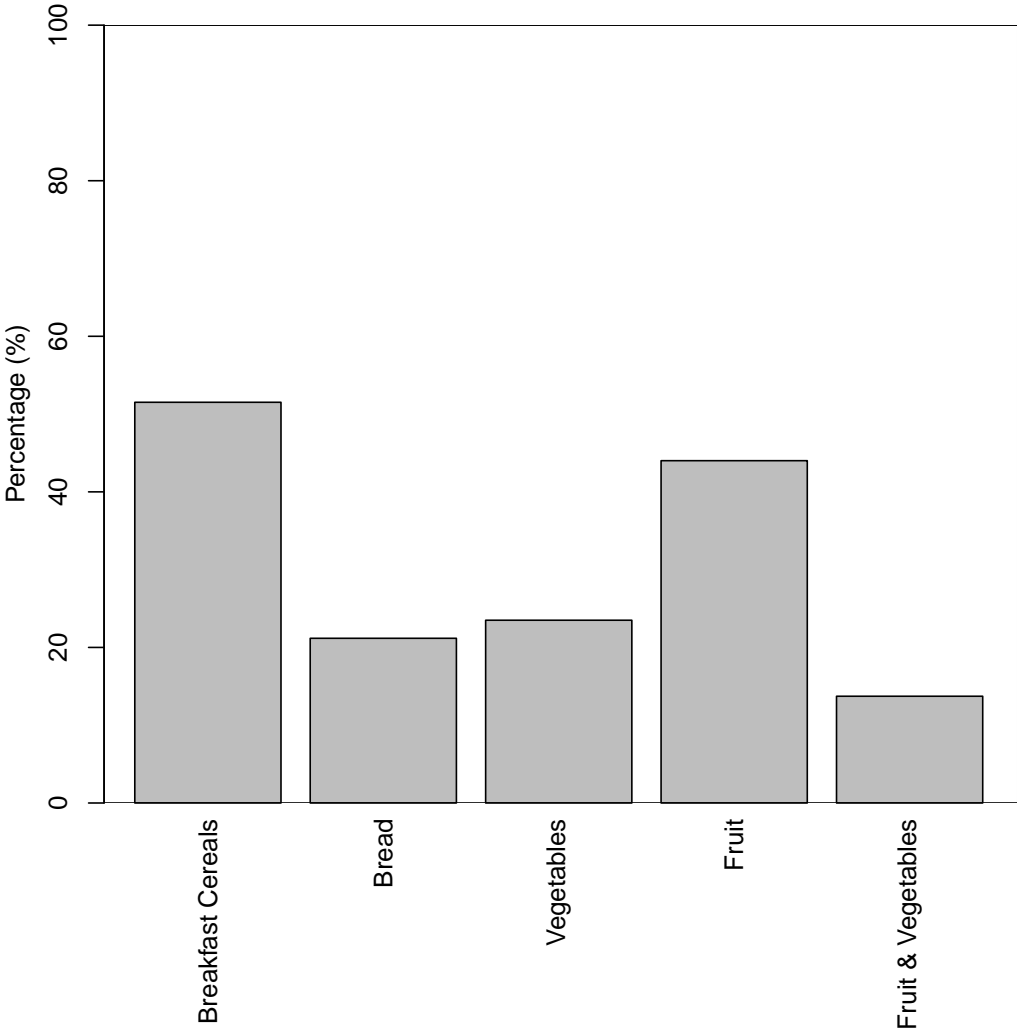
The percentage contribution to total iron intake by selected food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over.





**Figure 21b**

The proportion of zero consumption days by selected food groups using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over.



convergence did not occur and the procedure stopped running. In some cases estimation for individual parameters could not be completed and the remaining parameters were returned by the software, when this happened the results from the entire model were excluded thus results used throughout are only from models where all parameters were estimated with the desired convergence threshold. The time taken to fit 50 models varied from 15 hours and 13 minutes when ran with 5 quadrature points to 577 hours and 42 minutes with 20 quadrature points, though there was minimal difference between the means of the estimated standard errors comparing 5 quadrature points to the other three scenarios (see **Appendix E**). As a result of these findings, 5 quadrature points was deemed sufficient for model estimates and was used in all cases with the two-part model.

### **3.5 Standard error of parameter estimates**

To estimate the standard errors of the parameter estimates of the two-part model, bootstrap resampling was carried out as detailed in section 3.3.2. The number of bootstrap samples carried out and presented in the results shown in **Tables 8b-e** was 50. Choosing an appropriate number of bootstrap samples to carry out is a pragmatic trade-off dependent on the complexity of the model being fitted and hence the time taken against the convergence of the bootstrap estimates (Andrews and Buchinsky, 1996). To examine the extent to which the standard error accuracy was affected the two-part model was run with 300 bootstrap samples. Then the standard errors were calculated by taking the average of either 50, 100, 200 and 300 bootstrap samples representing 4 possible bootstrap scenarios, the results are shown in **Tables 10a** and **b** for vegetables only and for the other food groups in **Appendix D**.

### **3.6 Results**

**Table 7** displays the mean, standard deviation, median, interquartile range, minimum and maximum number of grams of food consumed each day for each of the chosen

food groups along with the percentage of days that contain no recorded intake. The parameter estimates of the fitted two-part model are displayed in **Tables 8a-e**. The tables show the estimated regression parameters from the two-part model for different food groups, with part 1 estimates, SE, and p-values relating to the propensity to consume iron from the given food and part 2 estimates, SE, and p-values relating to the mean intake of iron from the given food. Also displayed in the tables are the estimated shape  $k$  and scale  $\sigma$  parameters from the GG distribution and the random effects variance components, that is the variance of the random effects ( $\hat{\sigma}_u, \hat{\sigma}_v$ ) and their covariance. Correlations between part 1 and 2 are given by  $\widehat{cov}(u_i, v_i) / \hat{\sigma}_u \hat{\sigma}_v$  where a positive (negative) value indicates a positive (negative) correlation between the propensity of consuming iron from the food and the amount of food consumed. The model regression coefficients of the first part are the log odds ratio of consumption and those from the second part, given that the intake was positive, indicate the differences in the mean for a one unit increase on the explanatory variable, where it is continuous or the mean difference between categories for categorical variables which are conditional upon the random effects. For illustration, the impact of varying the shape  $k$  and scale ( $\sigma$ ) parameters of the GG is shown in Figure 16 where for example a  $k$  value of 0 and a  $\sigma$  value of 1 gives the Log-normal distribution and a  $k$  value of 1 and a  $\sigma$  value of 0.6 gives the Weibull distribution.

Females had greater odds of consuming iron from breakfast cereals (OR=1.5, 95% CI:1.2, 1.8,  $p < 0.001$ ) but consumed smaller amounts when compared to males ( $\hat{\gamma} = -0.11$ , 95% CI:-1.2, -0.1,  $p < 0.001$ ). Similar results were found for iron from vegetables but not iron from bread where males had higher odds of consuming and consumed greater quantities than females.

All age groups consumed greater amounts of iron from bread, vegetables and the fruit and vegetables food groups compared to the reference group 1.5-3y, though the odds of consuming varied. Odds ratios for consuming iron from breakfast cereals varied across age groups with the 4-10y age group having a higher odds ratio (OR=2.5, 95% CI:1.3, 4.7,  $p = 0.007$ ) whereas the 19-64y group had a lower odds ratio (OR=0.4, 95% CI:0.2, 0.7,  $p = 0.003$ ) compared to the youngest reference group. The odds

**Table 7**

Descriptive characteristics for selected foods using data from NDNS RP Years 1-4, (2008-2012).

	Breakfast Cereals	Bread	Vegetables	Fruit	Fruit & Vegetables
Mean (g)	23.3	76.1	127.5	91.9	219.4
sd (g)	38.8	67.5	137.3	122.5	198.5
Median (g)	0.0	70.0	90.0	50.0	179.3
Interquartile range	36.0	82.0	177.5	150.0	243.1
Min	0.0	0.0	0.0	0.0	0.0
Max (g)	490.2	675.6	1271.7	2306.5	2476.5
Zeros (%)	51.5	21.2	23.5	44.0	13.7

ratio of consuming iron from fruit was lower in the 11-18y group (OR=0.5, 95% CI:0.3, 0.9,  $p = 0.02$ ) but higher in the oldest 65+y age group (OR=2.4, 95% CI:1.4, 3.9,  $p < 0.001$ ), however in both cases the amount consumed was higher compared to the 1.5-3y reference group ( $\hat{\gamma}=0.20$ , 95% CI:0.0, 0.4,  $p = 0.046$  &  $\hat{\gamma}=0.56$ , 95% CI:0.4, 0.7,  $p < 0.001$  respectively)

When examining the regression coefficients for NSSEC, many groups differed from the reference higher managerial and professional occupations group. Those in the lower managerial and professional occupations (OR=0.6, 95% CI:0.4, 0.7,  $p < 0.001$ ,  $\hat{\gamma}=-0.07$ , 95% CI:-0.1, -0.01,  $p=0.01$ ), intermediate occupations (OR=0.5, 95% CI:0.4, 0.8,  $p=0.0015$ ,  $\hat{\gamma}=-0.13$ , 95% CI:-0.2, -0.05,  $p=0.001$ ) and semi-routine occupations (OR=0.6, 95% CI:0.5, 0.8,  $p=0.002$ ,  $\hat{\gamma}=-0.09$ , 95% CI:-0.1, -0.03,  $p=0.01$ ) classes all had significantly lower odds ratios for the consumption of iron from bread and consumed lower amounts where it was consumed. Iron from fruit intake also showed a significantly lower odds ratio and significantly smaller amounts were consumed in the lower managerial and professional occupations (OR=0.6, 95% CI:0.5, 0.8,  $p < 0.001$ ,  $\hat{\gamma}=-0.15$ , 95% CI:-0.2, -0.05,  $p=0.003$ ), intermediate occupations (OR=0.4, 95% CI:0.3, 0.6,  $p < 0.001$ ,  $\hat{\gamma}=-0.16$ , 95% CI:-0.3, -0.02,  $p=0.02$ ), small employers and own ac-

count workers (OR=0.7, 95% CI:0.5, 1.0,  $p=0.02$ ,  $\hat{\gamma}=-0.13$ , 95% CI:-0.2, -0.01,  $p=0.04$ ), lower supervisory and technical occupations (OR=0.4, 95% CI:0.3, 0.6,  $p<0.001$ ,  $\hat{\gamma}=-0.22$ , 95% CI:-0.3, -0.1,  $p<0.001$ ), semi-routine occupations (OR=0.3, 95% CI:0.2, 0.4  $p<0.001$ ,  $\hat{\gamma}=-0.23$ , 95% CI:-0.3, -0.1,  $p<0.001$ ) and routine occupations (OR=0.3, 95% CI:0.2, 0.4,  $p<0.001$ ,  $\hat{\gamma}=-0.29$ , 95% CI:-0.4, -0.2,  $p<0.001$ ) classes compared to the reference group. The routine occupations group also had lower odds ratios and consumed lower amounts of iron from vegetables (OR=0.4, 95% CI:0.3, 0.5,  $p<0.001$ ,  $\hat{\gamma}=-0.15$ , 95% CI:-0.27, -0.03,  $p=0.01$ ) and the fruit and vegetables combined categories (OR=0.4, 95% CI:0.3, 0.5,  $p<0.001$ ,  $\hat{\gamma}=-0.19$ , 95% CI:-0.3, -0.07,  $p<0.001$ ) when compared to the higher managerial and professional reference group. Along with the semi routine occupations group which also had a lower odds ratio for the consumption of iron from fruit and vegetables and consumed smaller amounts (OR=0.4, 95% CI:0.3, 0.5,  $p<0.001$ ,  $\hat{\gamma}=-0.15$ , 95% CI:-0.3, -0.03,  $p=0.002$ ) compared to the reference group.

Correlations between the random effects in parts 1 and 2 ranged from a fairly weak 0.12 for iron from breakfast cereals to highly correlated values of 0.92 and 0.94 in the fruit & vegetables group and the bread groups respectively, strong correlations between the probability of consuming iron and the amount consumed were also seen for the separate fruit and vegetable groups of 0.65 and 0.78 respectively, highlighting that these correlations are important and should be considered though correlated random effects.

**Table 8a**

Estimated parameters of the two-part model for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Part 1			Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	0.41	0.10	<0.001	-0.11	0.03	<0.001
Age Group	1.5 -3y (Reference)						
	4-10y	0.91	0.33	0.007	0.48	0.09	<0.001
	11-18y	-0.10	0.33	0.75	0.84	0.09	<0.001
	19-64y	-0.88	0.29	0.003	0.37	0.08	<0.001
	65y and older	0.80	0.31	0.01	-0.002	0.08	0.98
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	0.02	0.16	0.92	0.03	0.05	0.48
	Intermediate occupations	0.10	0.22	0.64	0.01	0.07	0.87
	Small employers & own account workers	-0.003	0.20	0.99	-0.01	0.06	0.88
	Lower supervisory & technical occupations	0.005	0.21	0.98	-0.01	0.06	0.91
	Semi-routine occupations	-0.10	0.19	0.60	0.01	0.06	0.80
	Routine occupations	-0.40	0.20	0.047	0.01	0.06	0.90
	Never worked	-0.001	0.36	0.998	-0.002	0.11	0.99
	Other	0.50	0.39	0.20	-0.002	0.12	0.99
$\hat{k}$ , GG distribution shape parameter					2.93	0.07	<0.001
$\hat{\sigma}$ , GG distribution scale parameter					0.01	0.05	<0.001
		Estimates			SE		p-value
Variance components	$\hat{\sigma}_u$	6.01			0.21		<0.001
	$\hat{\sigma}_v$	0.32			0.02		<0.001
	$\widehat{cov}(u_i, v_i)$	0.24			0.06		<0.001

**Table 8b**

Estimated parameters of the two-part model for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Part 1			Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	-0.38	0.08	<0.001	-0.24	0.02	<0.001
Age Group	1.5 -3y (Reference)						
	4-10y	0.38	0.27	0.15	0.32	0.06	<0.001
	11-18y	-0.04	0.25	0.89	0.49	0.06	<0.001
	19-64y	0.01	0.23	0.95	0.58	0.05	<0.001
	65y and older	1.01	0.25	<0.001	0.44	0.05	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.56	0.13	<0.001	-0.07	0.03	0.01
	Intermediate occupations	-0.61	0.18	0.0015	-0.13	0.04	<0.001
	Small employers & own account workers	-0.26	0.16	0.12	-0.07	0.03	0.03
	Lower supervisory & technical occupations	-0.47	0.17	0.004	-0.04	0.04	0.26
	Semi-routine occupations	-0.48	0.15	0.002	-0.09	0.03	0.01
	Routine occupations	-0.54	0.16	<0.001	-0.05	0.03	0.19
	Never worked	-0.99	0.28	<0.001	-0.04	0.06	0.55
	Other	-0.45	0.31	0.14	0.04	0.07	0.54
$\hat{k}$ , GG distribution shape parameter					0.50	0.10	<0.001
$\hat{\sigma}$ , GG distribution scale parameter					1.22	0.05	<0.001
		Estimates		SE		p-value	
Variance components	$\hat{\sigma}_u$	1.93		0.15		<0.001	
	$\hat{\sigma}_v$	0.11		0.01		<0.001	
	$\widehat{cov}(u_i, v_i)$	0.20		0.02		<0.001	

**Table 8c**

Estimated parameters of the two-part model for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Part 1			Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	0.40	0.076	<0.001	-0.08	0.03	0.01
Age Group	1.5 -3y (Reference)						
	4-10y	0.01	0.24	0.97	0.28	0.10	0.01
	11-18y	-0.51	0.23	0.03	0.43	0.10	<0.001
	19-64y	0.41	0.21	0.053	0.79	0.09	<0.001
	65y and older	1.00	0.23	<0.001	0.69	0.10	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.40	0.13	0.002	-0.07	0.05	0.17
	Intermediate occupations	-0.89	0.17	<0.001	-0.10	0.07	0.14
	Small employers & own account workers	-0.19	0.16	0.22	-0.05	0.06	0.42
	Lower supervisory & technical occupations	-0.69	0.16	<0.001	-0.10	0.06	0.09
	Semi-routine occupations	-1.10	0.14	<0.001	-0.09	0.06	0.15
	Routine occupations	-1.03	0.15	<0.001	-0.15	0.06	0.01
	Never worked	-1.00	0.27	<0.001	-0.12	0.11	0.29
	Other	-0.40	0.30	0.19	-0.12	0.12	0.32
$\hat{k}$ , GG distribution shape parameter					0.76	0.09	<0.001
$\hat{\sigma}$ , GG distribution scale parameter					0.30	0.05	<0.001
		Estimates		SE		p-value	
Variance components	$\hat{\sigma}_u$	1.59		0.12		<0.001	
	$\hat{\sigma}_v$	0.21		0.04		<0.001	
	$\widehat{cov}(u_i, v_i)$	0.26		0.04		<0.001	



**Table 8d**

Estimated parameters of the two-part model for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-12)

		Part 1			Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	0.13	0.08	0.11	-0.02	0.03	0.55
Age Group	1.5 -3y (Reference)						
	4-10y	0.07	0.27	0.78	0.15	0.10	0.12
	11-18y	-0.64	0.27	0.02	0.20	0.10	0.046
	19-64y	-0.23	0.24	0.34	0.34	0.08	<0.001
	65y and older	0.86	0.26	<0.001	0.56	0.09	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.52	0.13	<0.001	-0.15	0.05	0.003
	Intermediate occupations	-0.93	0.18	<0.001	-0.16	0.07	0.02
	Small employers & own account workers	-0.36	0.16	0.02	-0.13	0.06	0.04
	Lower supervisory & technical occupations	-0.80	0.16	<0.001	-0.22	0.06	<0.001
	Semi-routine occupations	-1.16	0.15	<0.001	-0.23	0.06	<0.001
	Routine occupations	-1.18	0.16	<0.001	-0.29	0.06	<0.001
	Never worked	-1.01	0.29	<0.001	-0.13	0.12	0.28
	Other	-0.39	0.31	0.21	-0.13	0.11	0.25
$\hat{k}$ , GG distribution shape parameter					0.60	0.07	<0.001
$\hat{\sigma}$ , GG distribution scale parameter					0.22	0.05	<0.001
		Estimates		SE		p-value	
Variance components	$\hat{\sigma}_u$	2.22		0.13		<0.001	
	$\hat{\sigma}_v$	0.38		0.02		<0.001	
	$\widehat{cov}(u_i, v_i)$	0.55		0.04		<0.001	

**Table 8e**

Estimated parameters of the two-part model for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Part 1			Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	0.58	0.09	<0.001	-0.02	0.03	0.41
Age Group	1.5 -3y (Reference)						
	4-10y	0.15	0.30	0.62	0.30	0.09	<0.001
	11-18y	-0.46	0.28	0.10	0.36	0.09	<0.001
	19-64y	0.49	0.26	0.06	0.79	0.08	<0.001
	65y and older	1.15	0.28	<0.001	0.76	0.09	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.23	0.15	0.12	-0.08	0.05	0.06
	Intermediate occupations	-0.85	0.19	<0.001	-0.10	0.06	0.09
	Small employers & own account workers	-0.17	0.18	0.36	-0.03	0.06	0.56
	Lower supervisory & technical occupations	-0.67	0.18	<0.001	-0.09	0.06	0.06
	Semi-routine occupations	-1.02	0.17	<0.001	-0.15	0.06	0.002
	Routine occupations	-1.01	0.17	<0.001	-0.19	0.06	<0.001
	Never worked	-0.99	0.31	0.001	-0.11	0.11	0.26
	Other	-0.38	0.35	0.28	-0.11	0.11	0.27
$\hat{k}$ , GG distribution shape parameter					0.91	0.10	<0.001
$\hat{\sigma}$ , GG distribution scale parameter					0.25	0.05	<0.001
		Estimates		SE		p-value	
Variance components	$\hat{\sigma}_u$	1.66		(0.15		<0.001	
	$\hat{\sigma}_v$	0.23		0.02		<0.001	
	$\widehat{cov}(u_i, v_i)$	0.35		0.04		<0.001	

### 3.6.1 Two-part model comparator analysis

To place the results in context is somewhat difficult as this chapter addresses current gaps in the analysis of episodically consumed foods and, as discussed in Section 1.12.1, estimation of episodically consumed foods is not currently carried out in the NDNS RP. Therefore to demonstrate the two-part model, introduced here, the amount of iron consumed in the various food groups is compared to estimated intakes from a regression model estimating mean intakes of iron from the same food groups whilst adjusting for the NDNS RP survey design, as is currently done in the NDNS RP report, but not considering the skewed distribution of intakes and not considering the within-person variation arising from the repeated measurements of the NDNS RP participants. Both models adjust for sex, age group and NSSEC as in the two part model given as

$$\mu_i = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{AgeGroup}_i + \beta_3 \text{NSSEC}_i \quad (18)$$

where  $\mu_i$  is the mean amount of iron consumed from the food group by participant  $i$ , sex is a categorical variable with values: males and females, age group is a categorical variable with categories: 1.5-3y; 4-10y; 11-18y; 19-64y; and 65y and older and NSSEC is a categorical variable containing the groups: higher managerial and professional occupations; lower managerial and professional occupations; intermediate occupations; small employers and own account workers; semi-routine occupations; routine occupations; never worked; and other. The limitations of the survey weighted regression approach are that it does not deal with the high frequency of zeros and it fails to separate the within- and between-person variation. This means that the variance component of the model is incorrect which can lead to biased results.

**Table 9a** compares the mean and standard error of iron consumed from breakfast cereals estimated by survey weighted regression and the two-part model and highlights that statistically significant differences may be missed. For example the iron intake in those aged 11-18y was not significantly higher than the reference group in the survey weighted regression model whereas it was found to be significantly higher in the two-part model (Survey weighted regression: 0.15, 95% CI:-0.05, 0.35, p-value=0.15

vs Two-part model: 0.84, 95% CI:0.66, 1.02, p-value<0.001). Moreover results from the survey weighted regression showed a statistically significant association between iron intake from breakfast cereals in those aged 19-64y compared to the reference group though this was a different direction of association to the two-part model (Survey weighted regression: -0.32, 95% CI:-0.50, -0.14, p-value<0.001 vs Two-part model: 0.37, 95% CI:0.21, 0.53, p-value<0.001).

When the amount of iron consumed from bread, as estimated by the two-part model, was compared with estimates using the survey weighted regression model (**Table 9b**) four of the NSSEC groups no longer showed a statistically significant difference compared to the reference group. These were the lower managerial & professional occupations group (Survey weighted regression: -0.01, 95% CI:-0.13, 0.11, p-value=0.90 vs Two-part model: -0.07, 95% CI:-0.13, -0.01, p-value=0.01); the intermediate occupations group (Survey weighted regression: -0.08, 95% CI:-0.26, 0.10, p-value=0.37 vs Two-part model: -0.13, 95% CI:-0.21, -0.05, p-value <0.001); the small employers & own account workers group (Survey weighted regression: 0.04, 95% CI:-0.12, 0.20, p-value=0.62 vs Two-part model: -0.07, 95% CI:-0.13, -0.01, p-value=0.01) and those in semi-routine occupations group (Survey weighted regression: -0.04, 95% CI:-0.18, 0.10, p-value=0.54 vs Two-part model: -0.09, 95% CI:-0.15, -0.03, p-value=0.01);

Estimates by survey weighted regression for iron intake from vegetables were found to be statistically significantly different from the reference group though not in estimates by the two-part model. These were the sex category with females different from the male reference group (Survey weighted regression: -0.04, 95% CI:-0.12, 0.04, p-value=0.26 vs Two-part model: -0.08, 95% CI:-0.14, -0.02, p-value=0.01) and five NSSEC groups different from the reference group higher managerial & professional occupations (**Table 9c**). These were the lower managerial & professional occupations group (Survey weighted regression: -0.17, 95% CI:-0.29, -0.05, p-value=0.006 vs Two-part model: -0.07, 95% CI:-0.17, 0.03, p-value=0.17); the intermediate occupations group (Survey weighted regression: -0.29, 95% CI:-0.45, -0.13, p-value<0.001 vs Two-part model: -0.10, 95% CI:-0.24, 0.04, p-value=0.14); the lower supervisory & technical occupations group (Survey weighted regression: -0.24, 95% CI:-0.40, -0.08,

p-value=0.001 vs Two-part model: -0.10, 95% CI:-0.22, 0.02, p-value=0.09); the semi-routine occupations group (Survey weighted regression: -0.27, 95% CI:-0.43, -0.11, p-value=0.001 vs Two-part model: -0.09, 95% CI:-0.21, 0.03, p-value=0.15) and those in the never worked category (Survey weighted regression: -0.32, 95% CI:-0.54, -0.10, p-value=0.006 vs Two-part model: -0.12, 95% CI:-0.34, 0.10, p-value=0.29).

Comparing the amount of iron consumed from fruit in the two-part model with the survey regression (**Table 9d**) showed three cases where statistically significant differences were found relative to the reference group in the survey weighted regression estimates that were not observed in the two-part model estimates. These were for females (Survey weighted regression: 0.04, 95% CI:0.02, 0.06, p-value=0.005 vs Two-part model: -0.02, 95% CI:-0.08, 0.04, p-value=0.55); for those in the age group 4-10y (Survey weighted regression: -0.06, 95% CI:-0.10, -0.02, p-value=0.003 vs Two-part model: 0.15, 95% CI:-0.05, 0.35, p-value=0.12) and those in the never worked NSSEC group (Survey weighted regression: -0.10, 95% CI:-0.18, -0.02, p-value=0.02 vs Two-part model: -0.13, 95% CI:-0.37, 0.11, p-value=0.28). In addition the direction of the difference in the amount of iron consumed changed in age groups 11-18y (Survey weighted regression: -0.18, 95% CI:-0.22, -0.14, p-value<0.001 vs Two-part model: 0.20, 95% CI:0, 0.40, p-value=0.046) relative to the reference group 1.5-3y.

Estimates for the amount of iron consumed from fruit and vegetables showed some differences between those for the two-part model and those from the survey regression model (**Table 9e**). The iron intake from fruit and vegetables for those in the age group 11-18y was significantly higher than the reference group and this was not seen in the comparator analysis (Survey weighted regression: 0.05, 95% CI:-0.03, 0.13, p-value=0.25 vs Two-part model: 0.36, 95% CI:0.18, 0.54, p-value<0.001). Furthermore four NSSEC groups showed statistically significant differences compared to the reference group in the model estimates from the survey weighted regression model that did not occur in the two-part model. These were the lower managerial and professional occupations group (Survey weighted regression: -0.24, 95% CI:-0.38, -0.10, p-value=0.001 vs Two-part model: -0.08, 95% CI:-0.18, 0.02, p-value=0.06); the intermediate occupations group (Survey weighted regression: -0.38, 95% CI:-0.56, -

0.20, p-value<0.001 vs Two-part model: -0.10, 95% CI:-0.22, 0.02, p-value=0.09); the lower supervisory and own account workers group (Survey weighted regression: -0.39, 95% CI:-0.57, -0.21, p-value<0.001 vs Two-part model: -0.09, 95% CI:-0.21, 0.03, p-value=0.06) and those in the never worked category (Survey weighted regression: -0.42, 95% CI:-0.67, -0.17, p-value=0.002 vs Two-part model: -0.11, 95% CI:-0.33, 0.11, p-value=0.26).

**Table 9a**

Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Survey weighted model			Two-part model: Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	-0.26	0.07	<0.001	-0.11	0.03	<0.001
Age Group	1.5-3y (Reference)						
	4-10y	0.54	0.10	<0.001	0.48	0.09	<0.001
	11-18y	0.15	0.10	0.15	0.84	0.09	<0.001
	19-64y	-0.32	0.09	<0.001	0.37	0.08	<0.001
	65y and older	-0.18	0.11	0.09	-0.002	0.08	0.98
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.04	0.11	0.72	0.03	0.05	0.48
	Intermediate occupations	0.07	0.15	0.63	0.01	0.07	0.87
	Small employers & own account workers	-0.22	0.14	0.12	-0.01	0.06	0.88
	Lower supervisory & technical occupations	-0.08	0.13	0.55	-0.01	0.06	0.91
	Semi-routine occupations	-0.10	0.13	0.42	0.01	0.06	0.80
	Routine occupations	-0.24	0.13	0.06	0.01	0.06	0.90
	Never worked	-0.28	0.16	0.07	-0.002	0.11	0.99
	Other	-0.23	0.23	0.31	-0.002	0.12	0.99

**Table 9b**

Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Survey weighted model			Two-part model: Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	-0.46	0.04	<0.001	-0.24	0.02	<0.001
Age Group	1.5-3y (Reference)						
	4-10y	0.46	0.05	<0.001	0.32	0.06	<0.001
	11-18y	0.56	0.05	<0.001	0.49	0.06	<0.001
	19-64y	0.78	0.05	<0.001	0.58	0.05	<0.001
	65y and older	0.71	0.06	<0.001	0.44	0.05	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.01	0.06	0.90	-0.07	0.03	0.01
	Intermediate occupations	-0.08	0.09	0.37	-0.13	0.04	<0.001
	Small employers & own account workers	0.04	0.08	0.62	-0.07	0.03	0.03
	Lower supervisory & technical occupations	0.06	0.09	0.52	-0.04	0.04	0.26
	Semi-routine occupations	-0.04	0.07	0.54	-0.09	0.03	0.01
	Routine occupations	0.03	0.09	0.77	-0.05	0.03	0.19
	Never worked	-0.12	0.12	0.34	-0.04	0.06	0.55
Other	0.03	0.14	0.84	0.04	0.07	0.54	



**Table 9c**

Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Survey weighted model			Two-part model: Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	-0.04	0.04	0.26	-0.08	0.03	0.01
Age Group	1.5-3y (Reference)						
	4-10y	0.19	0.04	<0.001	0.28	0.10	0.01
	11-18y	0.23	0.04	<0.001	0.43	0.10	<0.001
	19-64y	0.73	0.04	<0.001	0.79	0.09	<0.001
	65y and older	0.64	0.04	<0.001	0.69	0.10	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.17	0.06	0.006	-0.07	0.05	0.17
	Intermediate occupations	-0.29	0.08	<0.001	-0.10	0.07	0.14
	Small employers & own account workers	-0.12	0.08	0.11	-0.05	0.06	0.42
	Lower supervisory & technical occupations	-0.24	0.08	0.001	-0.10	0.06	0.09
	Semi-routine occupations	-0.27	0.08	0.001	-0.09	0.06	0.15
	Routine occupations	-0.35	0.08	<0.001	-0.15	0.06	0.01
	Never worked	-0.32	0.11	0.006	-0.12	0.11	0.29
Other	-0.25	0.13	0.06	-0.12	0.12	0.32	

**Table 9d**

Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Survey weighted model			Two-part model: Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	0.04	0.01	0.005	-0.02	0.03	0.55
Age Group	1.5-3y (Reference)						
	4-10y	-0.06	0.02	0.003	0.15	0.10	0.12
	11-18y	-0.18	0.02	<0.001	0.20	0.10	0.046
	19-64y	-0.06	0.02	0.001	0.34	0.08	<0.001
	65y and older	0.11	0.33	<0.001	0.56	0.09	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.07	0.03	0.01	-0.15	0.05	0.003
	Intermediate occupations	-0.10	0.03	0.002	-0.16	0.07	0.02
	Small employers & own account workers	-0.09	0.03	0.003	-0.13	0.06	0.04
	Lower supervisory & technical occupations	-0.14	0.03	<0.001	-0.22	0.06	<0.001
	Semi-routine occupations	-0.15	0.03	<0.001	-0.23	0.06	<0.001
	Routine occupations	-0.20	0.03	<0.001	-0.29	0.06	<0.001
	Never worked	-0.10	0.04	0.02	-0.13	0.12	0.28
	Other	-0.01	0.06	0.81	-0.13	0.11	0.25

**Table 9e**

Estimated parameters of a survey weighted regression model and from part 2 (Amount) of a two-part model, for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012).

		Survey weighted model			Two-part model: Part 2		
		Estimates	SE	p-value	Estimates	SE	p-value
Sex	Males (Reference)						
	Females	0.00	0.04	0.95	-0.02	0.03	0.41
Age Group	1.5-3y (Reference)						
	4-10y	0.13	0.04	0.001	0.30	0.09	<0.001
	11-18y	0.05	0.04	0.25	0.36	0.09	<0.001
	19-64y	0.67	0.04	<0.001	0.79	0.08	<0.001
	65y and older	0.75	0.06	<0.001	0.76	0.09	<0.001
NS-SEC	Higher managerial & professional occupations (Reference)						
	Lower managerial & professional occupations	-0.24	0.07	0.001	-0.08	0.05	0.06
	Intermediate occupations	-0.38	0.09	<0.001	-0.10	0.06	0.09
	Small employers & own account workers	-0.21	0.09	0.015	-0.03	0.06	0.56
	Lower supervisory & technical occupations	-0.39	0.09	<0.001	-0.09	0.06	0.06
	Semi-routine occupations	-0.42	0.09	<0.001	-0.15	0.06	0.002
	Routine occupations	-0.55	0.09	<0.001	-0.19	0.06	<0.001
	Never worked	-0.42	0.13	0.002	-0.11	0.11	0.26
	Other	-0.27	0.17	0.11	-0.11	0.11	0.27

### 3.6.2 Number of bootstrap samples for SE estimation

**Tables 10a** and **10b** display the point estimates of regression coefficients for the models for iron intake from vegetables along with standard errors taken from the average of 50, 100, 200 and 300 bootstrap samples to 5 decimal places and **Tables 11a** and **b** display the corresponding percentage differences for the standard errors from models estimating iron intake from vegetables between 50 bootstrap samples and 100, 200 and 300 bootstrap samples respectively. It can be seen from these tables that the standard errors change very little when the number of bootstrap samples is increased. Across both Tables, 11a and b, an increase in the percentage difference as the number of bootstrap samples increases would suggest that a higher number of bootstrap samples are required, though there doesn't appear to be an overall trend in this direction observed, and whilst the percentage difference for some estimates does increase as the number of bootstrap samples increases, for example across age groups, there are a similar number of cases where the percentage difference decreases as the number of bootstrap samples increases, indicating that standard errors given as a average of 300 bootstrap samples are similar to those from 50 bootstrap samples. Moreover the percentage differences tend to be around 1% in Table 11b with smaller differences of around 0.3% seen in Table 11a when comparing standard errors estimated from 300 bootstrap samples with those estimated from 50 bootstrap samples. Similar results are found for the other food groups (**Tables 24e** and **f** in **Appendix D**) and therefore for the sake of computational brevity 50 bootstrap samples are recommended and were used in Tables 8a-9e.

**Table 10a**

Estimated regression parameters of the two-part model for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	0.40	0.07606	0.07595	0.07592	0.07598
Age Group	1.5 -3y (Reference)					
	4-10y	0.01	0.24245	0.24042	0.24117	0.24123
	11-18y	-0.51	0.23309	0.23120	0.23201	0.23209
	19-64y	0.41	0.21071	0.20866	0.20953	0.20960
	65y and older	1.00	0.22857	0.22685	0.22745	0.22759
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.40	0.12629	0.12590	0.12578	0.12590
	Intermediate occupations	-0.89	0.16396	0.16406	0.16407	0.16414
	Small employers & own account workers	-0.19	0.15656	0.15596	0.15596	0.15622
	Lower supervisory & technical occupations	-0.69	0.15578	0.15565	0.15562	0.15575
	Semi-routine occupations	-1.10	0.14235	0.14212	0.14197	0.14207
	Routine occupations	-1.03	0.14988	0.14923	0.14933	0.14942
	Never worked	-1.00	0.26341	0.26227	0.26201	0.26292
	Other	-0.40	0.29787	0.29532	0.29718	0.29726

**Table 10b**

Estimated regression parameters of the two-part model for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	-0.08	0.03119	0.03154	0.03146	0.03150
Age Group	1.5 -3y (Reference)					
	4-10y	0.28	0.10342	0.10484	0.10488	0.10499
	11-18y	0.43	0.10139	0.10269	0.10281	0.10290
	19-64y	0.79	0.08977	0.09106	0.09118	0.09129
	65y and older	0.69	0.09540	0.09678	0.09681	0.09695
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.07	0.04901	0.04960	0.04944	0.04952
	Intermediate occupations	-0.10	0.06696	0.06778	0.06764	0.06764
	Small employers & own account workers	-0.05	0.06020	0.06091	0.06078	0.06092
	Lower supervisory & technical occupations	-0.10	0.06262	0.06337	0.06320	0.06330
	Semi-routine occupations	-0.09	0.05848	0.05928	0.05908	0.05914
	Routine occupations	-0.15	0.06211	0.06274	0.06262	0.06269
	Never worked	-0.12	0.11226	0.11335	0.11313	0.11349
	Other	-0.12	0.11787	0.11983	0.12030	0.12033
$\hat{k}$ , GG distribution shape parameter		0.76	0.09302	0.09113	0.09179	0.09154
$\hat{\sigma}$ , GG distribution scale parameter		0.30	0.04902	0.04820	0.04805	0.04836
Variance components	$\hat{\sigma}_u$	1.59	0.12002	0.11862	0.11882	0.11914
	$\hat{\sigma}_v$	0.21	0.03890	0.04050	0.03920	0.03916
	$\widehat{cov}(u_i, v_i)$	0.26	0.04462	0.04507	0.04465	0.04459

**Table 11a**

Percentage difference between standard error estimates for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
		SE (50)	SE (100)	SE (200)	SE (300)	
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	0.40	0.07606	0.14	0.18	0.11
Age Group	1.5 -3y (Reference)					
	4-10y	0.01	0.24245	0.84	0.53	0.50
	11-18y	-0.51	0.23309	0.81	0.46	0.43
	19-64y	0.41	0.21071	0.98	0.56	0.53
	65y and older	1.00	0.22857	0.76	0.49	0.43
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.40	0.12629	0.31	0.40	0.31
	Intermediate occupations	-0.89	0.16396	-0.06	-0.07	-0.11
	Small employers & own account workers	-0.19	0.15656	0.38	0.38	0.22
	Lower supervisory & technical occupations	-0.69	0.15578	0.08	0.10	0.02
	Semi-routine occupations	-1.10	0.14235	0.16	0.27	0.20
	Routine occupations	-1.03	0.14988	0.43	0.37	0.31
	Never worked	-1.00	0.26341	0.43	0.53	0.19
Other	-0.40	0.29787	0.86	0.23	0.20	

**Table 11b**

Percentage difference between standard error estimates for iron intake from vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	-0.08	0.03119	-1.12	-0.86	-0.99
Age Group	1.5 -3y (Reference)					
	4-10y	0.28	0.10342	-1.36	-1.40	-1.51
	11-18y	0.43	0.10139	-1.27	-1.39	-1.48
	19-64y	0.79	0.08977	-1.43	-1.56	-1.68
	65y and older	0.69	0.09540	-1.44	-1.47	-1.61
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.07	0.04901	-1.20	-0.87	-1.04
	Intermediate occupations	-0.10	0.06696	-1.22	-1.01	-1.01
	Small employers & own account workers	-0.05	0.06020	-1.17	-0.96	-1.19
	Lower supervisory & technical occupations	-0.10	0.06262	-1.19	-0.92	-1.08
	Semi-routine occupations	-0.09	0.05848	-1.36	-1.02	-0.01
	Routine occupations	-0.15	0.06211	-1.01	-0.82	-0.93
	Never worked	-0.12	0.11226	-0.97	-0.77	-1.09
	Other	-0.12	0.11787	-1.65	-2.04	-2.07
$\hat{k}$ , GG distribution shape parameter		0.76	0.09302	2.05	1.33	1.60
$\hat{\sigma}$ , GG distribution scale parameter		0.30	0.04902	1.69	2.00	1.36
Variance Components	$\hat{\sigma}_u$	1.59	0.12002	1.17	1.00	0.74
	$\hat{\sigma}_v$	0.21	0.03890	-4.03	-0.77	-0.67
	$\widehat{cov}(u_i, v_i)$	0.26	0.04462	-1.00	-0.07	0.07



### 3.7 Discussion

This chapter presented a two-part mixed-effects model capable of estimating the variability arising from semi-continuous data sampled under a complex sampling design applied to iron intake from selected food groups. This showed that females differed from males in both the amount of iron consumed, and the probability of consuming iron from bread, breakfast cereals, vegetables, and fruit and vegetables (Tables 8a-e). Iron intakes from these food groups were expected to be higher in males than females due to higher food intakes in general. Differences occurred between the reference age group, 1.5-3y, and the older age groups. In most cases larger amounts of iron from the selected food groups were consumed although the propensity to consume the foods was not always found to be higher. Iron intake from breakfast cereals was lower in the 19-64y age group along with iron from fruit and from vegetables in those aged 11-18y. Increasing iron intake through advice to increase meat and fish consumption has been shown to be ineffective at its goal in infants (Penny et al., 2005) whereas breakfast cereals are consumed by infants and increasing iron intake through increased breakfast cereal consumption may be easier to achieve.

The NSSEC categories represent a gradient where typically those in a higher group have greater material wealth than those of a relatively lower group, here almost all NSSEC groups had a lower propensity to consume and lower consumption amounts compared to those in the highest NSSEC group: higher managerial and professional occupations (Tables 8b-e). In the case of iron from fruit this tended to decrease in a linear fashion. Similar findings have been seen in the EPIC-Norfolk cohort examining over 22,000 men and women, showing that fruit and vegetable intake was lower in the manual social class vs the non-manual and those living in relatively more deprived areas (Shohaimi et al., 2004). Individuals living in more deprived areas, and this relates to those in lower NSSEC classes who have lower material wealth, have been found to be less aware of the importance of fruit and vegetable intake and are less likely or willing to act upon advice to change (McIntosh et al., 1990).

Here individual foods were combined into broad groups which has the advantage of allowing for a simple interpretation of intake between groups, though it may fail to describe the variation in the type of foods within each group. For example, there is evidence to suggest that vegetable intake varies by socio-economic position with those in higher social classes consuming a greater variety of fruits and vegetables than individuals in lower classes (Darmon and Drewnowski, 2008) and may therefore consume a greater variety of micronutrients. Similarly the proportion of breakfast cereals coming from the "high fibre breakfast cereals" category is lower in age groups 4-10y and 11-18y compared to older age groups (Bates et al., 2014b) and these typically have lower levels of added sugars. While the regression coefficients presented here compared overall consumption of the food groups across socioeconomic groups, it is not possible to disentangle qualitative differences in the choice of single foods that form each group.

The food groups were chosen, in part, as a demonstration of modelling semi-continuous data with a high proportion of zero values but mainly as they were found to be the main sources of iron currently consumed in the NDNS RP and that increasing intake of these foods is in line with dietary advice. The *Eatwell Plate* (NHS Choices, 2011), for example, suggests a third of food intake should come from starchy carbohydrates including bread and breakfast cereals and a third from fruit and vegetables. Using the findings presented here recommendations can be made to groups to increase their iron intake by increasing intakes of recommended foods.

The food groups breakfast cereals and bread are high in iron as flour is fortified with iron. However iron fortification is not without controversy as it can lead to iron overload in at-risk individuals, there has been some evidence to suggest associations with increased risk of diabetes (Forouhi et al., 2007), cancer (Huang, 2003) and cardiovascular disease (Danesh and Appleby, 1999). Furthermore in Denmark, ending the iron fortification of flour appeared to have little impact upon population iron levels (Osler et al., 1999) though this is not seen everywhere (Sadighi et al., 2008; Huang et al., 2009), nevertheless in the UK iron fortification contributes a sizeable proportion of iron to the diet and increasing the range of foods fortified with iron could see a reduction in iron deficiency.

To compare the effectiveness of the introduced methods, a second piece of analysis was presented using the amount part (part 2) of the novel methods presented here. This is based on the generalised gamma distribution and includes a random intercept and accounts for the survey weighting and design. Estimates given by the amount model were presented alongside those using methods currently used in the NDNS RP, namely a survey weighted regression model that does not consider measurement error nor the skewed intake distribution. The magnitude and statistical significance of the regression coefficients for sex, age and NSSEC groups differed between the novel amount part model and the traditional survey weighted regression method. This includes the finding of the amount part model that females consumed significantly less iron from vegetables when compared to males that was not found in the survey weighted regression model and that females consumed significantly more iron from fruit compared to males as estimated by the survey weighted regression, a finding that was not repeated in the amount part model. Similarly, differences in the amount of iron between NSSEC groups estimated by the two methods were found. For example, with the exception of those in the 'other' group, all NSSEC groups were found to consume significantly lower amounts of iron from fruit and vegetables compared to those in the higher managerial and professional group as estimated by the survey weighted regression model whereas only those in the semi- and routine occupations were found to consume lower amounts of iron from fruit and vegetables compared to the reference group and that the absolute amount of iron estimated by the survey weighted regression model was approximately twice as low as the estimated amounts given by the amount part model. As a result, associations indicated by the incorrectly specified survey weighted model may be spurious and should be interpreted with care.

## **4 Quantile regression of dietary intake in complex sample surveys**

As detailed in the aims of this thesis (Section 1.15), of interest currently is the modelling of usual intake of the diet. There are different challenges in estimating the intake of foods and nutrients. This chapter serves to complement Chapter 3 that presented methods for modelling the mean intake of foods and nutrients, by presenting methods to examine the tails of intake distributions. By examining the proportion of the population with low intakes of a particular nutrient, governments and public health agencies are better informed to introduce or amend policies to reduce the number of people at risk of deficiency. This chapter extends a recently developed quantile regression approach for clustered data to include a complex sample design in the estimation. The approach is illustrated by describing patterns of dietary iron intake in the UK compared to reference values. The extension to include a complex sample design allows this model to be used to estimate intake collected from survey data, such as the NDNS RP collected using complex sampling methods and shows that, compared to weighted linear mixed-effects regression is better capable of modelling skewed distributions.

### **4.1 Introduction**

To monitor dietary intake a more complete characterisation of the distribution is required to identify those at the lower end of the distribution who are more likely to be at risk of deficiency and conversely those at the high end being at increased risk of toxicity. This can be achieved by directly modelling the conditional quantiles of the distribution of dietary intake given the explanatory variables through quantile regression (Koenker and Bassett, 1978; Koenker and Hallock, 2001). Whilst linear regression models are capable of providing estimates of particular quantiles through the inverse of the CDF, this will fail to provide a complete picture of relationships between dietary intake and explanatory variables, as these are not easy to include in the quantile estimates. In addition, outliers often present in dietary data can have a large influence

on mean estimates, in which case, the median becomes a more robust measure of location.

Quantile regression using data sampled with simple random sampling is well established, and there have been methods for confidence interval estimation in the complex sample case (Francisco and Fuller, 1991; Dubnicka, 2007; Román-Montoya et al., 2008; Dodd, 1999; Lee and Eltinge, 1999). Geraci (2013) provided a recent review of quantile regression of survey data highlighting a limited number of studies. These included using quantile regression to estimate Body Mass Index using survey data collected in the US, where the survey weight was included in local-linear kernel weights required to produce smoothed growth curves (Li et al., 2010). Also mentioned was an application mapping quantile of expenditure in Ecuador (Geraci and Salvati, 2007). The multistage sampling and sampling weightings have different impacts upon the model estimates. The sampling weights affect both the point estimates and variance estimates as they adjust the sample back to the population but the clustering and strata will have an impact only upon the variance estimation as they deal with changes due to similarities (through clustering) and differences (strata) (see Sections 1.7 and 2.6 and Table 5).

Demonstrated here is the use of a recently developed linear mixed-effects quantile regression model based on the asymmetric Laplace distribution (Geraci and Bottai, 2007; Geraci, 2014) to obtain a more accurate description of the distribution of dietary data collected using a complex sample design. Whilst the asymmetric Laplace distribution is used for maximum likelihood estimation of the parameters, due to the equivalence of the estimating equations for the regression parameters with those from non-parametric quantile regression (Koenker and Bassett, 1978; Koenker and Hallock, 2001), the method is robust to distributional assumptions. A pseudo likelihood approach is used to incorporate the sampling weightings in the estimation, and bootstrap methods are employed to estimate standard errors of the parameters estimates that account for clustering and stratification. The proposed approach can be easily implemented through use of the "survey" (Lumley, 2004, 2014) and "lqmm" (Geraci, 2014; Geraci and Bottai, 2014) R packages. The method is illustrated by describing current

patterns of iron intake in the UK population using dietary data from the UK NDNS RP (Public Health England, 2014).

## 4.2 The asymmetric Laplace distribution as working model in quantile regression with a random intercept

A continuous random variable  $Y \in \mathbb{R}$  has an asymmetric Laplace distribution (ALD)  $\text{ALD}(\mu, \sigma, p)$  (Yu and Zhang, 2005) with parameters  $-\infty < \mu < \infty$ ,  $\sigma > 0$  and  $0 < p < 1$ , if its probability density is

$$f(y; \mu, \sigma, p) = \frac{p(1-p)}{\sigma} \exp\left\{-\frac{1}{\sigma} \rho_p(y - \mu)\right\}, \quad (19)$$

where  $\mu$  is a location parameter,  $\sigma$  is a scale parameter and  $p$  is an asymmetry parameter; and

$$\rho_p(y - \mu) = \begin{cases} p(y - \mu), & \text{if } (y - \mu) \geq 0 \\ (p - 1)(y - \mu), & \text{if } (y - \mu) < 0. \end{cases}$$

For a fixed value of  $p$ , the parameter of interest is  $\mu$  as it defines the  $p$ th quantile of the distribution and  $\sigma$  is a nuisance parameter. To model the  $p$ th quantile, denoted by  $\mu^{(p)}$ , of dietary intake  $y_{ij}$  of individual  $i, i = 1, \dots, n$  at day  $j, j = 1, \dots, n_i$  for a given vector of explanatory variables  $\mathbf{X}_{ij}$  and regression coefficients  $\boldsymbol{\beta}^{(p)}$  we specify

$$\mu^{(p)}(\mathbf{X}'_{ij}, u_i) = \mathbf{X}'_{ij} \boldsymbol{\beta}^{(p)} + u_i \quad (20)$$

where  $u_i$  is a random intercept with assumed normal distribution with zero mean and variance  $\sigma_u$ .

The expectation is given by

$$E(y) = \mu + \sigma \frac{1 - 2p}{p(1 - p)} \quad (21)$$

and the variance is given by

$$\text{VAR}(y) = \frac{\sigma^2(1 - 2p + 2p^2)}{(1 - p)^2 p^2} \quad (22)$$

Under the assumption that  $y_{ij}|u_i$  are independent, maximum likelihood estimation based on the working distribution  $\text{ALD}(\mu^{(p)}, \sigma^{(p)}, p)$  yields robust estimates of  $\boldsymbol{\beta}^{(p)}$  with respect

to distributional assumptions (Geraci and Bottai, 2007; Geraci, 2014). This follows from the equivalence between maximum likelihood estimation of the regression parameters in quantile regression based on the ALD and estimation based on a minimisation problem analogous to least squares but using the function  $\rho_p$  as a loss function (Koenker and Bassett, 1978).

Inference is based on the marginal likelihood function which is defined as the integral of the joint probability density of  $(\mathbf{y}, u_i)$  with respect to  $u_i$ . The joint probability density of  $(\mathbf{y}, u_i)$  is

$$p(\mathbf{y}, u_i; \boldsymbol{\beta}^{(p)}, \sigma, \sigma_u) = \prod_{i=1}^n p(\mathbf{y}_i, u_i; \boldsymbol{\beta}^{(p)}) = \prod_{i=1}^n \prod_{j=1}^{n_i} f(y_{ij}|u_i, \mathbf{X}_i; \boldsymbol{\beta}^{(p)}, \sigma) p(u_i; \sigma_u), \quad (23)$$

where  $f(y_{ij}|u_i, \mathbf{X}_i; \boldsymbol{\beta}^{(p)}, \sigma)$  denotes the asymmetric Laplace density function with parameter  $p$  fixed, location  $\mu^{(p)}(\mathbf{X}'_{ij}, u_i)$  with  $\boldsymbol{\beta}^{(p)}$  the vector of quantile regression parameters, scale parameter  $\sigma$ ; and  $p(u_i; \sigma_u)$  is the density function of the random effect  $u_i$ , typically a normal density with mean zero and variance  $\sigma_u$ . The  $i$ th contribution of observations  $\mathbf{y}_i$  to the marginal likelihood function is obtained by integrating out the random intercept in the joint distribution of  $(\mathbf{y}_i, u_i)$ , that is

$$L_{Y_i}(\boldsymbol{\beta}^{(p)}, \sigma, \sigma_u) = \int p(\mathbf{y}_i, u_i; \boldsymbol{\beta}^{(p)}, \sigma, \sigma_u) du_i.$$

Gauss-Hermite quadrature uses a fixed set of  $K$  ordinates and weights  $(v_k, w_k)$ ,  $k = 1, \dots, K$  to approximate the integral with respect to  $u_i$  as

$$L_{Y_i}(\boldsymbol{\beta}^{(p)}, \sigma, \sigma_u) \approx \sum_{k=1}^K w_k \exp\{\log(p(\mathbf{y}_i, v_k; \boldsymbol{\beta}^{(p)}, \sigma, \sigma_u))\},$$

and the marginal likelihood is then simply calculated as

$$L_Y(\boldsymbol{\beta}^{(p)}, \sigma, \sigma_u) = \prod_{i=1}^n L_{Y_i}(\boldsymbol{\beta}^{(p)}, \sigma, \sigma_u).$$

Gaussian quadrature has been found to be computationally less intensive and more efficient than a Monte Carlo EM procedure (Geraci and Bottai, 2007) previously proposed by the same authors Geraci (2014).

### 4.3 Statistical inference taking the complex sample design into account

The quantile regression model can be extended to describe dietary data collected using a complex survey design that involves multistage sampling. Often, the target population is divided into  $L$  strata, and within each stratum  $l$ ,  $l = 1, \dots, L$ ,  $N_l$  primary sampling units (PSUs) are sampled under a given probability sampling design. Subsequently,  $N_{kl}$  individuals are selected from the  $k$ th PSU in stratum  $l$  using SRS. Following selection, sampling weights  $w_{ikl}$  are calculated for individual  $i$  in PSU  $k$  of stratum  $l$  to adjust for unequal selection probability, to compensate for non-response rate and to reflect known population characteristics. Each selected individual then provides several records of dietary intake, generating clustered data at the individual level. The survey design for the NDNS RP is described in detail in Section 2.4 though briefly, the data comprise dietary intake records  $y_{ijkl}$  taken at day  $j$ , ( $j = 1, \dots, N_{ikl}$ ) by individual  $i$ , ( $i = 1, \dots, N_{kl}$ ) selected from the  $k$ th PSU in stratum  $l$ , and a vector of explanatory variables  $\mathbf{X}_{ijkl}$ .

### 4.4 Pseudo likelihood estimation

A pseudo likelihood approach can be used to incorporate the sampling weights. Note that the weights discussed in the previous section do not vary within individual; therefore, the weights can readily be incorporated and re-scaling is not required, unlike other contexts whereby the sampling weights might change at different levels of sampling, such as a weighting that adjusts for day of the week and therefore would vary within the individual (Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006).

### 4.5 Model selection

Model selection can be undertaken using a pseudo likelihood ratio test. The analytical distribution of the PLRT statistic requires the computation of the weighted information matrix, which can be challenging to obtain. Alternatively, Aerts and Claeskens (1999)



have shown that the parametric bootstrap method provides a consistent estimator for the distribution of the pseudo likelihood ratio test statistics in similar settings to that presented here involving models for clustered data, this approach is adopted here. This is undertaken by: generating  $R$  samples (with the same cluster structure and size as the observed sample) from the fitted null model; for each  $r$ th sample, pseudo ML estimates of the parameters are obtained from the null and alternative models, and the PLRT statistic  $t_r^*$  from the  $r$ th sample is calculated. The significance of the PLRT is calculated by Equation 24:

$$P_{boot} = \frac{1 + \#\{t_r^* \geq t\}}{R + 1} \quad (24)$$

where  $t$  denotes the observed PLRT statistic.

## 4.6 Estimation of standard errors

Bootstrap resampling can be used for estimation of uncertainty of the maximum pseudo likelihood estimates, taking into account the complex survey design. To obtain bootstrap estimates of the covariance matrix of the maximum pseudo likelihood estimate  $\hat{\boldsymbol{\theta}}^{(p)} = (\hat{\boldsymbol{\beta}}^{(p)}, \hat{\sigma}^{(p)})$ ,  $B$  bootstrap replicates were generated by random sampling from each PSU in each stratum, from each replicate the pseudo likelihood parameter estimate  $\hat{\boldsymbol{\theta}}_b^{(p)}$  ( $b = 1, \dots, B$ ) is obtained. The bootstrap estimate of the covariance matrix  $\hat{\boldsymbol{\theta}}^{(p)}$  is given by

$$\text{COV}(\hat{\boldsymbol{\theta}}^{(p)}) = \frac{A}{B} \sum_{b=1}^B (\hat{\boldsymbol{\theta}}_b^{(p)} - \hat{\boldsymbol{\theta}}^{(p)*}) (\hat{\boldsymbol{\theta}}_b^{(p)} - \hat{\boldsymbol{\theta}}^{(p)*})^T$$

where  $\hat{\boldsymbol{\theta}}^{(p)*} = \frac{1}{B} \sum_{b=1}^B \hat{\boldsymbol{\theta}}_b^{(p)*}$  and  $A$  is a scaling factor defined by  $A = \frac{\bar{M}_l}{M_{l-1}}$ , where  $\bar{M}_l$  is the average number of PSUs per stratum. The weights for each bootstrap replicate need to be adjusted according to the sampling method (Canty and Davison, 1999). For example, if  $N_l$  PSUs are sampled in stratum  $l$  then the adjusted weights are  $w_{ikl}^b = w_{ikl} k_{kl}^b$ , where  $w_{ikl}$  is the original weight for the  $i$ th individual in the  $k$ th PSU of the  $l$ th stratum, and  $k_{kl}^b$  is the number of repeated samples from the  $k$ th PSU of the  $l$ th stratum, in the  $b$ th bootstrap replicate.

## 4.7 Modelling of dietary iron consumption

The purpose of the analysis was to describe current patterns of dietary iron consumption in the UK population by age and sex and to compare with corresponding age and sex specific LRNI values for iron. The following analysis is thus concerned with a nutrient: dietary iron, and although supplements containing iron were recorded, they were not included in the analysis. Individual daily iron intake was taken as the total iron (mg/day) consumed in foods and drinks only (i.e. excluding supplements). Of the initial 6127 participants between 2008 and 2012, aged 65y or less, 18 were removed from the analysis because they had a missing value for the NSSEC variable which gave a total of 24232 observations of which 1.8% recorded three days with the remaining recording 4 days of intake. Differences in dietary intake have been shown at the weekend (Haines et al., 2003), therefore a binary variable (1=weekday, 0=weekend) was created to distinguish week day consumption from weekend days defined as Saturday and Sunday. The survey design consists of 795 PSUs clustered in 388 strata with an average cluster size of 9 and a maximum of 19. There are two PSUs with a single participant (lonely PSUs) which were joined with the nearest similar cluster. The sample weightings used to adjust for selection probability and non response ranged from 0.02 to 10.23, had a median (IQR) of 0.50 (0.96) and a mean (sd) of 0.94 (1.14). The difference in median and mean values reflects the over sampling of children and participants from Northern Ireland, Wales and Scotland as their contributions are deflated so they reflect the population of the UK. The  $p$ th quantile of dietary intake  $y_{ijkl}$ , taken at day  $j$ , ( $j = 1, \dots, N_{ikl}$ ) by individual  $i$ , ( $i = 1, \dots, N_{ikl}$ ) selected from the  $k$ th PSU in stratum  $l$ , is expressed as in Equation 20 by

$$\mu_{ijkl}^{(p)} = \beta_0^{(p)} + \beta_1^{(p)} \text{Sex}_{ikl} + \beta_2^{(p)} \text{Age}_{ikl} + \beta_3^{(p)} \text{Age}_{ikl}^2 + \beta_4^{(p)} \text{Age}_{ikl}^3 + \beta_5^{(p)} \text{NSSEC}_{ikl} + \beta_6^{(p)} \text{Weekday}_{ijkl} + u_i, \quad (25)$$

where  $u_i$  is a random intercept with assumed normal distribution with zero mean and variance  $\sigma_u$ .

## 4.8 Results

**Table 12** presents the participants' weighted characteristics, including socio-economic status (NSSEC), age and proportion of diary records taken during the weekend (Saturday or Sunday) by sex. Although each participant recorded their intake for 4 days only, the survey was designed to ensure that a reasonable proportion of weekend days were captured, yielding at least one third of weekend observations on average. There is a good mix of socio-economic status in the sample, which was achieved by the survey stratification by region.

The 2.5<sup>th</sup>, 25<sup>th</sup>, median, 75<sup>th</sup> and 97.5<sup>th</sup> quantiles were modelled with sex, age, weekday and NSSEC as explanatory variables (Equation 25). The distribution of iron intake is described through a weighted histogram in **Figure 22**. As expected, this shows a skewed distribution partly due to a number of outliers, which are checked by nutritionists and deemed feasible, but are often present in survey dietary data. The figure also shows the asymmetric Laplace distribution fitted to these data. The fit appears to be reasonably good.

**Figure 23** provides an empirical graphical representation of all the available data, presenting the empirical median iron intake for each participant alongside the estimated quantile regression curves with 95% confidence bands by age. It can be seen that the estimated curves represent the patterns of intake well. The mean and mean  $\pm 2sd$  from a reference population were used to set thresholds of iron intake in a population, and represent the estimated average requirement (EAR), LRNI(mean-2sd) and reference nutrient intakes RNI (mean+2sd) respectively (Scientific Advisory Committee on Nutrition (SACN), 1991). These are reference values set by the Committee on Medical Aspects of Food and Nutrition Policy (COMA) and endorsed by SACN. Comparison of the 2.5<sup>th</sup> quantile with the LRNI recommendations according to sex and age can help to identify those who are at risk of dietary deficiencies (Figures 24 and 25). Inspection of the observed median intake by age suggested that the relationship of age could be non-linear (Figure 23). Quadratic and cubic terms of age were added to assess this. A bootstrap PLRT showed that a model including the cubic term fitted significantly better

**Table 12**

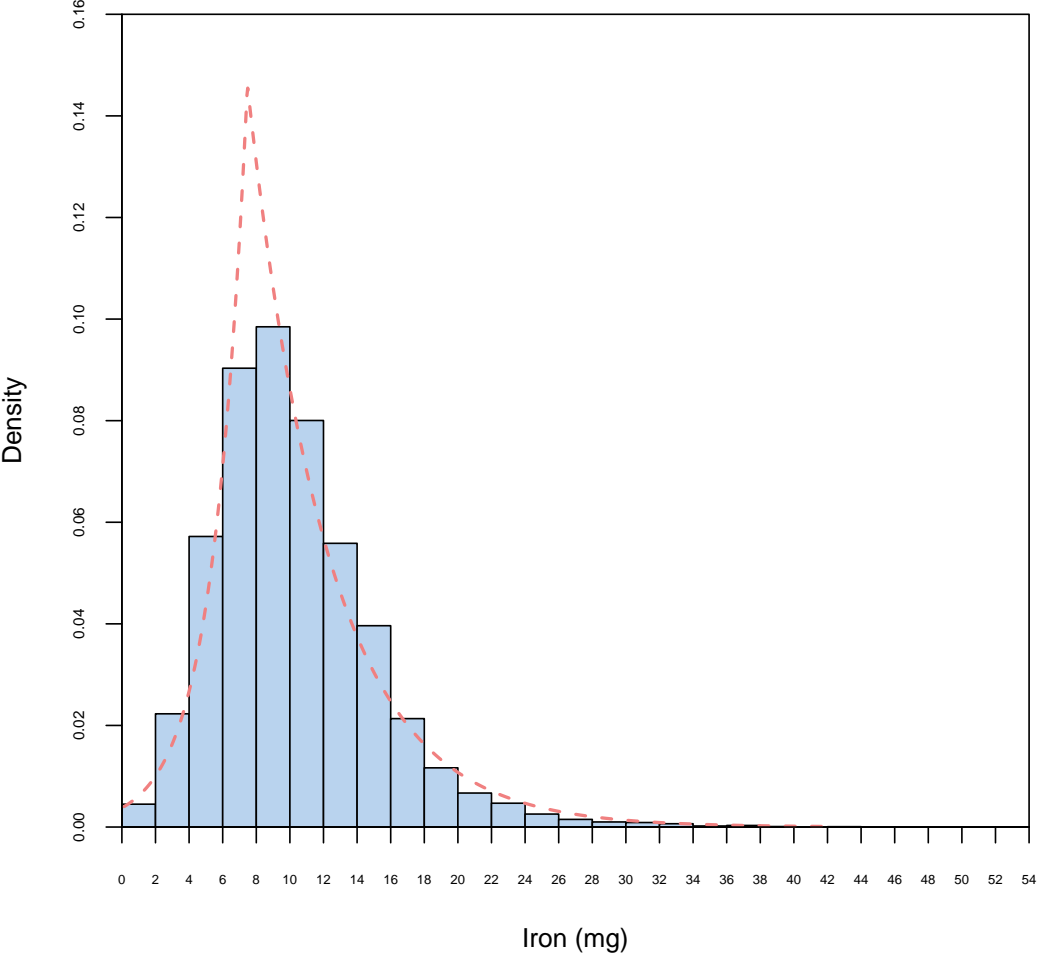
Weighted demographic characteristics of 6109 participants aged 65y and under in the NDNS Rolling Programme Y1-4 (2008-2012)

	Females	Males	All
Number of participants (%)	3249 (50.9)	2860 (49.1)	6109
Age (years), (mean, SD)	34.0 (18.1)	33.4 (18.1)	33.7 (18.1)
NS-SEC (%)			
Higher managerial and professional occupations	462 (14.2)	502(17.4)	971 (15.3)
Lower managerial and professional occupations	869 (26.7)	783 (26.5)	1653 (26.7)
Intermediate occupations	271 (8.3)	237 (7.9)	508 (8.5)
Small employers and own account workers	394 (12.1)	289 (10.0)	678 (11.4)
Lower supervisory and technical occupations	329 (10.1)	274 (10.7)	601 (10.2)
Semi-routine occupations	430 (13.2)	342 (12.5)	769 (12.7)
Routine occupations	329 (10.1)	329 (11.7)	660 (11.1)
Never worked	83 (2.6)	53 (1.7)	134 (2.2)
Other	83 (2.6)	52 (1.6)	134 (1.9)
Weekend consumption days (%)	31.7	31.9	31.8

NS-SEC is the National Statistics socio-economic classification.

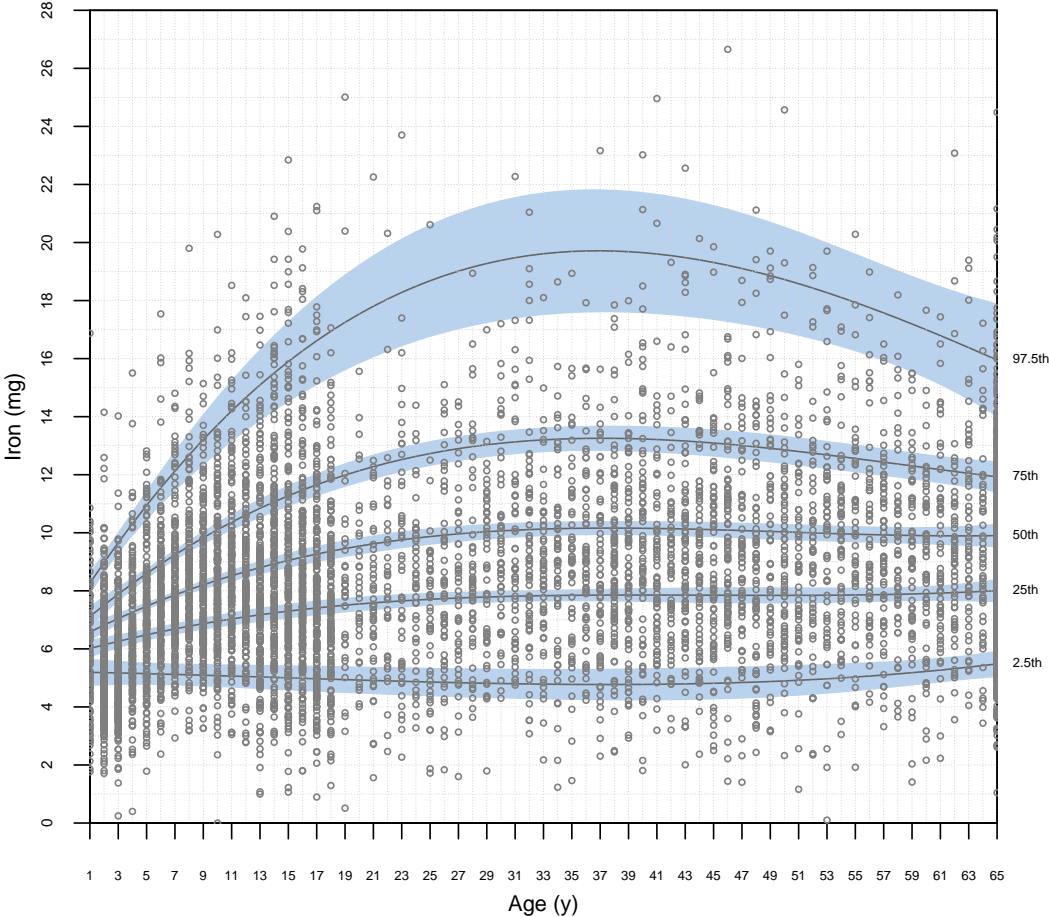
**Figure 22**

Weighted histogram of mean daily iron intake for 6109 participants aged 65 and under in the NDNS Y1-4 (2008-12) (shaded area) and fitted asymmetric Laplace distribution (dashed line).



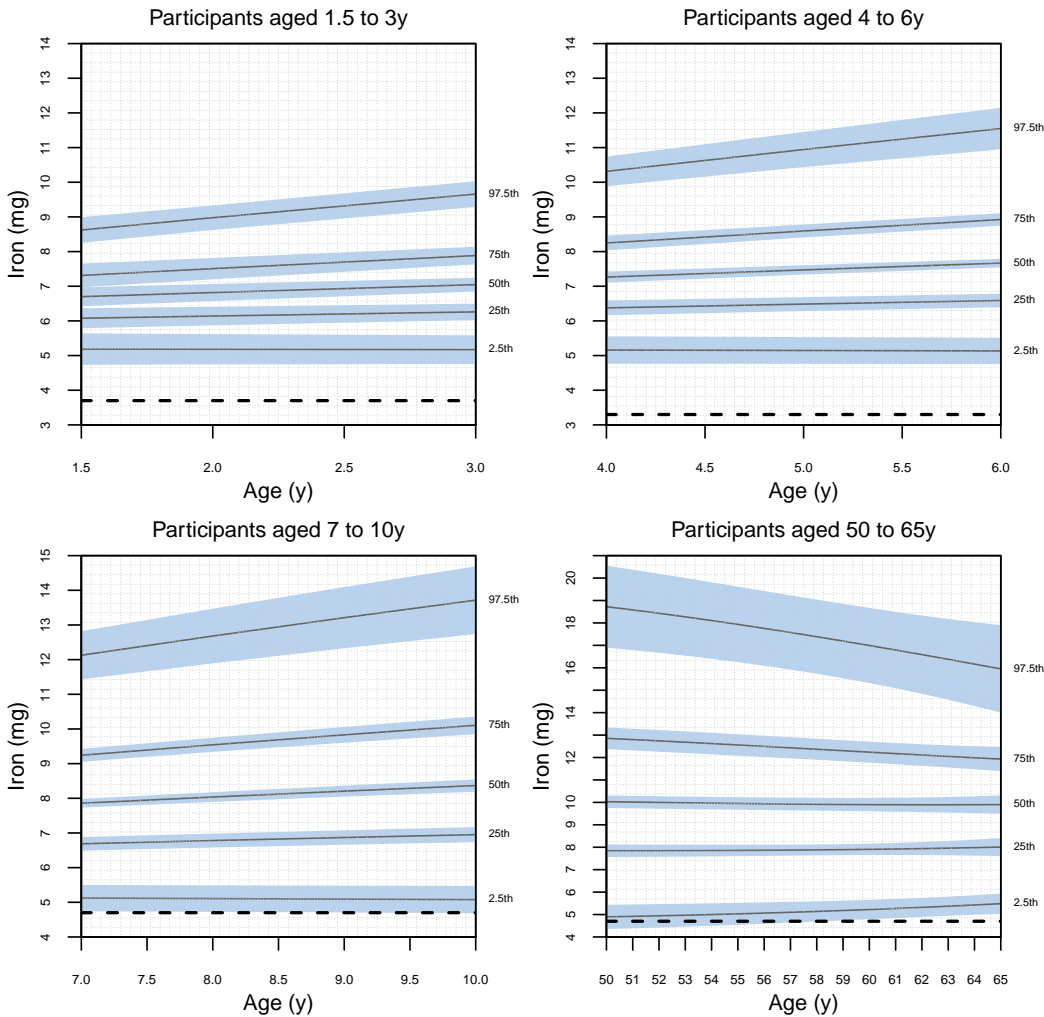
**Figure 23**

Estimated quantile iron intake with 95% confidence bands and observed median individual intake.



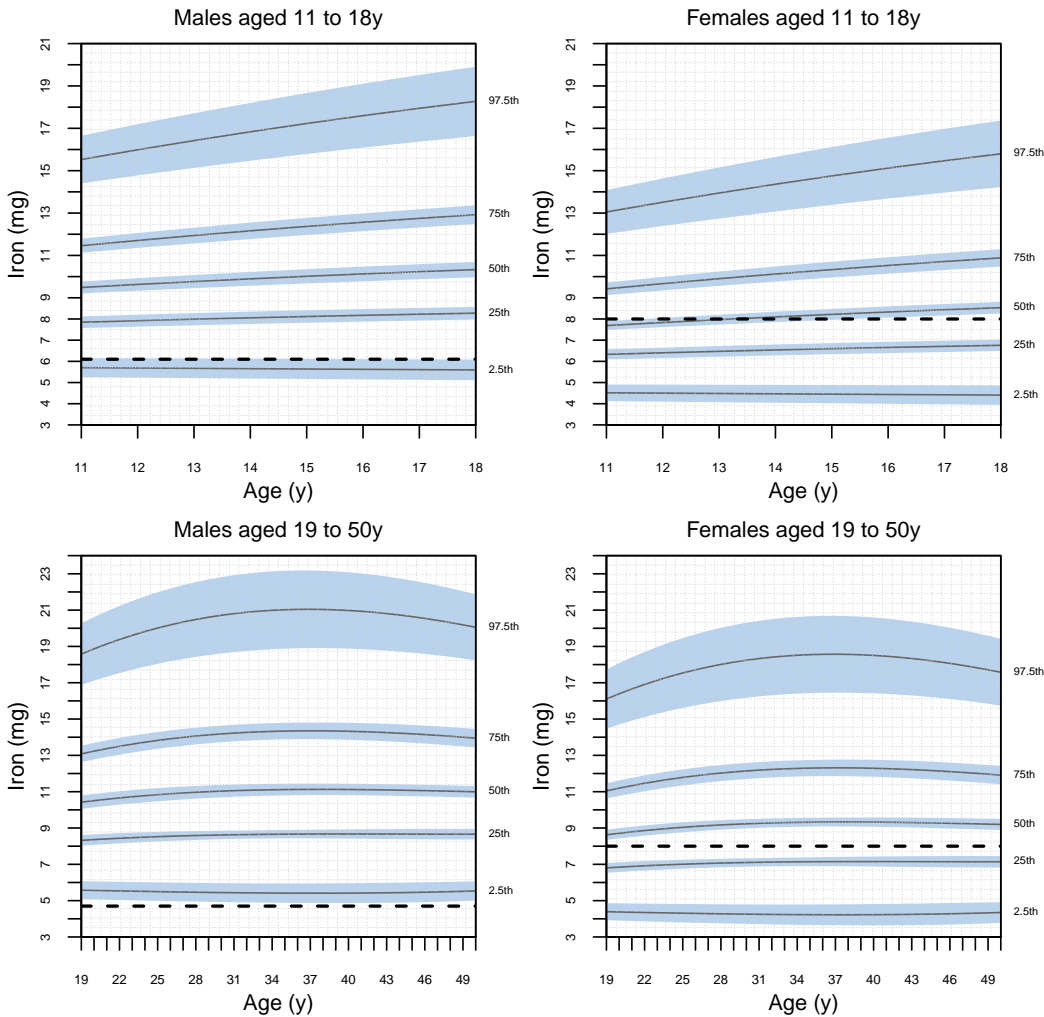
**Figure 24**

Estimated iron intake quantiles with 95% confidence bands by age groups with LRNI recommendations (broken lines).



**Figure 25**

Estimated iron intake quantiles with 95% confidence bands by age groups with LRNI recommendations (broken lines) that differ by sex.





that a model with linear and quadratic terms only (PLRT statistic= 0.00672, Degrees of freedom=1, p-value <0.001).

**Figures 24 & 25** show the 2.5<sup>th</sup>, 25<sup>th</sup>, median, 75<sup>th</sup> and 97.5<sup>th</sup> estimated quantiles of iron intake by each year of age including reference values. The LRNI is the same for males and females for ages 1-3y, 4-6y, 7-10y and 50-65y (Scientific Advisory Committee on Nutrition (SACN), 1991); in such cases plots of quantiles are presented for both sexes in Figure 24. Sex specific LRNI values are provided for age groups 11-18y and 19-50y separately in Figure 25. Estimates are given by the solid line and the shaded area indicates 95% confidence bands. The plots show the large proportion of females below the LRNI for iron with the dashed LRNI line going between the 25<sup>th</sup> and 50<sup>th</sup> quantiles in Females aged 19 to 50y and close to the 50<sup>th</sup> quantile throughout the entire range of 11 to 18y age group. In the majority of other groups the LRNI is close to, or below, the predicted 2.5<sup>th</sup> quantile.

**Table 13** presents the estimated regression coefficients and standard errors when modelling the 2.5<sup>th</sup>, 25<sup>th</sup>, median, 75<sup>th</sup> and 97.5<sup>th</sup> quantiles of iron intake. Age and its square had significant regression coefficients across quantiles, with the exception of the lowest quantile (2.5<sup>th</sup>) along with the regression coefficient of Age cubed except for the two lowest quantiles (2.5<sup>th</sup> 25<sup>th</sup>). Males consumed significantly higher amounts of iron across all quantiles of intake when compared to females (reference group). Examining the impact of NSSEC showed a trend towards decreased iron intake compared to the reference group of higher managerial and professional occupations. All quantiles showed significantly lower intakes in the intermediate occupations, lower supervisory and technical occupations, semi-routine occupations, routine occupations and the never worked groups compared to the reference group. The lower managerial and professional occupations group had significantly lower intake across all quantiles with the exception of the highest (97.5<sup>th</sup>) quantile and those in the small employers and own account workers group had significantly lower iron intakes in the lowest quantile only, compared to those in the higher managerial and professional occupations group. Also iron intakes were higher on weekends (Saturday - Sunday) across quantiles compared to intake on week days (Monday - Friday).

**Table 13**

Estimated regression parameters for 2.5th, 25th, 50th, 75th and 97.5th quantiles and 95% confidence intervals for dietary iron intake in the UK.

The model used for quantile regression estimation was the linear mixed-effects quantile regression with ALD with SE estimated using bootstrap.

	2.5th	95%CI	25th	95%CI	50th*	95%CI	75th	95%CI	97.5th	95%CI					
Age	-0.009	0.06	-0.08	0.136	0.20	0.07	0.254	0.32	0.19	0.418	0.50	0.33	0.75	0.90	0.60
Age <sup>2</sup>	-0.043	0.18	-0.27	-0.307	-0.06	-0.55	-0.54	-0.30	-0.79	-0.842	-0.54	-1.14	-1.377	-1.06	-1.69
Age <sup>3</sup>	0.098	0.33	-0.13	0.229	0.48	-0.02	0.359	0.59	0.13	0.469	0.75	0.19	0.65	0.92	0.38
Gender: Females (Reference)															
Gender: Males	1.188	1.44	0.94	1.517	1.76	1.28	1.797	2.08	1.51	2.083	2.37	1.80	2.475	2.81	2.14
Weekday: Weekend (Reference)															
Weekday: Weekday	0.024	0.18	-0.13	0.148	0.27	0.02	0.29	0.44	0.14	0.383	0.54	0.22	0.502	0.72	0.29
Higher managerial and professional occupations (Reference)															
Lower managerial and professional occupations	-0.576	-0.26	-0.90	-0.493	-0.08	-0.91	-0.436	-0.01	-0.86	-0.434	-0.01	-0.86	-0.318	0.18	-0.82
Intermediate occupations	-0.753	-0.37	-1.14	-0.734	-0.21	-1.26	-0.723	-0.23	-1.22	-0.699	-0.14	-1.26	-0.656	-0.07	-1.24
Small employers and own account workers	-0.526	-0.03	-1.03	-0.473	0.13	-1.07	-0.447	0.15	-1.04	-0.398	0.11	-0.90	-0.305	0.26	-0.87
Lower supervisory and technical occupations	-1.043	-0.62	-1.47	-0.97	-0.51	-1.43	-0.971	-0.52	-1.42	-0.925	-0.34	-1.51	-0.911	-0.34	-1.48
Semi-routine occupations	-1.13	-0.83	-1.43	-1.09	-0.63	-1.55	-1.054	-0.66	-1.45	-0.98	-0.49	-1.47	-0.934	-0.47	-1.40
Routine occupations	-1.281	-0.87	-1.69	-1.273	-0.67	-1.88	-1.24	-0.64	-1.84	-1.202	-0.67	-1.73	-1.055	-0.46	-1.65
Never worked	-1.365	-0.12	-2.61	-1.308	-0.32	-2.30	-1.301	-0.23	-2.37	-1.303	-0.18	-2.43	-1.235	-0.04	-2.43
Other	-0.356	0.63	-1.35	-0.351	0.43	-1.13	-0.343	0.45	-1.13	-0.338	0.52	-1.20	-0.335	0.55	-1.22

\*The pseudo likelihood ratio test statistic is 0.00672, Degrees of freedom=1, p<0.001 for the regression coefficient of age cubic

#### 4.8.1 Quantile regression comparator analysis

To place the results of the above analysis into context a further set of estimated iron intake figures are presented in **Table 14**. These were produced using by a weighted linear mixed-effects regression model that includes the survey weighting but does not adjust for the other elements of the complex survey design, that is the strata and PSU, nor does it adjust for the skewed distributions seen in dietary data, though the mixed-effects model does allow the within- and between-person variation to be included (Bates et al., 2014c).

The mean iron intake was specified as

$$\mu_{ij} = \beta_0 + \beta_1 \text{Sex}_i + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i^2 + \beta_4 \text{Age}_i^3 + \beta_5 \text{NSSEC}_i + \beta_6 \text{Weekday}_{ij} + u_i,$$

where  $\mu_{ij}$  is the mean iron intake for person  $i$  on day  $j$ ,  $i = 1, \dots, 6109, j = 1, \dots, n_i$ , Sex is a binary variable (male or female), Age was the individual's age measured in years, Age<sup>2</sup> and Age<sup>3</sup> are quadratic and cubic age terms, respectively, NSSEC is the individual's NSSEC group and Weekday is a binary variable indicating Saturday or Sunday (0) or Monday - Friday (1) for individual  $i$  on day  $j$ . The random intercept  $u_i$  was assumed to have a normal distribution centred at zero.

Comparing the results of the quantile regression models with those of the weighted linear mixed-effects regression model (Table 14) showed some differences in findings. Overall results were similar, though in the small employers and own account workers, the median iron intake was lower than those in the higher managerial and professional occupations group though not significantly lower (Quantile regression -0.45, 95% CI:0.15, -1.04) whereas the mean intake was significantly lower (weighted linear mixed-effects regression -0.45, 95% CI:-0.12, -0.79).

**Table 14**

A comparison of regression parameters estimated for the 50<sup>th</sup> quantile (median) and mean along with 95% confidence intervals for iron intake in NDNS RP Y1-4 (2008-2012) participants using linear mixed-effects quantile regression with ALD with SE estimated using bootstrap and a weighted linear mixed-effects model.

	Quantile regression			weighted linear mixed-effects regression		
	50th	95%CI		Mean	95%CI	
Age	<b>0.254</b>	<b>0.19</b>	<b>0.32</b>	<b>0.237</b>	<b>0.21</b>	<b>0.27</b>
Age <sup>2</sup>	<b>-0.54</b>	<b>-0.79</b>	<b>-0.30</b>	<b>-0.412</b>	<b>-0.50</b>	<b>-0.32</b>
Age <sup>3</sup>	<b>0.359</b>	<b>0.13</b>	<b>0.59</b>	<b>0.207</b>	<b>0.14</b>	<b>0.28</b>
Gender: Females (Reference)						
Gender: Males	<b>1.797</b>	<b>1.51</b>	<b>2.08</b>	<b>1.782</b>	<b>1.61</b>	<b>1.95</b>
Weekday: Weekend (Reference)						
Weekday: Weekday	<b>0.29</b>	<b>0.14</b>	<b>0.44</b>	<b>0.274</b>	<b>0.18</b>	<b>0.37</b>
Higher managerial and professional occupations (Reference)						
Lower managerial and professional occupations	<b>-0.436</b>	<b>-0.86</b>	<b>-0.01</b>	<b>-0.386</b>	<b>-0.66</b>	<b>-0.11</b>
Intermediate occupations	<b>-0.723</b>	<b>-1.22</b>	<b>-0.23</b>	<b>-0.721</b>	<b>-1.08</b>	<b>-0.36</b>
Small employers and own account workers	-0.447	-1.04	0.15	<b>-0.453</b>	<b>-0.79</b>	<b>-0.12</b>
Lower supervisory and technical occupations	<b>-0.971</b>	<b>-1.42</b>	<b>-0.52</b>	<b>-0.86</b>	<b>-1.20</b>	<b>-0.52</b>
Semi-routine occupations	<b>-1.054</b>	<b>-1.45</b>	<b>-0.66</b>	<b>-0.903</b>	<b>-1.22</b>	<b>-0.58</b>
Routine occupations	<b>-1.24</b>	<b>-1.84</b>	<b>-0.64</b>	<b>-1.26</b>	<b>-1.59</b>	<b>-0.93</b>
Never worked	<b>-1.301</b>	<b>-2.37</b>	<b>-0.23</b>	<b>-1.281</b>	<b>-1.86</b>	<b>-0.70</b>
Other	-0.343	-1.13	0.45	-0.588	-1.26	0.08

**BOLD** indicates statistical significance at the p<0.05 level

## 4.9 Simulation

The performance of the linear mixed-effects quantile regression based on the ALD and its implementation in the lqmm R package have been previously assessed extensively through simulation (Geraci and Bottai, 2014) where 23 scenarios (**Table 15**) were examined that varied by cluster size and sampling distributions, using data generated according to

$$y_{ij} = (\beta_0 + u_i) + (\beta_1 + v_i)x_{ij} + \beta_2 + z_{ij} + (1 + \gamma x_{ij})e_{ij} \quad (26)$$

where  $\beta = (100, 2, 1)'$ ,  $u_i$  and  $v_i$  are random effects specific to each cluster,  $x_{ij} = \delta_i + \xi_{ij}$ , where  $\delta \sim N(0, 1)$ ,  $\xi \sim N(0, 1)$  and  $z_{ij} \sim \text{Binomial}(1, 0.5)$  and  $\sigma = 10$ . They generated multiple datasets which were modelled with the estimated beta coefficients compared to those used to simulate the data and they reported the relative bias and coverage. Relative bias is taken as the relative difference in the simulated estimate to the parameter that was specified, thus a relative bias of 0 would indicate that the simulated estimate contains no bias and coverage is given by the percentage of times the confidence intervals include, or cover, the initial parameters used in simulation. In general, relative bias of the model estimates was low (see Table 2. Geraci (2014)), although in some scenarios (19,20,21) it rose to around 0.5 for some values. The 90% coverage rate was higher than 90% in all cases though values as low as 90.2% were reported (Table 3. Geraci (2014)). The results of this simulation give confidence that estimates given by the lqmm method are accurate and from that basis further simulation work was carried out to test the extension of the methods to complex survey data.

To examine the performance of maximum pseudo likelihood estimation in combination with bootstrap estimates of variance for the analysis of complex survey data a Monte Carlo simulation study was carried out described in **Figure 26**. This involved generating 100 data sets with predefined parameters that varied by 6 different factors then performing 200 bootstrap replications to calculate the relative bias and coverage of estimated values for three parameters (Intercept,  $x$  and  $y$ ) of three quantiles (50<sup>th</sup>, 75<sup>th</sup> and 90<sup>th</sup>) which were then compared to the parameters used to simulate the data.

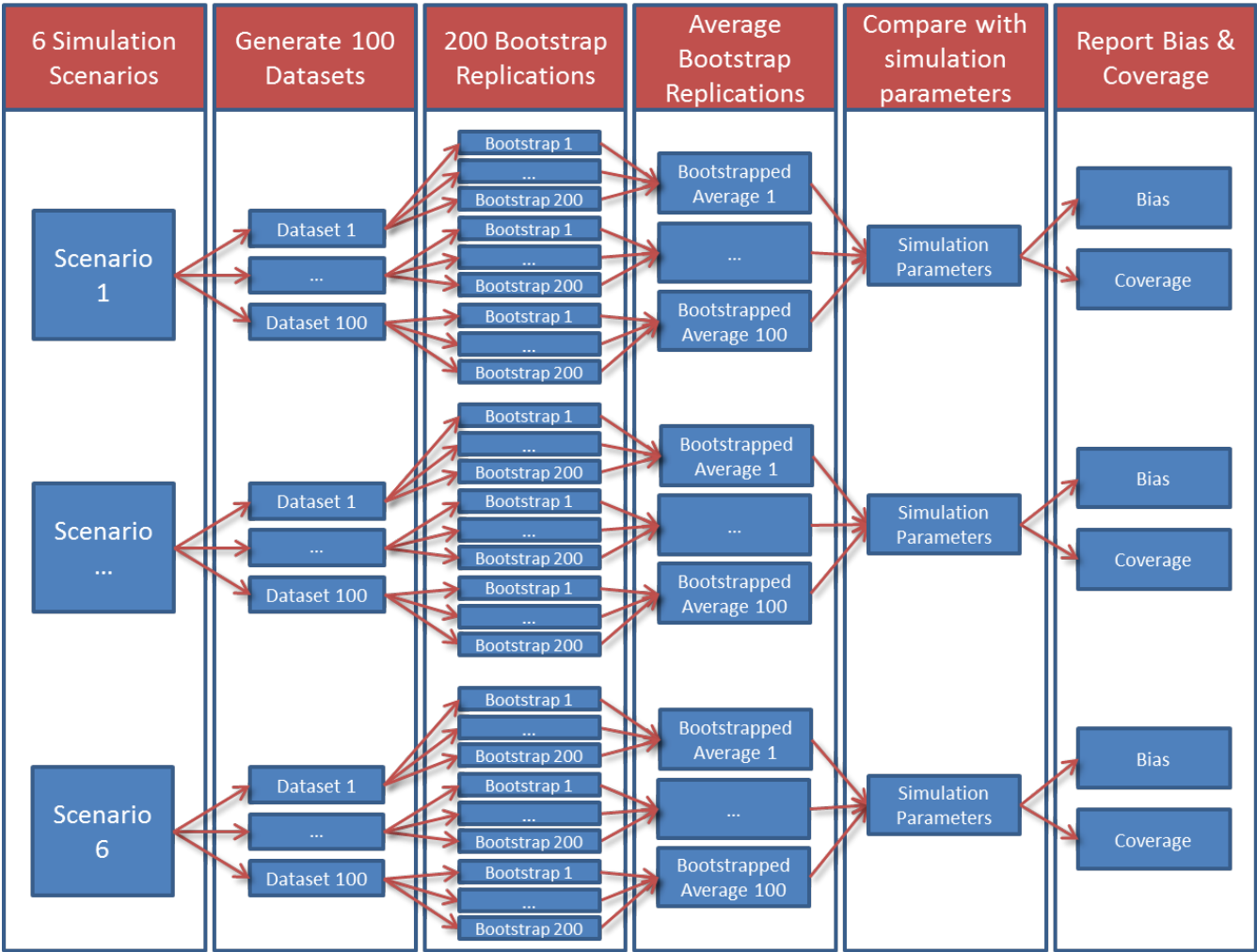
**Table 15**

Simulation study scenarios used to evaluate the performance of the lqmm package adapted from (Geraci, 2014)

Model Description	$(n, M)$	$u$	$v$	$e$	$\gamma$
Location shift symmetric	(5,50)	$N(0, 5)$	-	$N(0, 5)$	0
Location shift symmetric	(5,50)	$t_3$	-	$N(0, 5)$	0
Location shift heavy tailed	(5,50)	$N(0, 5)$	-	$t_3$	0
Location shift heavy tailed	(5,50)	$t_3$	-	$t_3$	0
Location shift asymmetric	(5,50)	$N(0, 5)$	-	$\chi_2^2$	0
Location shift asymmetric	(5,50)	$t_3$	-	$\chi_2^2$	0
Location shift symmetric $cor(u, v) = 0$	(5,50)	$N(0, 5)$	$N(0, 5)$	$N(0, 5)$	0
Location shift heavy tailed $cor(u, v) = 0$	(5,50)	$t_3$	$t_3$	$t_3$	0
Location shift symmetric	(10, 100)	$N(0, 5)$	-	$N(0, 5)$	0
Location shift symmetric	(20, 200)	$N(0, 5)$	-	$N(0, 5)$	0
Location shift heavy tailed $cor(u, v) = 0$	(10, 100)	$t_3$	$t_3$	$t_3$	0
Location shift heavy tailed $cor(u, v) = 0$	(20, 200)	$t_3$	$t_3$	$t_3$	0
Location shift heavy tailed $cor(u, v) = 0$	(20, 200)	$t_3$	$t_3$	$\chi_2^2$	0
Heterscedasctic symmetric	(10, 100)	$N(0, 5)$	-	$N(0, 50)$	0.25
Heterscedasctic heavy tailed	(10, 100)	$t_3$	-	$t_3$	0.25
Heterscedasctic asymmetric	(10, 100)	$t_3$	-	$\chi_2^2$	0.25
Location shift symmetric $cor(u, v) > 0$	(10, 100)	$N(0, 5)$	$N(0, 5)$	$N(0, 5)$	0
Location shift symmetric $cor(u, v) < 0$	(10, 100)	$N(0, 5)$	$N(0, 5)$	$N(0, 5)$	0
Location shift heavy trailed $cor(u, v) > 0$	(10, 100)	$t_3$	$t_3$	$t_3$	0
Location shift heavy trailed $cor(u, v) < 0$	(10, 100)	$St_3$	$St_3$	$t_3$	0
Location shift heavy trailed $cor(u, v) > 0$	(10, 100)	$St_3$	$St_3$	$t_3$	0
Location shift with 5% contamination	(10, 100)	$N(0, 5)$	-	$\chi_2^2 + N(0, 50)$	0
Location shift with 5% contamination	(10, 100)	$N(0, 5) + N(0, 50)$	-	$\chi_2^2$	0

**Figure 26**

Schematic illustration of a simulation study carried out to examine the coverage and relative bias of maximum pseudo likelihood estimation with bootstrapped variance estimates



**Table 16**

Scenarios for the simulation study.

	Random effects	Number of subjects	Number of strata	Number of repeated measures
(1)	Normal	200	5	2
(2)	Normal	200	5	4
(3)	Normal	1000	25	4
(4)	Asymmetric Laplace	200	5	2
(5)	Asymmetric Laplace	200	5	4
(6)	Asymmetric Laplace	1000	25	4

The six different scenarios (**Table 16**) varied by the number of subjects that were included, the number of observations per subject and the random effect distribution: either Gaussian or Laplacian. The scenarios proposed here are reflective of real national complex survey data, in terms of design and large variability observed in dietary data, with the number of measures fixed at 2 and 4 to reflect two 24HR and the NDNS RP's four day diaries. The number of subjects chosen was either 200 and 1000 and was selected to examine the performance of the methods under the most challenging conditions, in terms of fewer data points, likely to be seen. The former sample size is considered to be at the lower end of the number of subjects typically recruited in national surveys employing complex sampling methods but is greater than the minimum sample size used by Souverein et al. (2011) who used 150 and 500 samples and Laureano et al. (2016) who used 150, 300 and 500 samples respectively, in their simulation studies of national survey data.

The data were simulated as follows: intake  $Y_{ijkl}$  for individual  $i$  on occasion  $j$  in PSU  $k$  in strata  $l$  were simulated from  $N(\mu_{ijkl}, 6)$  where  $\mu_{ijkl}$  is given by

$$\mu_{ijkl} = \beta_0 + \beta_1 x_{ijkl} + \beta_2 z_{ijkl} + u_{ikl} + v_{kl} \quad (27)$$

where  $x_{ijkl}$  was sampled from  $N(0, 1)$  and  $z_{ijkl}$  from a Bernoulli distribution with probability of success 0.5; there are two random effects, first  $u_{ikl}$  was defined at the individual



level and was sampled from either  $N(0, 6)$  or  $ALD(\mu, \sigma, p)$ , with  $\mu = 0$ ,  $\sigma = 3/\sqrt{2}$  and  $p = 0.5$ . The choice of these parameters for the ALD is such that they match those when the random effect is normally distributed, i.e. the mean is zero and the standard deviation 6 (Yu and Zhang, 2005). The second random effect  $v_{kl}$  was defined at the PSU level and was simulated from  $N(0, 1)$ . The regression coefficients:  $\beta_0 = 10$ ,  $\beta_1 = 1$ ,  $\beta_2 = 0.5$ , were chosen to be reflective of realistic variances seen in dietary survey data and yielded quite small means of the outcome variable relative to its variance, in contrast to the parameters  $\beta = (100, 2, 1)'$  and  $\sigma = 10$  used by Geraci (2014). To illustrate, the coefficient of variance of the simulated data used here, calculated by  $CV = \sigma/\mu$ , of 0.59 is similar to values seen in the NDNS RP and falls within the range reported by the NDNS RP for iron intake for all sources (Table 5.33a of the NDNS RP report (Bates et al., 2014a)), which ranged from 0.30 in girls aged 4-10y to 0.85 in women aged 19-64y whereas, the values chosen by (Geraci, 2014) gave considerably smaller CVs for example the CV for scenario 1 (Table 15), of 0.10, which is outside the reported range for iron intake.

Probability weights for each cluster were simulated from  $N(1, 0.25)$ . Strata varied in size (Table 16) and included 2 PSUs per stratum. Numerical integration of the individual random intercept was carried out using adaptive Gaussian quadrature for normal random effects and Gauss-Laguerre quadrature for the asymmetric Laplace random effects, and in each case 20 quadrature points were specified. 100 datasets were created and from each dataset, 200 bootstrap replicates were taken. It took an average of approximately 110 milliseconds for a single model when  $p = 0.5$  to estimate median regression using a windows 7 i5 laptop with 16Gb Ram, though the time taken is dependent upon the number of subjects, the number of repeated measures and the sparsity of the data for extreme quantiles. Relative bias and coverage of 90% confidence intervals are presented in **Tables 17** and **18**. Three parameters were estimated for three quantiles (median, 75<sup>th</sup> and 90<sup>th</sup>) under each scenario. This gives 54 estimated values in total, and on the whole, the methods performed well with the average coverage close to 90% and scenarios 1,3 and 6 had a coverage probability of 90% or higher and scenario 4 close to 90% coverage at 89%. Scenarios 2 and 5 had the low-

est average coverage at 88% and 86% respectively and the majority of values had a coverage probability or equal to 90% coverage except for the intercepts, which may not be surprising given the large variation with which the data were generated. The average coverage probabilities for the intercept in scenarios 2 and 5 at the median had the lowest rates of coverage at 79% and 78% respectively. These scenarios both had the lowest number of subjects (200) and the highest number of repeated measures (4) but varied in terms of their random effects distributions (Normal and Asymmetric Laplace). When the number of repeated measures was reduced, as in scenarios 1 and 4, whilst keeping the number of subjects constant the coverage probabilities for the intercepts increased to 85% and 83% and increased again to 87% and 91% as the number of subjects included rose to 1000 (scenarios 3 and 6) suggesting that 200 subjects and 4 repeated measures is insufficient for reliable coverage. In the majority of cases the model performs well with a mean bias of 0.049 overall. Parameters with the greatest relative bias were those for  $z$ , estimated at the 90<sup>th</sup> which is at the extremes of the distribution and may be due to relatively fewer data points. The average relative bias varies by scenario ranging from 0.099 and 0.111 in scenarios 1 and 4 with 200 subjects and 2 repeated measures only differing by random effects distributions to 0.015 and 0.013 in scenarios 3 and 6 with 1000 subjects and 4 repeated measures, again only differing by random effects distribution indicating that the relative bias is reduced as the number of subjects and repeated measures increases. In the application of these methods above using NDNS RP data the number of strata per PSU is comparable and the number of repeated measures is the same at 4 but the number of subjects (6109) is greater than the 1000 sampled here, thus the coverage probability and relative bias can be expected to be comparable or better than the values reported in the simulation.

**Table 17**

Coverage probability of 90% confidence intervals calculated from simulated data under six different scenarios.

	Median			75 <sup>th</sup> quantile			90 <sup>th</sup> quantile			Average coverage
	Intercept	x	z	Intercept	x	z	Intercept	x	z	
(1)	85	91	92	87	90	92	91	95	91	90
(2)	79	89	82	87	91	84	94	99	90	88
(3)	87	84	93	83	94	93	83	99	94	90
(4)	83	90	90	87	91	91	87	92	90	89
(5)	78	90	85	85	87	81	92	93	87	86
(6)	91	93	91	93	87	90	83	96	92	91

**Table 18**

Relative bias of estimated parameters from data simulated under six different scenarios.

	Median			75 <sup>th</sup> quantile			90 <sup>th</sup> quantile			Average bias
	Intercept	x	z	Intercept	x	z	Intercept	x	z	
(1)	0	-0.024	-0.106	-0.001	0.002	0.089	0.052	-0.014	0.891	0.099
(2)	0.008	0.027	-0.072	-0.022	0.067	0.072	0.026	0.060	0.563	0.081
(3)	0.004	0.018	-0.046	-0.009	0.005	0.021	0.033	0.019	0.092	0.015
(4)	0.005	0.009	0.005	0.004	0.039	0.066	0.066	0.025	0.776	0.111
(5)	-0.018	-0.014	0.155	-0.025	-0.069	0.187	0.020	-0.027	0.582	0.088
(6)	0.001	-0.002	-0.051	-0.009	-0.010	0.052	0.046	-0.023	0.117	0.013

## 4.10 Discussion

This chapter presented a novel approach to quantile regression of repeated measures data collected under a complex survey design. This was illustrated through quantile plots showing the effect of age and sex upon dietary iron intake using data collected by the NDNS RP. The NDNS RP is the only source of high quality nationally representative dietary survey data in the UK, and as such, is used as the evidence base for policy change and implementation. Using these data an approach that is capable of providing a more precise characterisation of dietary intake for subgroups of the UK population is demonstrated. By accounting for measurement error through a random intercept, multistage sampling and survey weighting using bootstrap resampling and a pseudo likelihood approach, it is possible to examine the association between the outcome variable and the explanatory variables at the extremes of the distribution. Due to the increased risk of deficiency, the LRNI is contrasted with actual quantiles of iron intakes to highlight the high number of people below this recommendation. The greatest subgroup of individuals below the LRNI are females aged 11-50y, which is in line with other surveys, though is of concern as this range covers those most likely to be childbearing and the impact of low iron during pregnancy and breast feeding is severe for both mother and infant. Participants in the 11-18y age group are likely to be experiencing a period of growth and development with associated increase in muscle mass and blood volume, both of which require iron. In adolescent females there are increased losses of iron with menstruation and females, compared to males, are likely to consume less iron due to weight loss diets. Iron deficiency can cause lethargy and shortness of breath leading to reduced physical activity. This plays an important role in bone mineralisation through bone loading; reduced activity can cause lower bone mineral density and a lower peak bone mass which in turn leads to increased risk of osteoporosis (Troy et al., 2018). The cost to the NHS from hip fractures is estimated at £1.1Billion per annum (Leal et al., 2016). Moreover iron deficiency has been linked to cardiovascular disease (von Haehling et al., 2015) which is also a major economic burden. Perhaps of greatest concern is the impact of iron deficiency for females between 11-50y on the infant during pregnancy. The prevalence seen here is similar to

that found in a recent study of pregnant females (Barroso et al., 2011). Low iron can lead to increased rates of premature births, infection and lower birth weights.

Iron intakes estimated by a weighted linear mixed-effects regression method estimating mean intake showed that members of the small employers and own account workers had significantly lower iron intakes compared to those in the higher managerial and professional occupations groups, though this wasn't found in the results of the quantile regression model estimated at the median. The difference is likely due the impact of outliers shown iron intake impacting upon the distribution of iron intakes (e.g. Figure 22). This can have implications where one group is erroneously identified as having low iron intakes and resources are diverted towards the group to increase their iron intakes.

Fitting the model relies heavily on computing power and as a consequence can take many hours for a model using ~6000 subject clusters to converge. To reduce this time it is possible to parallelise the code to exploit the multiple cores available available on most computers. Each iteration took approximately 4 minutes using a Windows i5 machine with 16Gb of RAM. This time is dependent also upon the number of parameters being estimated. Similarly the speed of convergence is dependent on the data points at the quantile of interest. There is also a pragmatic trade-off between convergence of estimates and the time taken; we have used 500 bootstrap replications.

Semi-parametric QR models can occasionally demonstrate crossing of the quantile curves. This usually occurs when plotting a number of curves with few observations, which can lead to the implausible case in which the 95th percentile is lower than the 90th, for example. This is because the quantiles are estimated separately. Bondell et al. (2010) presented a simple constrained version of quantile regression to avoid quantile crossing. Wu and Liu (2009) suggested the quantiles are fitted sequentially with the subsequent quantile curve restricted from crossing the previous one. This, however, is dependent on the order in which the curves are fitted. In a parameterised setting, quantiles can be estimated simultaneously and in relation to the distribution, which has the advantage that crossing will not occur (Stasinopoulos and Rigby, 2007).

## 5 Iron prescription costs across the UK

The previous chapters have introduced and proposed methods for dealing with the challenges that arise when estimating the mean or quantiles of skewed dietary data sampled under a complex sampling plan that includes sampling weights. These were addressed by modelling intakes of episodically consumed nutrients through a two part logistic - generalised gamma regression model with correlated random effects (Chapter 3) and modelling quantiles of intake using an asymmetric Laplace distribution (Chapter 4). In these two chapters modelling of iron intake was used to illustrate the methods. In this chapter I shall use national electronic health records to examine the amount spent upon the treatment of iron deficiency in the UK and adjust for iron intake using dietary iron taken from the NDNS RP.

### 5.1 Introduction

Iron deficiency impacts the UK population to a large extent with approximately 4.7 million individuals deficient in iron (WHO and CDC, 2008). Iron deficiency disproportionately affects certain groups with 12.9% of young children (1.5-3y), 5.8% of women aged 19-64 and 13.5% of the elderly (65+y) reporting iron levels below the haemoglobin threshold for iron deficiency (**Table 19** in Section 5.5).

### 5.2 Iron deficiency

Iron deficiency occurs when iron stores become depleted and are no longer sufficient to meet iron requirements often a result of low iron intake from the diet (Killip et al., 2007). Infants, children and women of reproductive age are more likely to be deficient because of increased iron requirements during growth and excretion through menstruation particularly in those with higher rates of bleeding. Similarly those with inflammatory bowel disease suffering from blood loss in the stomach or intestine are at increased risk (Gasche et al., 2004).

Individuals are typically classified as being iron deficient using blood measures such as haemoglobin or serum ferritin. Although an individual can be classified as having an increased risk of iron deficiency if their dietary intake is below the LRNI (Scientific Advisory Committee on Nutrition (SACN), 1991). This is an age and sex specific value that represents an adequate intake for 2.5% of the population.

### **5.3 Iron deficiency consequences**

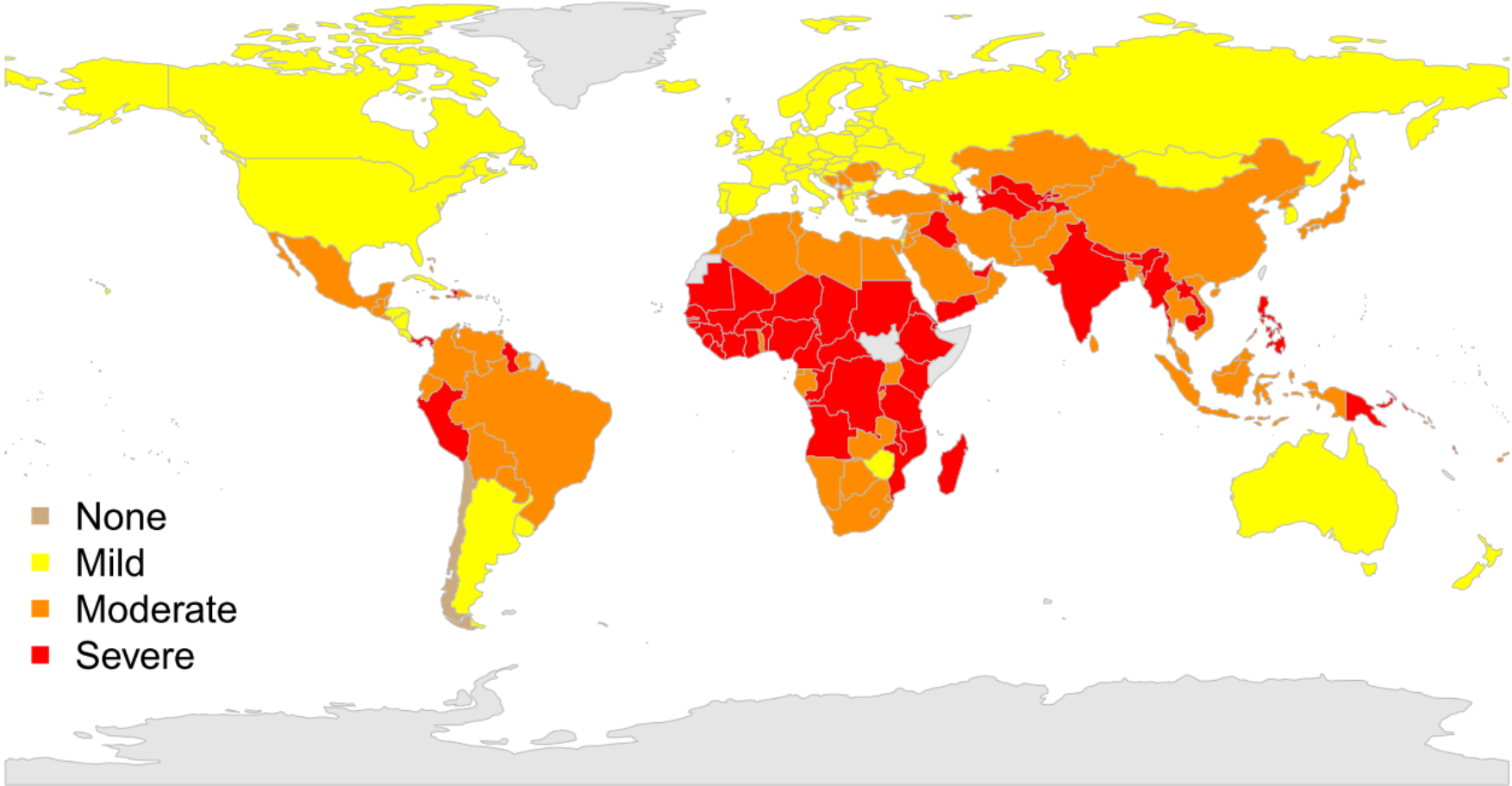
Iron deficiency can lead to impaired physical and cognitive development in children, increased risks for mother and neonate, and can cause delayed central nervous system development (Beard, 2007). It can also lead to impaired immune status and thus increased morbidity (Oppenheimer, 2001) and has been shown to lead to reduced work capacity (Haas and Brownlie, 2001).

### **5.4 Global iron deficiency prevalence**

Iron deficiency affects a sizeable portion of the developed and developing world (McLean et al., 2009). Globally it is estimated that approximately 1.1 billion individuals are anaemic due to iron deficiency (World Health Organisation, 2001). Across the world an estimated 17,000 deaths per annum are due to iron deficiency (WHO, 2012). **Figure 27** displays the extent to which iron deficiency impacts upon health in non-pregnant women of childbearing age by country, and highlights that although deficiency is most severe in developing nations such as large parts of Africa, Asia and Latin America it is still of mild concern in developed countries including the UK.

Figure 27

The global extent of the public health problem of iron deficiency anaemia in non-pregnant woman of child bearing age.



Countries are coloured according to level of public health concern. Countries without data are coloured gray. Adapted from WHO (2012).



## **5.5 UK iron deficiency prevalence**

In the UK, a number of groups have been found to have high levels of iron deficiency (Nicholson et al., 2014). 5.7% of girls aged 4-10y were classified as iron deficient according to their haemoglobin levels, in females aged 11-18y this rose to 7.4%, in females aged 19-64y this was 9.9% and 12.3% in females aged 65+y. In males aged 65+y, 15.2% were classified as below the haemoglobin threshold, Overall 12.9% of children aged 1.5 to 3 years and 13.5% of adults aged 65+y were below the threshold for haemoglobin (Nicholson et al., 2014) (Table 19). Across the UK it is estimated that 4.7 million individuals are deficient in iron (WHO and CDC, 2008).

**Table 19**

Haemoglobin levels (g/dL) and status by age and sex from NDNS RP Years 1-4 (2008-2012)

	Males				Females				Overall				
Age Group	4-10	11-18	19-64	65+	4-10	11-18	19-64	65+	1.5-3	4-10	11-18	19-64	65+
Mean (g/dL)	13.0	14.4	14.9	14.4	12.8	13.1	13.2	13.2	12.0	12.1	13.8	14.0	13.7
Median (g/dL)	13.0	14.3	15.0	14.5	12.8	13.2	13.3	13.5	12.0	12.1	13.7	14.1	13.8
sd (g/dL)	0.94	1.10	0.88	1.49	0.83	0.89	0.98	1.21	0.89	0.98	1.20	1.24	1.46
2.5 <sup>th</sup> percentile (g/dL)	11.2	12.4	13.1	11.3	10.5	11.1	11.2	9.5	10.6	10.6	11.6	11.4	10.5
97.5 <sup>th</sup> percentile (g/dL)	15.4	16.4	16.5	16.8	14.7	14.8	15.1	15.0	14.0	14.0	16.2	16.4	16.4
Below Hb Threshold <sup>a</sup> (%)	3.1	1.8	1.5	15.2	5.7	7.4	9.9	12.3	12.9	4.4	4.5	5.8	13.5
n	138	261	562	143	116	255	778	200	50	254	536	1340	343

<sup>a</sup> Haemoglobin lower thresholds given by Scientific Advisory Committee on Nutrition (2011): 1.5-4y males <11g/dL, 1.5-4y females <11g/dL, 5-11y males <11.5g/dL, 5-11y females <11.5g/dL, 12-14y males <12g/dL, 12-14y females <12g/dL, 15y+ males <13g/dL, 15y+ females (non-pregnant) <12g/dL

Adapted from NDNS RP Table 6.1 (Nicholson et al., 2014).

## 5.6 Iron deficiency treatment

Iron is present in many foods either naturally occurring or added as a fortificant. However when dietary sources fail to meet iron requirements, treatment for iron deficiency in moderate cases is through oral supplementation. This is in tablet form commonly supplied as ferrous sulphate, ferrous fumarate and ferrous gluconate (Allen, 2002). Doses between 100-200mg are advised 3 times a day typically as ferrous sulphate, and may be prescribed prophylactically in those at increased risk of iron deficiency. Increases of between 100-200mg /100 mL over 3-4 weeks are typically observed and once haemoglobin values have reached the reference range, continued treatment is advised for a further 3 months to replenish the iron stores (Joint Formulary Committee, 2017).

Iron supplementation is associated with side effects which may include increased flatulence, abdominal discomfort or pain, nausea, constipation, loose or discoloured stools (Cancelo-Hidalgo et al., 2013) and these can have an impact upon adherence to a treatment regimen. Rates of adherence as low as 47% have been reported in some groups (Lacerte et al., 2011). However the degree to which adverse events are reported is dependent on the form of iron given and has been shown to vary from around 31% for ferrous gluconate and ferrous sulfate to 47% for ferrous fumarate (Lacerte et al., 2011).

Iron status is dictated by dietary iron intake and when intake fails to meet iron requirements and symptoms of iron deficiency present themselves, the first line treatment is iron supplementation prescribed by a GP (Joint Formulary Committee, 2017). GPs are clustered within health boards who are, amongst other things, responsible for the commissioning of services and the amount spent on prescriptions. In health boards with financial constraints it is possible that treatments are prioritised at the expense of iron deficiency leading to a disparity in the amount spent of iron deficiency medication across health boards. For example The beechwood medical practice issued notice that it will not be issuing prescriptions for many items including vitamin and mineral supplements in a bid to save money (Beechwood Medical Practice, Bristol CCG, 2016), and

similarly Mid Essex CCG has issued a statement that due to a £464 million deficit they are proposing cuts for items patients such as gluten-free items that patients could purchase for cheaper than the CCG cost (Mid Essex CCG, 2015). This variation can have an impact upon the patient whose treatment can then become dependent on where they live in the country: the so-called "postcode lottery". Therefore I shall describe iron prescriptions in the UK, estimating the median amount spent on iron medication at the health board level to highlight variation in prescribing practice across the UK and in particular across health boards and present the findings in easily interpretable maps.

Expenditure on iron prescriptions is likely to depend directly on the population levels of iron deficiency, or inversely on iron bioavailability. This information can be obtained from the NDNS RP. When comparing intakes of iron to population reference levels such as the RNI or LRNI as carried out in Chapter 4 it is appropriate to simply sum the iron content of each food eaten, however when trying to determine the bioavailability of iron, a physiological method that considers the effect of the interaction of foods and nutrients on iron absorption is required. This is because the amount of iron available for uptake in the body is dependent on other nutrients that can either have a positive or negative impact on uptake (Zijp et al., 2000). Dietary iron comes in two forms namely haem and non-haem which come from animal and non-animal sources respectively. Haem iron is the more stable form of iron and is not affected by nutrients consumed at the same time and makes up approximately 10% of total dietary iron intake (Reddy et al., 2006). In contrast, the amount of non-haem iron available for uptake is impacted by the intake of calcium, vitamin C phytate, polyphenols and tannins and also meat fish and poultry intake, some of which improve availability and some decrease availability, see **Table 20**.

Research examining variations in medication prescriptions is fairly limited particularly when mapping prescription data. This has been carried out for prescriptions in England (Rowlingson et al., 2013), Australia (Mullins et al., 2009), the US (Allen et al., 2010; Ashton et al., 1999; Sargen et al., 2012), Taiwan (Cheng et al., 2011) and between countries (Domanski et al., 2004). Rowlingson et al. (2013) examined prescriptions for diabetes medication in England combined with diabetes incidence rates taken from

**Table 20**

Dietary factors affecting the absorption of Iron

Increase Iron Absorption	Decrease Iron Absorption
Vitamin C	Polyphenols
Red meat	Calcium
Fish	Phytate
Poultry	Tannins

the GP Quality and Outcomes Framework (QOF). The QOF is a programme created in 2004 to standardise care provision given by GP practices dependent upon their performance. GP practices are rewarded financially for meeting achievement indicators across a range of clinical areas including chronic disease, heart failure and diabetes. For example there are four indicators for heart failure: that include prescribing the appropriate medication and ensuring that a certain number of patients are referred to exercise-based rehabilitation programmes (NHS Digital, 2015). Using the prescription and QOF data GP practices were categorised to identify those whose prescribing practice differed significantly from the expected spending for a GP practice when adjusted for the age and sex distribution of the practice along with the total number of patients. However the QOF does not monitor the incidence of those suffering from iron deficiency or iron deficiency anaemia. For diseases that are assessed, the QOF provides a powerful dataset when combined with prescription information allowing for detailed modelling of incidence of disease in England.

As a public funded organisation the NHS has released information on all prescriptions dispensed in the UK with the aim of increased spending transparency. Using these data I present an analysis examining the median amount spent on prescriptions for iron medication by each of the 235 health boards in the UK and rank the health boards in to quintiles of spending whilst adjusting for patient demographics and dietary iron intake. Levels of expenditure are then presented in a choropleth (a map coloured with intensity related to ranking) which easily allows health boards with higher and lower

spending rates to be identified. Further analysis highlighting the impact of modelling bioavailable iron is included as a comparator.

## **5.7 Methods**

### **5.7.1 The data**

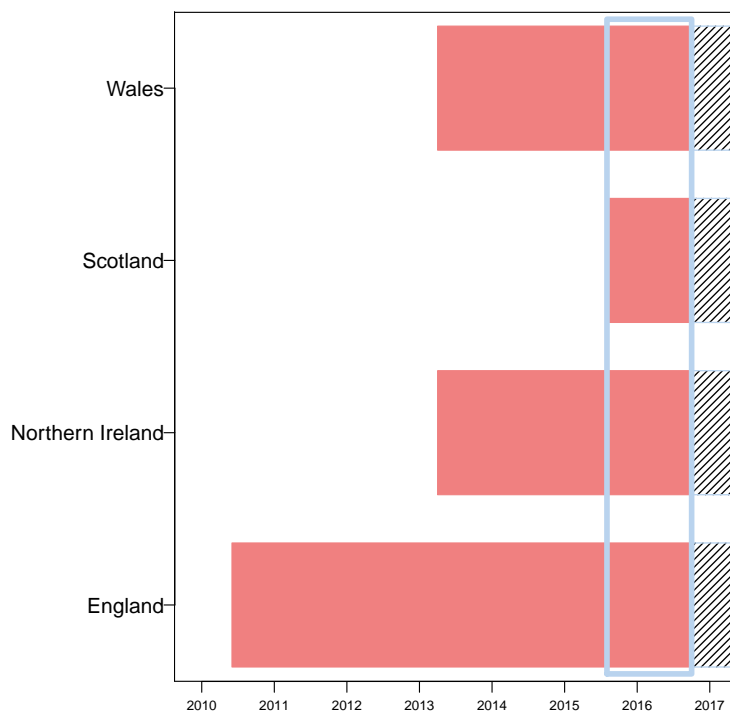
Data files containing the number of prescriptions dispensed, cost per prescription and type of medication were downloaded from the English, Northern Irish, Scottish and Welsh governmental websites. Data on the patients that received the prescriptions was not available thus, overall GP practice level data was used. Separate data files that contained the total number of registered patients at each GP practice split into age and sex groups were also downloaded. Patients' age groups were available in the categories 0-4y, 5-15y, 16-44y, 45-64y and 65+y and a further set of data files that contained Index of Multiple Deprivation (IMD) scores were used. The IMD values were based on the postcode of the GP practice. Each GP practice has a unique identifying number which was used to link the prescribing data with the registered patient data and the IMD data, each GP had a IMD ranking from 1 to 9461. A small number of records did not have patient information and were therefore excluded. These records were missing due to the prescription coming from either an out of hours service, a hospital or in some cases the GP practice had closed suggesting the patient had been issued the prescription then kept until it was dispensed. The four countries vary in the availability of prescription data with data available for England from June 2010. However to allow for the entire UK to be examined the prescriptions presented here are at the earliest point that all four countries' data are available simultaneously which is from October 2015 until July 2016 (see **Figure 28**).

### **5.7.2 Health boards**

Each country has a different name for the cluster of GPs. In England there are 209 regions known as Clinical Commissioning Groups (CCG)s which were created in 2013

**Figure 28**

Timeline depicting the availability and selection of prescription data by country.



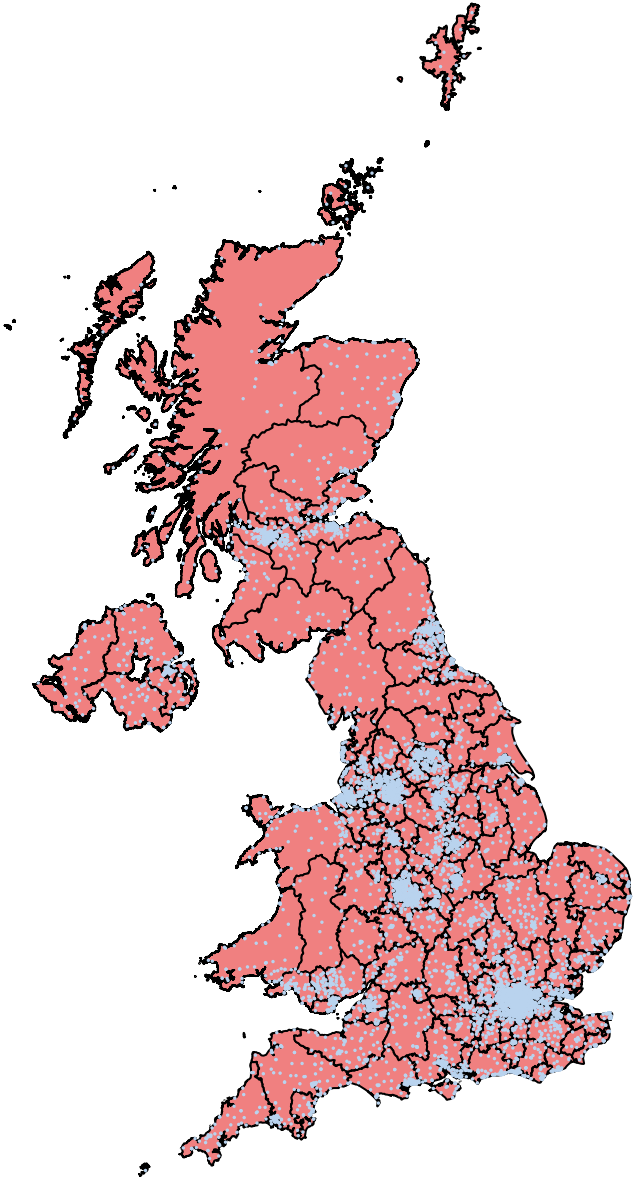
and replaced Primary Care Trusts (PCT)s. In Wales they are known as health boards and there are 7 of them, in Northern Ireland there are 5, known as regional trusts and in Scotland there are 14 regional health boards so for ease they shall be referred to as health boards throughout. **Figure 29** shows the location of each GP practice with a dispensed prescription for iron during the data collection period. Health boards are delineated by black borders. Practices are clustered within health boards and the health board is responsible for the budget assigned to each practice.

### 5.7.3 Data sources

**Figure 30** displays the location of the governmental webpages used to download data. Prescription, registered patients and IMD data were downloaded from open source repositories managed by each country and are available from the following locations. From the English [data.gov.uk](http://data.gov.uk) website, Northern Ireland from the open data NI, from the Scottish information services division part of NHS National Services Scotland and

**Figure 29**

The distribution of UK GP Practices by postcode, with black borders indicating health board boundaries. GP surgeries indicated by a blue dot.



Welsh data came from the primary care services website. The number of patients registered at a GP practice in England were downloaded from the NHS digital website which contains data from January 2016. IMD values were obtained based on the post-code of the GP practice from the open data communities website which is part of the



Ministry of Housing, Communities & Local Government and contains a lookup table that returns the corresponding IMD value.

**Figure 30**

Sources of data for England, Northern Ireland, Scotland and Wales

	England	Northern Ireland	Scotland	Wales
Prescription Data	NHS Digital	Open Data NI	NHS National Services Scotland	NHS Wales Primary Care Services
Index of Multiple Deprivation (IMD)	Open Data Communities .org	NI Statistics & Research Agency	Scottish Government	Welsh Government
Patient Numbers	NHS Digital	NI Statistics & Research Agency	NHS National Services Scotland	Welsh Government
Practice Location	NHS Digital	NI Statistics & Research Agency	NHS National Services Scotland	Welsh Government
Map Coordinates	NHS Digital	Open Data NI	GADM	GADM

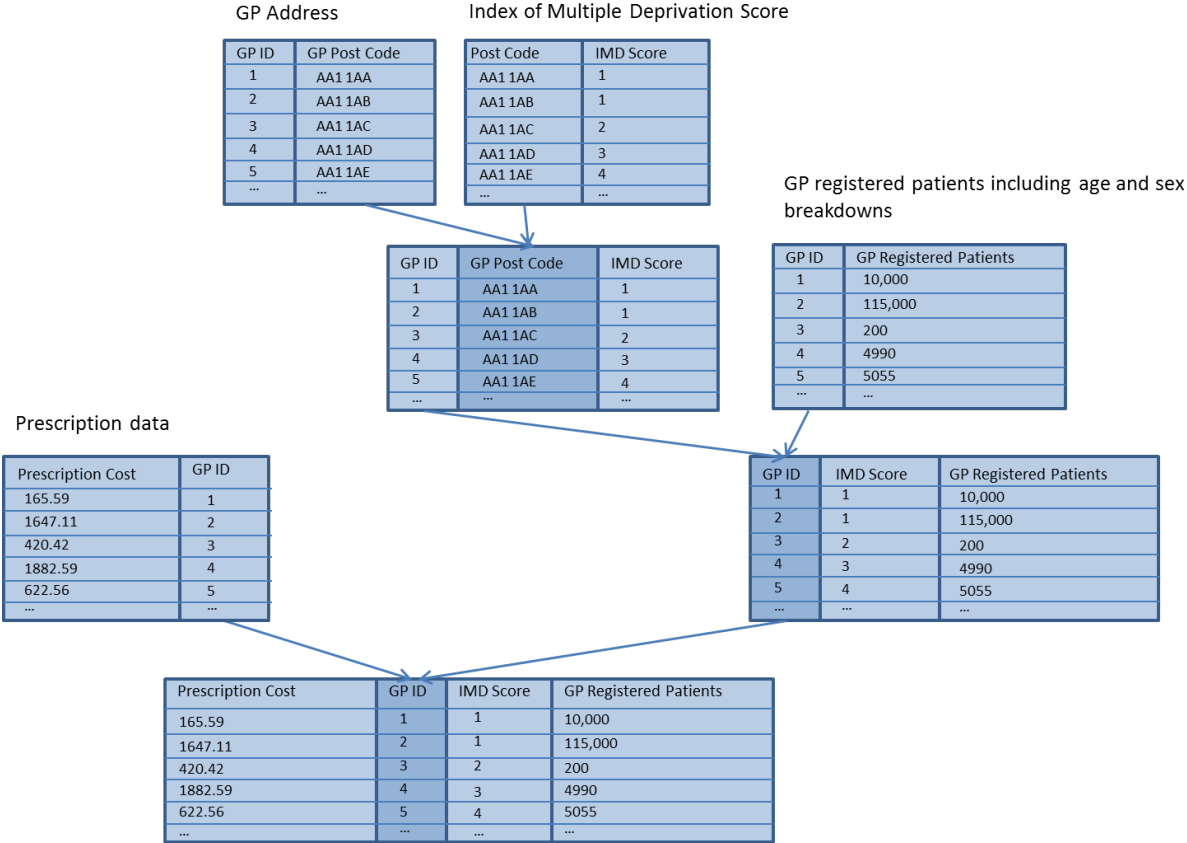
GP addresses for England were taken from NHS Digital, data containing the number of patients registered at Scottish GP practices were extracted from IDS Scotland and values were taken as of 1st of October 2015. Scottish IMD information was obtained from the Scottish government website and were matched to GP practice postcodes found at the IDS Scotland website.

Welsh GP practice addresses and GP registered patient numbers were found at Welsh government website along IMD information and the addresses of GP practices. Patient information regarding the age and sex distribution of patients registered at GP practices

in Northern Ireland were requested from the health and social care Northern Ireland team as only the total number of registered patients were available on the website. Northern Irish IMD was not available at the postcode level but was provided in small geographical units (Lower Layer Super Output Area (LSOA)). Then the LSOA of the GP practice was found by converting the postcode from <http://mapit.mysociety.org/postcode/#> (where # is the GP practice postcode). The IMD scores were then found from the Northern Ireland statistics and research agency, the latest available IMD data is from 2010. URLs are given in full in **Appendix K**.

**Figure 31**

Schematic diagram illustrating the merging of files containing iron prescriptions, GP addresses, Index of Multiple Deprivation ranking and registered patients



**Figure 31** illustrates the process of merging the four files to create a final dataset used for analysis. Index of Multiple Deprivation was included based on the postcode assigned to the IMD and this was used to merge with the file containing the GP address. This was then merged with a file containing the total number of registered patients

broken down by age and sex using the GP ID which exists in both files. This file in turn was then merged, again by GP ID, to the file containing prescription information for each GP surgery giving a final data set ready for analysis (see **Appendix L** for a sample of the data).

#### **5.7.4 Mapping health boards**

To present the health boards and GP practices geographically, a map of the UK was created with health boards delineated by combining maps from each country (see Figure 30 for sources). In England a health board map was available from the NHS digital website, in Northern Ireland this come from Open Data NI whereas maps for Scotland and Wales came from the database of global administrative areas (GADM)

[http://biogeo.ucdavis.edu/data/gadm2.8/rds/GBR\\_adm2.rds](http://biogeo.ucdavis.edu/data/gadm2.8/rds/GBR_adm2.rds) which included local authority boundaries that matched the boundaries of the health boards. Some adjustment was made to the names of the local authority areas to ensure that they matched the names of the corresponding health boards e.g. the Orkney Islands became Orkney. In addition "and" was changed to "&", the list of adjustments and code used to create the maps is provided in **Appendix J**.

#### **5.7.5 Number of patients by GP practice at the time of prescription**

The number of patients registered at each GP practice is available on a quarterly basis whereas the prescription data is updated monthly, therefore it was not possible to match patients across the duration of data collection. Furthermore the prescription date corresponds to the date the medication was issued not the date the prescription was written. In most cases the difference is expected to be minimal although it is possible that a patient may have moved to a different health board since the prescription was issued, particularly if it was a repeat prescription. This was seen in one case where a prescription was dispensed a number of months after the GP practice had closed. Therefore there is a possibility that patients may be registered at one GP whilst collecting prescriptions issued, and paid for, by another GP practice. To minimise this

possibility registered patient numbers were taken at a single time point towards the start of the period of data analysis in January 2016.

### **5.7.6 Index of Multiple Deprivation (IMD)**

The Index of Multiple Deprivation (IMD) is a measure of the deprivation suffered by the population based on multiple indicators. The indicators used to compose the IMD are: income deprivation; employment deprivation; health deprivation and disability; education, skills and training deprivation; crime; barriers to housing and services; and living environment deprivation. These indicators are periodically assessed in surveys carried out in each of the UK countries and from the results a score for each local area is created then ranked within the country. In England there are 32,844 lower-layer super output areas (LSOAs) each with their own ranking. In Scotland there are 6,942 data zones for the IMD collected in 2016, in Wales the IMD values were last collected in 2014 and there are 1909 LSOAs and in Northern Ireland rankings were last compiled in 2010 at the super output area of which there are 890. As the IMD is based on an, albeit small, area not all people living in highly deprived area will be deprived and conversely there may be people suffering from personal deprivation living in areas assigned a low deprivation ranking (Department for Communities and Local Government, 2015).

### **5.7.7 Index of Multiple Deprivation by GP practice**

The IMD scores were based on the postcode of the GP practice, although this may not necessarily correspond to the IMD of the patients who use the service as patients may travel to the GP practice from areas with a different IMD score. GP practices were ranked by ascending IMD within each country. This ranking was then standardised to percentages relative to the country of origin to provide an overall UK IMD ranking in which indices from the four countries are combined. This was done by taking the ranking for each GP practice and converting it to a percentage by  $(r/N) * 100$  where  $r$  is the IMD ranking of the LSOA and  $N$  is the number of ranked IMDs. Once the rankings for each country have been converted to a percentage they are combined and

then ordered creating a single IMD ranking. IMD is of interest as individuals with lower rankings are thought to have poorer diets and consequently lower levels of iron in their diet than those with a higher ranking (Nelson et al., 2007), this is then hypothesised to have an impact upon the number of iron prescriptions and thus the cost incumbent upon the practice. However the number of prescriptions will be related to the proportion of patients with a lower IMD rank therefore IMD rank at the GP practice level was adjusted by the number of patients at each practice.

### **5.7.8 Prescription of iron medication**

The prescription datasets contain information on all medications prescribed by each GP practice and so to extract iron prescription records, all codes starting with a British National Formulary (BNF) code of 090101 were selected, representing medications used in the treatment of Iron Deficiency Anaemia (IDA) and given either orally or parenterally. These include iron in various forms: ferrous sulphate, ferrous fumarate and ferrous gluconate. Iron is recommended prophylactically for women of child bearing age, in individuals with iron malabsorption and in low birth weight infants (Joint Formulary Committee, 2017).

### **5.7.9 Iron bioavailability**

Estimation of iron bioavailability requires adjustment for the interaction between the iron available in foods and other nutrient components. Rickard et al. (2009) utilised test meal data and physiological knowledge of iron absorption to derive a non-linear expression that estimates available iron in terms of enhancing or inhibiting dietary factors of iron absorption. This was implemented to estimate the amount of available iron consumed per NDNS RP participant. The NDNS RP dataset contains information on the intake of calcium, vitamin C, haem and non-haem iron, along with disaggregated meat, fish and poultry consumption. Tannin values were taken as 30mg per 200ml of tea (Hallberg and Hulthen, 2000) and phytate values were taken from the EWL in-house composition database with foods matched to NDNS RP foods. Not all foods had

a corresponding food identification code and so some phytate values were imputed based on either the closest possible food, the closest possible lower level food group or the closest possible higher level food group, in that order. When scaled in this manner the amount of available iron drops substantially to reflect the various interactions but more accurately represents iron bioavailability. Bioavailable iron is calculated as follows: Bioavailable Iron (mg) = (Percentage available NH × NH) + (0.25 × HI) where

$$\text{Percentage available NH} = 22.42 \times \frac{(1 + \log(1 + 0.0056 \times VC))(1 + \log(1 + 0.0008 \times AT))}{(1 + \log(1 + 0.0008 \times C))(1 + \log(1 + 0.0033 \times P))(1 + \log(1 + 0.0004 \times PO))(1 + \log(1 + 0.0424 \times NH))} \quad (28)$$

VC is vitamin C (mg), AT (animal tissue) is red meat, fish and poultry (g), C is calcium (mg), P is phytate (mg), PO is polyphenols from tea (mg), NH is non-haem iron (mg) and HI is haem iron (mg). Each individual in the NDNS RP was then assigned an iron intake based upon their estimated bioavailability.

**Table 21**

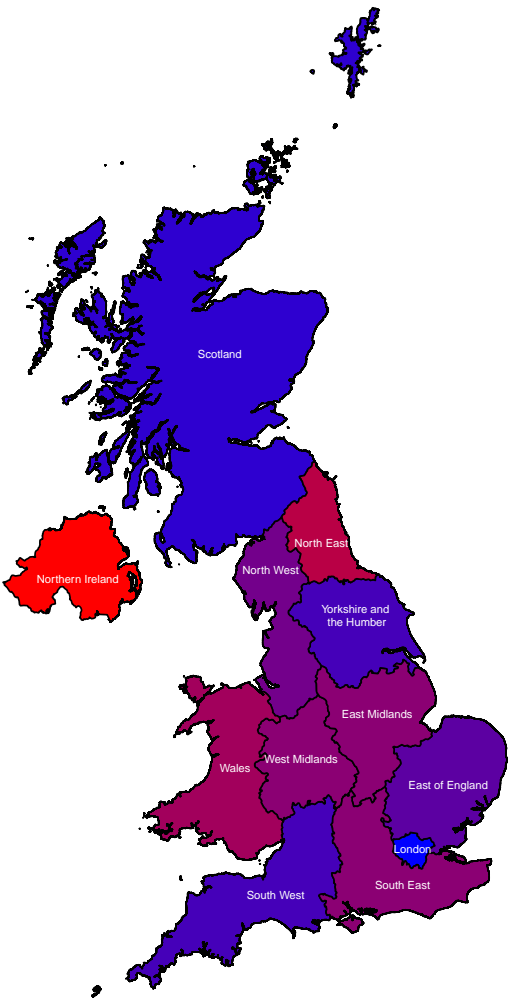
Weighted dietary iron intake adjusted using an algorithm to adjust for concurrent food intake (g/dL).

Sex	Age Group	Mean	SD	Median	IQR
Male	0-3y	2.12	1.39	1.79	1.30
	4-14y	3.40	1.91	3.04	1.98
	15-44y	4.53	3.19	3.74	3.17
	45-64y	4.95	4.02	4.07	3.28
	65+	4.29	2.97	3.60	2.76
Female	0-3y	1.97	1.29	1.72	1.24
	4-14y	2.85	1.50	2.58	1.71
	15-44y	3.82	5.13	2.90	2.44
	45-64y	5.21	9.85	3.44	2.99
	65+	4.00	3.13	3.13	2.43

**Figure 32**

Mean dietary Iron intake by UK government office region. Data taken from NDNS RP Y1-4

- (12.6mg) London
- (11.3mg) Scotland
- (11.2mg) South West
- (10.8mg) Yorkshire and the Humber
- (10.4mg) East of England
- (9.9mg) North West
- (9.7mg) West Midlands
- (9.7mg) East Midlands
- (9.6mg) South East
- (9mg) Wales
- (8.8mg) North East
- (7.1mg) Northern Ireland



### 5.7.10 Dietary Iron

There are discrepancies in dietary bioavailable iron intake across regions within the NDNS RP as shown in **Figure 32** which displays the average dietary iron intake by GOR and shows large differences in iron intake ranging from 7.1mg in Northern Ireland to 12.6mg in London. Furthermore dietary intake varies by age and sex groups, with intakes increasing across age groups and male iron intakes higher than females in all age groups except those aged 45-64y. Daily male iron intake increases from 2.12mg in 0-3y to 4.95mg in those aged 45-64y before reducing slightly to 4.29mg in those aged 65+y. Similarly daily iron intakes in females rose from 1.97mg in the 0-3y age group to 5.21mg in the 45-64y then dropped to 4mg in the eldest age group (**Table 21**). These differences are likely to reflect the difference in total food consumption that increases with age, and also suggests that intakes are skewed somewhat as mean intakes are higher than median values in all cases. As the amount spent upon iron medication is thought to be related to iron status which itself is related to bioavailable dietary iron intake, median iron intakes for age, sex and regional groups were estimated using a linear quantile mixed-effects model (Geraci, 2014; Geraci and Bottai, 2007) with NDNS RP Years 1-4 dietary data.

The distribution of the bioavailable iron as calculated using the algorithm is shown in **Figure 33** (9 values greater than 20mg were removed from the Figure). Clearly bioavailable iron is skewed and has outlier observations, these issues can be accommodated well by use of the linear quantile mixed-effects model. The median regression for bioavailable iron intake was specified as

$$\mu_{ij}^{(0.5)} = \beta_0^{(0.5)} + \beta_1^{(0.5)} \text{Age}_i + \beta_2^{(0.5)} \text{Sex}_i + \beta_3^{(0.5)} \text{Region}_i + u_i,$$

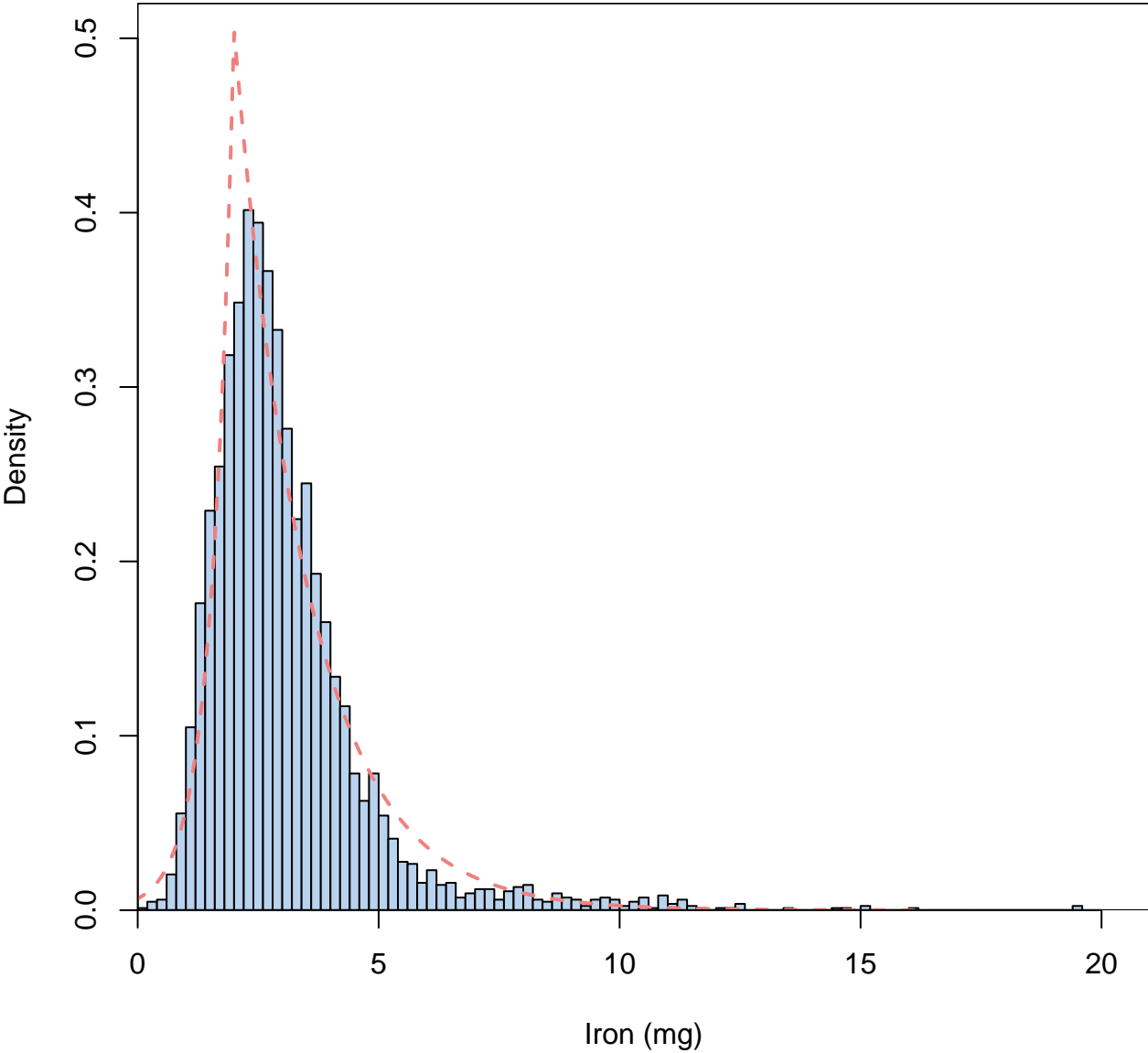
where  $\mu_{ij}^{(0.5)}$  is the median bioavailable iron intake for person  $i$  on day  $j$ ,  $i = 1, \dots, 6828, j = 1, \dots, n_i$ , Age was the individual's age measured in years, Sex is a binary variable (male or female) and region is the government office region the individual lives in. The random intercept  $u_i$  was assumed to have a normal distribution centred at zero.

These median values were used to assign an estimated median dietary iron intake to each GP practice that was reflective of the practice's location and the age and sex



**Figure 33**

Iron (mg) intake as determined through an algorithm that considers nutrient interactions for individuals in the NDNS RP Y1-4 (2008-2012) and fitted asymmetric Laplace distribution (dashed line).



split of its registered patients. The GOR of the GP practice was used along with the percentage of individuals in each age and sex group to estimate a weighted average of the median iron intake from the NDNS RP data. This involved taking the number of patients registered at the GP practice and determining the proportion in each of the 10 age and sex groups. The proportion for each group was then multiplied by the corresponding group median intake. The sum of these values was then taken as a proxy of the iron intake for the GP practice. The median and mean intakes of dietary iron by age and sex are given in Table 21

## 5.8 Statistical analysis

The distribution of the amount spent on iron prescriptions by health board showed a number of large values skewing the distribution to the right (**Figure 34**), because of this median regression is a more appropriate method as it is not impacted by outliers unlike linear regression. Therefore quantile regression based on an asymmetric Laplace distribution (Chapter 4) was used to estimate the median amount spent on iron prescriptions by the  $j^{\text{th}}$  GP practice in the  $i^{\text{th}}$  health board,  $i = 1, \dots, 235$ ,  $j = 1, \dots, n_i$  with the median regression specified by

$$\mu_{ij}^{(0.5)} = \beta_0^{(0.5)} + \beta_1^{(0.5)} \text{BioavailableIron}_{ij} + \beta_2^{(0.5)} \text{IMD}_{ij} + \beta_3^{(0.5)} \text{TotalPatients}_{ij} + u_i,$$

where BioavailableIron is the estimated iron intake for each GP practice, IMD of the GP practice was included due to the relationship with health literacy whereby those in relatively more deprived areas are less likely to seek medical treatment (Rowlands et al., 2013). TotalPatients is the number of patients registered at each GP practice, and IMD is the IMD of the GP practice. The random intercept  $u_i$  was assumed to have a normal distribution centred at zero.

A second median regression model was estimated without bioavailable iron i.e.

$$\mu_{ij}^{(0.5)} = \beta_0^{(0.5)} + \beta_1^{(0.5)} \text{IMD}_{ij} + \beta_2^{(0.5)} \text{TotalPatients}_{ij} + u_i,$$

to assess the impact of the information on iron bioavailability on the parameters estimates.

The models were estimated using the "lqmm" package (Geraci, 2014) in R, standard errors were estimated through bootstrapping with 50 repetitions. A sample of the data is given in **Table 29** in **Appendix L**.

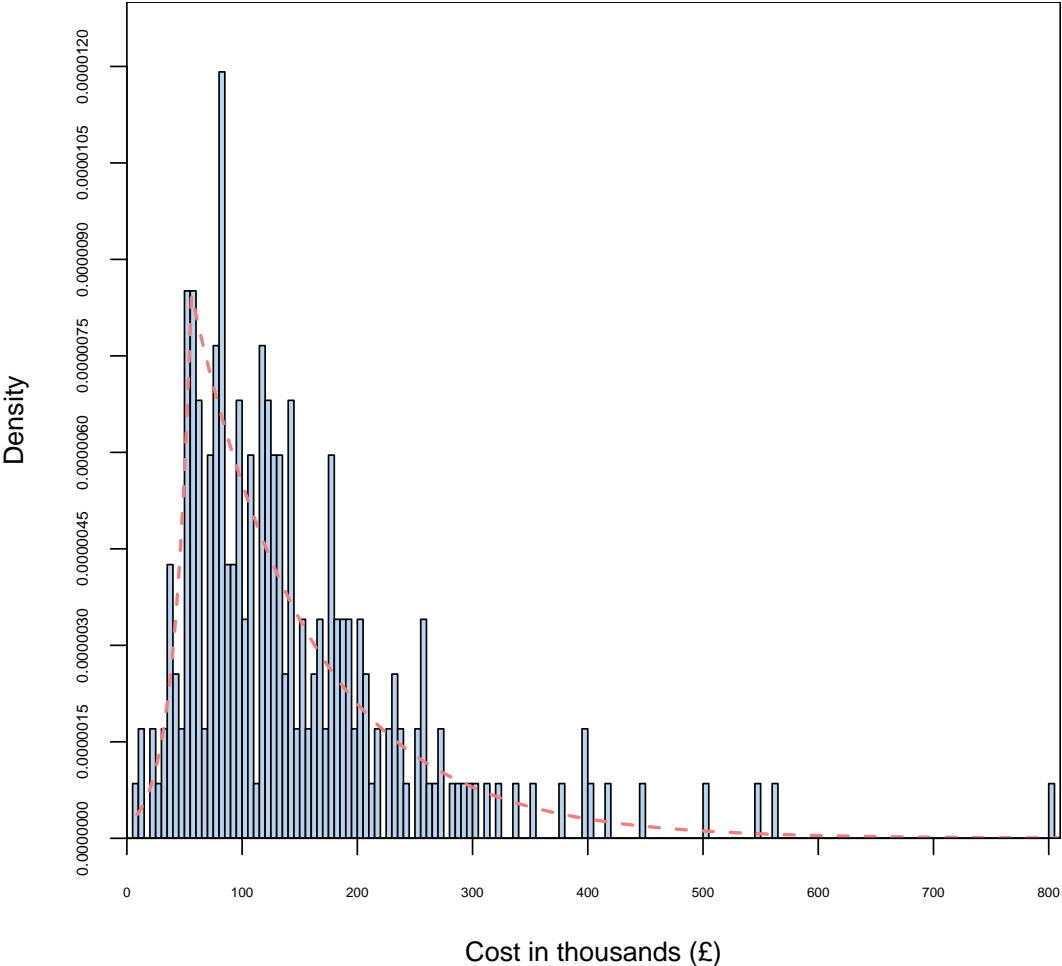
## 5.9 Results

**Figure 35** displays a choropleth of the UK with health boards ranked into quintiles according to the median amount spent upon iron prescriptions, adjusted for total population and the age and sex of registered patients. Predicted expenditure for each health board was determined using the predicted medians then from these, expenditure quintiles were calculated. This figure shows health boards coloured according to their relative expenditures with a darker colour indicating increased spending. A further plot showing the impact that of including bioavailable iron intake is displayed as **Figure 36**. The estimated regression coefficients along with standard errors are displayed in full for both models (excluding and including bioavailable iron) in **Table 30** in **Appendix M**.

The overall median expenditure over the 12 month period per health board was £120,002 (Inter-quartile range: £77,430, £183,700). Table 30 shows that the size of health board as indicated by the total number of registered patients was a small but significant predictor of the amount spent (excluding bioavailable iron: 0.40 95% CI:0.38, 0.41,  $p < 0.001$ ; including bioavailable iron: 0.39 95% CI:0.38, 0.41,  $p < 0.001$ ). An increase in iron intake, as estimated by iron bioavailability, was a strong and significant predictor (-2395.4 95% CI:-2990.4, -1800.4,  $p < 0.001$ ) of reduced spending on iron prescriptions. Also a mild but significant relationship between increasing IMD and a decrease in the amount spent was seen (excluding bioavailable iron: -0.04 95% CI:-0.05, -0.03,  $p < 0.001$ ; including bioavailable iron: -0.04 95% CI -0.04, -0.03,  $p < 0.001$ ). A large number (105) of health boards had statistically significant coefficients determined by  $\alpha < 0.05$  so for brevity those below  $\alpha < 0.001$  will be discussed here. There were 46 health boards with statistically significant coefficients for the amount spent when bioavailable iron was not modelled; of these 36 had negative values indicating

**Figure 34**

Total amount spent on Iron prescriptions by health boards in the UK from Oct 15 - Sept 16  
overlay with an asymmetric Laplace distribution.



low spending and the remainder were higher than expected compared to the reference board: Lincolnshire West, which was chosen as it had the median spending amount of £120,002, of these Rotherham (4619, 95% CI:3121, 6117,  $p<0.001$ ) and Bradford City (4056, 95% CI:1917, 6195,  $p<0.001$ ) were particularly high. In the model including bioavailable iron there were 105 healthboards with significantly different spending on iron compared to the reference. Of these all but 2 had lower than expected spending with Crawley (1626, 95% CI:796, 2455,  $p<0.001$ ) and Ipswich & East Suffolk (2198, 95% CI:1354, 3043,  $p<0.001$ ) having higher than expected spending compared to the reference health board Lincolnshire West.

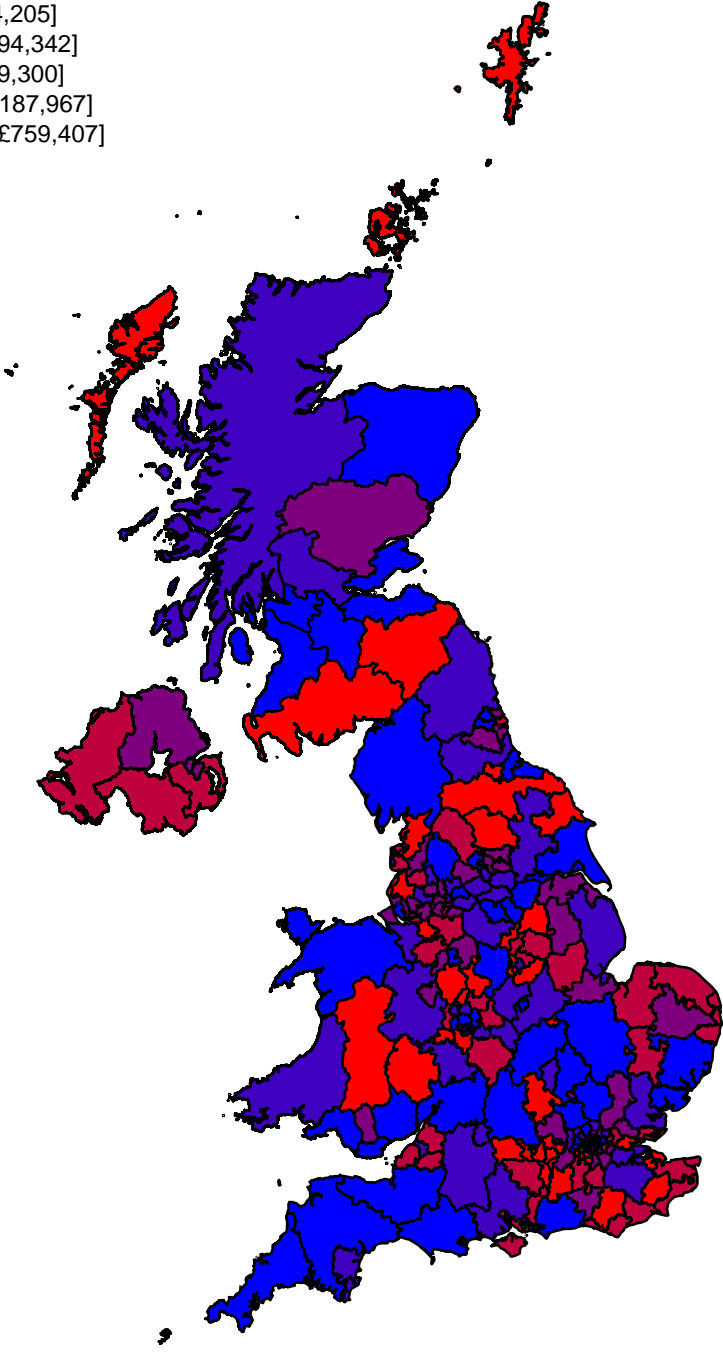
All of the Northern Irish health boards had negative coefficients which equates to a lower than expected spending on iron prescriptions both when bioavailable iron was not included and the estimated spending was lower when bioavailable iron was included. Of the 14 Scottish health boards, 5 reported statistically significant estimates, Borders (-2037, 95% CI:-2634, -1439,  $p<0.001$ ), Dumfries & Galloway (-2025, 95% CI:-2625, -1425,  $p<0.001$ ), Orkney (-1332, 95% CI:-1964, -699,  $p<0.001$ ), Tayside (-2063, 95% CI:-2689, -1437,  $p<0.001$ ) and Western Isles (-1190, 95% CI:-1826, -553,  $p<0.001$ ) were all found to have lower spending compared to the reference health board in the model excluding bioavailable iron. When bioavailable iron was included in the model 12 of the health boards had lower than expected spending on iron prescriptions of these Tayside (-3055, 95% CI:-3562, -2548,  $p<0.001$ ), Borders (-2988, 95% CI:-3537, -2439,  $p<0.001$ ) and Dumfries & Galloway (-2943, 95% CI:-3498, -2388,  $p<0.001$ ) were particularly low. Estimates for the Welsh health boards were statistically significantly different from the reference category in two cases when bioavailable iron was excluded: Cwm Taf (-1523, 95% CI:-2170, -877,  $p<0.001$ ) and Powys (-1469, 95% CI:-2190, -747,  $p<0.001$ ); these two health boards were also the only two out of the six Welsh health boards that had statistically significantly lower spending on iron prescriptions when bioavailable iron was included (Cwm Taf: -1987, 95% CI:-2539, -1435,  $p<0.001$ ) and Powys (-1813, 95% CI:-2366, -1260,  $p<0.001$ ). The variance of the asymmetric Laplace parameter  $\hat{\sigma}$  is given in Table 30 of 570.84 for the model including bioavailable

iron showing slightly larger variance than that from the model excluding bioavailable iron of 570.06.

**Figure 35**

Quintiles of median amount spent from Oct 15 - Sept 16 on iron prescriptions in each health board across the UK adjusted for IMD and the number of registered patients per health board.

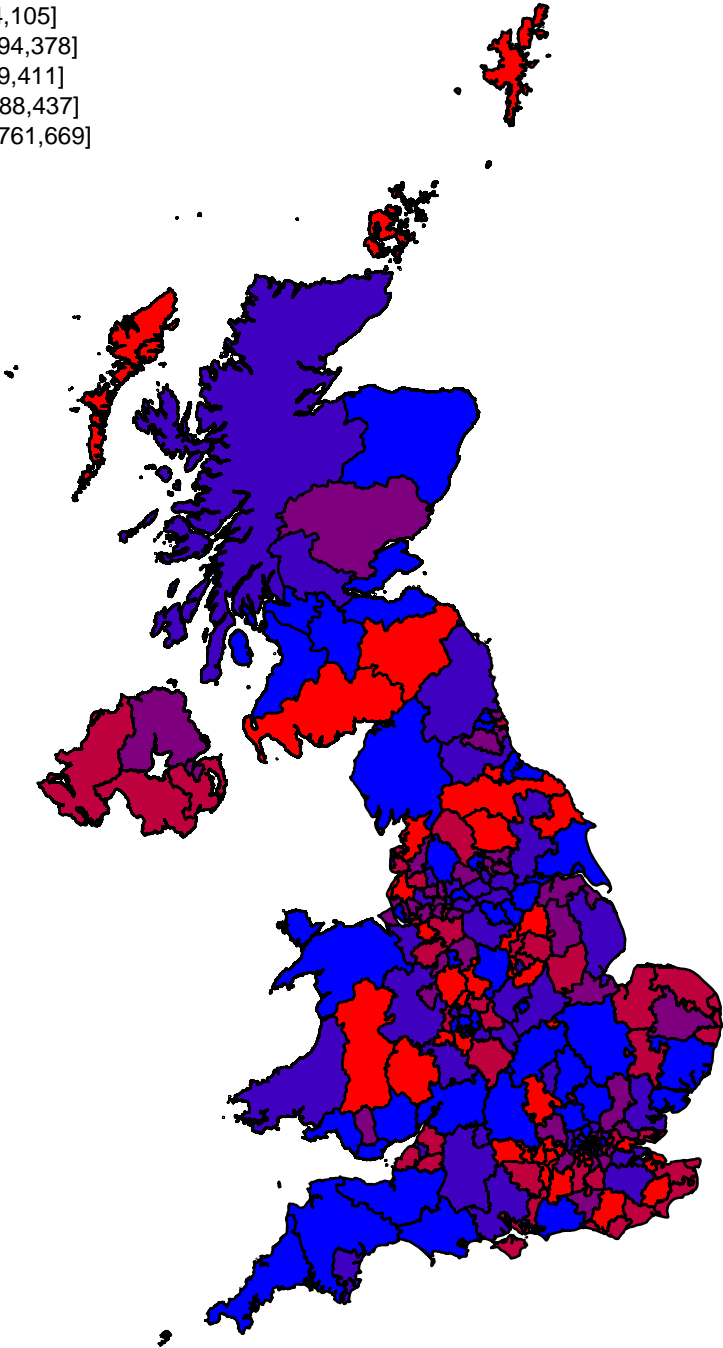
- Lowest Quintile [£6,499 – £64,205]
- Second Quintile (£64,205 – £94,342]
- Third Quintile (£94,342 – £129,300]
- Fourth Quintile (£129,300 – £187,967]
- Highest Quintile (£187,967 – £759,407]



**Figure 36**

Quintiles of median amount spent from Oct 15 - Sept 16 on iron prescriptions in each health board across the UK adjusted for IMD, bioavailable iron intake and the number of registered patients per health board.

- Lowest Quintile [£6,508 – £64,105]
- Second Quintile (£64,105 – £94,378)
- Third Quintile (£94,378 – £129,411)
- Fourth Quintile (£129,411–£188,437)
- Highest Quintile (£188,437–£761,669)





## 5.10 Discussion

This work has produced two maps that show the geographic distribution of current expenditure on prescriptions for iron in the UK by health boards both when the dietary bioavailable iron intake is considered (Figure 36) and when it is not (Figure 35). Presenting the findings in choropleth maps such as these allows the amount of spending by each health board to be easily examined, for example it is possible to see at a glance that spending in Northern Ireland is generally lower than that in the east of England as the colour of Northern Ireland health boards is lighter than those in the east of England. The statistical method used for the analysis of costs considered the clustered structure of the data, with prescriptions issued by GP practices nested within health boards, whilst adjusting for the number of patients registered at each health board, and bioavailable iron intake and deprivation status of registered patients. These methods are widely applicable and can be used to examine inequalities in prescribing rates of any medication prescribed. The findings show that out of the 235 health boards 105 had a significantly different expenditure than the reference board. Previous research that does not consider the impact the health boards has on GP prescribing rates may lead to biased findings. A strength of the current investigation is the inclusion of iron bioavailability as an explanatory variable as this makes the comparison among health boards fair and it improves the predictive power of the model. Each patient consultation results in a cost to the practice and health board, in health boards suffering financial constraints a consultation for iron supplements, that are available over the counter, may be seen as an unnecessary cost that can be minimised. If information regarding the financial status of each health boards in the UK were available then this could be included as an explanatory variable in the model to explain some of the variation in spending. Financial status of the health boards in England is measured using the QOF, which reports that out of the 209 health boards, 52 required improvement and 33 were rated inadequate according to CCG Planning and Assessment National Team (2016) (**Table 22**).

**Table 22**

English health boards financial assessment.

Financial status	Number of health boards
Outstanding	7
Good	117
Requires Improvement	52
Inadequate	33

Of the 7 health boards rated outstanding two spent significantly higher amounts on iron prescriptions when excluding bioavailable iron (Dudley: 1930, 95% CI:1089, 2771,  $p < 0.001$ ; Sandwell & West Birmingham 1916, 95% CI:1018, 2814,  $p < 0.001$ ) and 1 health board spent significantly less than the reference when bioavailable iron was included (Salford: -2213, 95% CI:-2923, -1503,  $p < 0.001$ ). Similarly, health boards that received an inadequate financial rating tended to have spent less on iron prescribing with 5 health boards spending significantly lower amounts on iron prescriptions although 1 health board, Ipswich & East Suffolk, spent significantly more on iron prescriptions compared to the reference in both models excluding and including bioavailable iron.

A strength of this work is the data used as it contained all dispensed prescriptions for iron medication over the 12 month period from October 2015 to September 2016. Along with the age and sex of patients registered at the GP practice and the IMD of the practice itself. This is a comprehensive data set allowing for an accurate description of the variation in the amount spent between Health boards. Furthermore the prescriptions recorded refer to medication that is dispensed to the individual rather than simply medication that is prescribed and not dispensed. All medication dispensed is recorded otherwise the reimbursement would not be received by the issuing practice. This does not ensure that the medication is taken as directed, however, as patients may not complete the course prescribed. Indeed, it is estimated that adherence to medication can be as low as 20% in chronic conditions, that is 80% of individuals do not complete the course of medication (Haynes et al., 2008; Dolce et al., 1991).

As the majority of the UK population are estimated to be registered with a GP this allows for the investigation of iron prescriptions accurately at a national level. The exact percentage of the population registered with a GP and therefore the validity in representing the UK is unclear as the number of people registered with a GP is higher than the estimated population of the UK. The reason for this is unclear but appears to stem from potential over counting by GPs and an under estimate of the UK population due to ambiguity over residency status (House of Commons Library, 2016). The over counting may be due to patients who have died, moved outside of England or are simply registered with another practice. To try to ensure list of registered patients is correct, patients who have not attended the GP surgery for 5 years are contacted and are removed from the practice registry if they fail to respond. This initiative was introduced in 2016 and so should eventually reduce the discrepancy between the population and patient numbers (Primary Care Support England, 2018).

For those who are time poor, buying supplements from the supermarket rather than a pharmacy may offer advantages, however these individuals will not be captured here. Advice to this effect has been issued by health boards to their practices (Mid Essex CCG, 2015; Beechwood Medical Practice, Bristol CCG, 2016; Iacobucci, 2017). The willingness to prescribe medication that may not be necessary or can be purchased over-the-counter could be related to the time and work constraints reported by many GPs. As an example, prescribing antibiotics rather than issuing advice has been found to result in fewer repeat visits and therefore a reduced workload (Little et al., 1997). There is a possibility, however, that if this becomes a policy for all health boards, those unable to access supplements over the counter may not receive the medication required.

Some interesting mapping has been carried out by Rowlingson et al. (2013) who looked at the variation in diabetes and attention deficit hyperactivity disorder medication in England. The resultant maps were finer in resolution than those presented here as the prescribing rate of each of the 8111 GP practices was plotted. This is an alternate approach that allows smoothing to be used to consider the impact of GP practices. However this does not consider the within health board dependency that impacts upon

prescribing practice; instead treating each GP practice as an independent entity and was carried out prior to the establishment of CCGs in England.

## 6 Discussion

This thesis developed and applied novel statistical methods to estimate usual intake of episodically and habitually consumed foods and nutrients collected from national complex surveys, using iron as an exemplar.

### 6.1 Summary of novel methods

The methodological contributions of the thesis include the use, extension and computational implementation of the two-part generalised gamma model and the quantile linear mixed-effects model to accommodate a multistage complex survey design. To the best of my knowledge, this work presents the first application of these models to dietary data collected from complex surveys. The methods were applied to address three different research questions within the field of dietary surveillance, with a focus on the topic of iron deficiency which is one of the most common nutrient deficiencies in the world.

The statistical methods developed in this thesis aimed to estimate the mean dietary intake as a function of explanatory variables and to estimate quantiles of the intake distribution also in terms of explanatory variables. These methods were developed to take into consideration: (i) the shape of the distribution of dietary intake; (ii) the sources of data variation for correct specification of the variance components of the underlying model, and (iii) the data sampling method to allow extrapolation of results into the target population.

Highlighted were three common features of dietary data that can affect the shape of the distribution of intake data collected from national surveys. These included a large number of zero observations for episodically consumed foods; non-symmetric distributions of consumption, partly due to outlier observations which are very common in survey data, and the fact that dietary intake can only take non-negative values. While the traditional methods of analysis based on the normal distribution are convenient and simple to implement, they are inadequate as they cannot handle the excess zeros or

non-symmetric distributions, and the domain of the normal distribution includes negative values. Others had already pointed out these issues, explained the consequences of undertaking a naive analysis based on the standard application of the normal distribution, offered alternative methods of analysis and adapted them to incorporate a multistage sample design (Tooze et al., 2006) (see Section 1.12).

These methods included the use of a two-part model to deal with the high frequency of zero observations found in episodically consumed foods where the first model component is specified to estimate the probability of consumption and the second to estimate the mean amount consumed, given a positive consumption. The first component is typically modelled with a mixed-effects logistic regression model and the second with a linear mixed-effects regression model.

The work presented here addressed the limitations of such methods, including NCI (Tooze et al., 2006, 2010) and SPADE (Dekkers et al., 2014) methods of dietary intake assessment used to analyse the US NHANES and the Dutch DNFCS survey data respectively. The NCI method uses a Box-Cox transformation to deal with the non-symmetric distribution in the second model component; however, this approach requires a back transformation of the results which can complicate their interpretation and hinder the reproducibility of analysis.

In contrast, both the generalised gamma distribution used in the two-part model for the analysis of episodically consumed foods, and the quantile regression model with an asymmetric Laplace distribution for the analysis of habitually consumed foods, offer a wide family of distributions with many different shapes including specific distributions for non-negative data. This allows the location parameter in the model using the generalised gamma distribution and the quantiles in the model using the asymmetric Laplace distribution to be directly expressed as functions of explanatory variables. Thus the proposed methods of analyses in this thesis are straightforward to carry out as no transformations for normality are required; they are based on models that provide good fit to the data; they are reproducible and the interpretation of results is transparent.

An important step in statistical modelling is the incorporation of the sources of data variation into the model. Failure to do this can lead to incorrect specification of the variance components of the model, which leads to bias in the model parameter estimates, incorrect estimated variance of model parameters, and incorrect inferences as a result. These issues have been largely discussed in the statistical literature, for example on methods for the analysis of repeated data and longitudinal data analysis (Diggle et al., 2002). The sources of variability of dietary intake collected from a 4-day diary were separated into between- and within-individual variability, where the latter includes day-to-day variation in food consumption and measurement error.

Again, the prevalent traditional analysis would take the average intake of observations from each individual and analyse these using a linear regression model, thus ignoring the two different sources of variability. This would lead to biased regression estimates and incorrect estimated variance of parameters estimates. Instead, here the within- and between-person variability are incorporated into the two-part model by including a random intercept in each model component to induce a correlation between any two observations taken from each individual. These random effects would represent an individual's propensity to consume, and their propensity to consume greater or smaller amounts of food. It seemed plausible that those who are more likely to consume, say alcohol, would also tend to have greater alcohol consumption on the days of consumption. To accommodate this the two random intercepts were allowed to be correlated. This is an important improvement over the method for estimation of episodically consumed foods presented by SPADE which does not include a correlation between the two parts of their two-part model. As shown in the modelling of intake of iron intake from vegetables (Figure 17 in Section 3.4.1), these correlations were estimated not to be zero. Furthermore, Su et al. (2009) showed that failing to take this correlation into account in a two-part model could lead to biased estimates of the regression parameters.

Similarly, a random effect at the individual level to accommodate the within- and between-individual variation was incorporated into a quantile regression model based on the recently developed linear quantile mixed-effects regression model which uses the asym-

metric Laplace distribution (Geraci and Bottai, 2007; Geraci, 2014). The extension of the statistical methods developed in this thesis to incorporate the complex sample design into the parameters estimation required three steps. Firstly, the likelihood function was multiplied by the survey weightings. Secondly, bootstrap methods were used for the estimation of the variance of the parameters estimates to account for the correlation among observations taken from the same PSU. Finally, the variance estimation of the model parameters was stratified by the survey strata.

### **6.1.1 Software implementation**

The computation implementation of this extension for the two-part model with a generalised gamma distribution was undertaken in SAS, as this program offers flexibility at maximising a user's defined likelihood function using Gaussian quadrature methods for the approximation of the integrals required to integrate out the random effects of the model. Likewise, bootstrap variance estimation was implemented using the same program. The estimation of model parameters in the linear quantile mixed-effects model is also complicated by the presence of random effects and is more computing intensive; however, the R package *lqmm* was used as it provided a robust method of estimation based on Gaussian quadrature numerical methods (Geraci, 2014).

## **6.2 Summary of findings**

Three key areas of research relevant to dietary surveillance were addressed using the methods developed here, and exemplified using data from the NDNS RP and electronic records of iron prescription in the UK. First, the mean of dietary intake across sex, age and socio-economic groups was modelled to assess the current dietary status of a population, using the two-part model with generalised gamma distribution. This was illustrated by modelling the mean consumption of iron intake from selected episodically consumed food groups using data from the NDNS RP in Chapter 3. The analyses showed that females were more likely to consume iron from breakfast cereals and



vegetables than males but consumed smaller amounts, in addition both the probability of consumption and the amount of iron consumed from bread was higher in males than females. All age groups consumed greater amounts of iron from bread, vegetables, and fruit and vegetables compared to those aged 1.5-3y. Of note, was the finding that overall, those in lower NSSEC groups had lower probabilities of consuming iron from fruit and vegetables and consumed lower amounts of iron from fruit and vegetables when compared to those in the higher managerial and professional group.

### **6.2.1 Novel approach compared with the traditional approach: two-part model**

A comparative analysis was undertaken using a survey weighted regression model that did not include a random effect. A number of differences were found between regression parameters that were either statistically significant in the two-part model and not in the survey weighted regression or vice versa, highlighting that under the wrong model assumptions differences in findings are observed. In particular, the amount of iron consumed from vegetables by females was significantly higher compared to males, as indicated by the two-part model, but this was not seen in the survey weighed regression model. Conversely, the intake of iron from fruit in females was significantly different to males in the survey weighted regression model but not in the two-part model. Iron intake from breakfast cereals and fruit and vegetables in those aged 11-18y, differed to the reference group in the two-part model but not in the survey weighted regression model, though the opposite finding was the case for iron intake from fruit in the 11-18y group with the survey weighted regression model reporting a significant difference in the amount consumed, whereas the two-part model found no difference. Differences were also found when comparing intakes by NSSEC groups between the two models with statistically significant differences reported in iron from bread in the two-part model though not in the survey weighted model and differences in intakes found in the survey weighted model for iron from vegetables, fruit, and fruit and vegetables that were not found in the two-part model.

## 6.2.2 Novel approach compared with the traditional approach: Quantile regression

Second, population groups with low or high levels of consumption of certain foods or nutrients were identified using linear quantile mixed-effects models, to inform policy. I demonstrated how quantile regression is well suited to describe the tails of the intake distribution in relationship to factors such as age, sex and socio-economic status. This was illustrated by comparing the quantiles of iron intake with LRNI recommendations using NDNS RP years 1-4 data, in Chapter 4. The analyses showed that older age and intake on a weekend day was associated with higher iron intakes in the 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup> and 97.5<sup>th</sup> quantiles, similarly, males consumed greater amounts of iron than females across all five quantiles (2.5<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup> and 97.5<sup>th</sup>). Iron intakes tended to be associated with NSSEC groups, with lower consumption seen in the intermediate occupations group, the lower supervisory and technical occupations group, semi-routine occupations group, the routine occupations group and those in the never worked group for all five quantiles. Those in the lower managerial and professional occupations group had lower iron intakes for the 2.5<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> quantiles. Also produced, were intake curves of the quantiles that provide a useful tool to quickly assess intake deficiency in a population. Using these plots it is straightforward to add intake thresholds such as the LRNI to visualise the proportion of the population at risk of iron deficiency. A comparison analysis was also presented highlighting the importance of using correct model assumptions. Median iron intakes, as estimated using the quantile regression methods presented here, were compared to mean iron intakes estimated using regression that did consider the between- and within-person variation and the survey weighting, though not the other elements of the complex survey design nor the skewed distribution of the data and, as a result of this misspecification, differences in the results were seen as significantly lower iron intake was reported for those in the semi routine occupations group by the comparator weighted regression mixed-effects model but not in the quantile regression case.

### **6.2.3 Expenditure on iron prescriptions by health boards**

Finally, the distribution of expenditure on iron deficiency medication across health boards in the UK was investigated as this can highlight inequalities and inform policy. This analysis was based on the linear quantile mixed-effects model. Expenditure on supplementation is likely to depend on the nutrient status of the population. For a fair comparison of expenditure among health boards, the methods developed in Chapter 4 were used to estimate nutrition status and this information was incorporated into a model of expenditure on prescriptions for iron across health boards in the UK. This analysis found that the number of patients registered at each health board was associated with an increase in spending on iron prescriptions, and that a higher index of multiple deprivation ranking along with an increase in regional iron intake were associated with a decrease in spending. In total 36 health boards had significantly lower spending and 10 were significantly higher ( $\alpha < 0.001$ ) than the reference health board, Lincolnshire west. In addition, choropleth maps were presented, coloured according to quintiles of spending on iron prescriptions. From these, for example 35, the lower than expected spending rates in Scotland and Northern Ireland are apparent along with the higher than expected spending in a belt across England, north of the midlands, and in Cornwall.

### **6.2.4 Implications of findings**

The results of comparisons between the methods developed here and the approaches that are currently used have shown statistically significant associations in the current methods that do not occur in the novel methods and vice versa. The traditional within-person mean approach is used in the NDNS RP to estimate intakes with the findings from the report used to inform policy. For example NDNS RP data was used in the Healthy Lives, Healthy People white paper (Department of Health, 2010) to recommend a reduction in sodium intake which led to the public health campaign around reducing sodium intakes, based on the association between sodium intake and heart disease, however the shape of this association has been shown to depend on the anal-

ysis method (He et al., 2018) highlighting the importance of correct methods. Sodium is a habitually consumed nutrient. The methods of analysis presented in Chapter 3 can be easily adapted to model intake of habitually consumed nutrients. Specifically, the amount of food consumed can be modelled using a mixed-effects regression model with a generalised gamma distribution. This modelling approach corresponds to using the second component of the two-part model described in Chapter 3 but without conditioning on whether there was consumption or not. The advantage of this method over standard analysis, is that it accounts for within-person variation in consumption, through the inclusion of a random intercept, and for the skewed distribution commonly observed in dietary intakes. This is achieved by using the generalised gamma distribution. Model misspecification can affect both the magnitude and the standard error of the regression parameters estimates. This has been demonstrated in the literature describing methods of analysis of clustered data.

### **6.3 Strengths and limitations of this research**

The strengths of this thesis include the development and use of contemporaneous statistical models to provide reliable and robust estimates of mean and quantile dietary intake collected from complex survey data. The models were implemented in R and SAS and code is provided for their implementation (Appendices G and I). The statistical analyses of iron intake and iron prescription expenditure demonstrated the utility of the methods, while addressing important dietary surveillance questions concerning iron deficiency. The analyses were based on the NDNS RP data which are the data of highest quality available on dietary consumption in the UK. The methods are generalizable to the analysis of data from any national survey with multistage sampling and readily available for implementation. There are some limitations of this work, the proposed methods can be computing intensive due to the bootstrapping sampling, however, compared to the effort required to undertake a national survey, this is a relatively minor disadvantage. The methods presented here assume that dietary intake is measured without systematic error and that any bias produced in the estimate of

the dietary component is random in nature and is due to an insufficient number of recorded measures. Thus increasing the number of diary days will minimise the difference between an individual's reported and true intake. It has been demonstrated that systematic error does indeed exist as energy intakes have been found to be lower than expected, compared to an objective measure of energy intake. The data used throughout this thesis were taken from years 1-4 of the NDNS RP and in years 1 and 3, energy intake was reported to be lower than DLW measured energy expenditure in all age groups. This ranged from 11% lower in females aged 4-10 to 36% lower in Females aged 16-49y (Lennox et al., 2014). Similarly, in a previous NDNS report surveying adults aged 19-64y (Henderson et al., 2004), energy intakes were compared to those measured using DLW and found that underreporting also occurred (Rennie et al., 2007). It was found that 75% of men and 77% of women overall were classified as under reporting energy intake. It was also found that the level of under reporting was found to be higher in obese participants suggesting that weight has an important role in the bias associated with reported intakes. Slightly lower, though still substantial, levels of under reported energy intake has been seen over the course of the NHANES (Archer et al., 2013). Using statistical methods such as those proposed here will reduce the assumed random bias; however, it is apparent that the systematic element remains and that dietary assessment methods need to evolve before a truly unbiased usual intake can be known.

Estimates of consumption based on consumers will often be unable to distinguish between those who are non-consumers for the period of reporting and those who will never consume the food regardless of how many days of intake are collected. The NDNS RP has recorded a limited number of questions regarding annual consumption of foods in the Computer Assisted Personal Interview (CAPI) interview which relate to the following food groups: Meat; poultry; Fish; Eggs; dairy products; Salad vegetables; Cooked green vegetables; Root vegetables; fruit; Nuts; Offal along with 16 foods which were chosen due to the high content of certain nutrients. As the questions only relate to a small section it is not possible to use this information to attempt to validate overall diary quality and, in addition, the quality of the data can be poor as the answers given

by some participants are contradictory. For example, approximately 5% of respondents stated that they avoided meat and fish consumption in the past 12 months yet recorded consuming these foods in the diary. Other methods recognise the advantage of including information on long term intake to distinguish between non- and never-consumers. The ISU method allows for adjustment where the reporting of food consumption is related to the collection day. They showed that NHANES participants were more likely to report higher intakes on the first day of data collection than on the second. This adjustment however can only be made in the probability part of their method and not in the amount part of the two-part model.

## **6.4 Comparison to previous results**

Alternate approaches to estimating linear quantile mixed-effects models in R have been proposed (Galarza and Lachos, 2015; Galarza et al., 2015). This is a method of quantile estimation using likelihood based inference determined by an EM algorithm in the qrLMM function in the qrLMM package that provides regression estimates using a Stochastic Approximation of the EM algorithm (SAEM) (Delyon et al., 1999). Using an expectation-maximisation algorithm offers a more precise estimation of the regression parameters and convergence occurs with fewer Monte Carlo EM samples ( $\leq 10$ ) than would be required for typical Monte Carlo EM (Meza et al., 2012). A comparison between fitting the linear quantile mixed-effects model using lqmm and qrLMM has been carried out (Galarza and Lachos, 2015), reporting lower root mean squared error with the qrLMM method compared to estimates from the lqmm method, particularly when estimating extreme quantiles. However, an important limitation of the qrLMM function is that currently, sampling weights cannot be included in the model. A further method of quantile estimation has been proposed using M-quantile regression, which has recently been extended to include a random effect (Tzavidis et al., 2016), although the interpretation of the linear quantile mixed-effects model is not synonymous with that of the M-quantile regression. Moreover, running a simple weighted model took much longer than the comparative linear quantile mixed-effects model fitted using lqmm and

returned non-convergence warnings using the R code kindly shared by Tzavidis et al. (2016). Because of the non-convergence warnings, the results were not considered for comparison. Similar to the qrLMM package, currently sampling weights cannot be included and random coefficients cannot be included along with the random intercept, unlike in the lqmm package.

In this thesis, the use of the lqmm package was coupled with the survey package to provide the first extension of quantile regression with random effects that can provide variance estimates for survey data with multistage clustering and sampling weights in a semi-parametric framework. The survey package (Lumley, 2014) allows for many different complex sampling designs and can include a finite population correction if required. Furthermore, there is flexibility in the resampling method with Jackknife, Balanced Repeated Replication (BRR), and Fay's modified BRR (Judkins, 1990), bootstrap (Canty and Davison, 1999) n-1 bootstrap and multistage rescaled bootstrap (Preston, 2009) all easily implementable allowing these methods to be used in the analysis of the NDNS RP dietary data. These methods can be extended with the implementation of a model selection criterion such as AIC or BIC, however, these are not easily available as they involve estimation of the precision matrix.

In the methods developed in this thesis, the variance estimation at the PSU level of the data sampled under multistage sampling has been carried out by bootstrap resampling, alternate approaches exist but bootstrap resampling has the advantage of flexibility as it can be used in almost all cases. The software used to estimate usual intakes in the US, the NCI method, uses BRR (Kish and Frankel, 1970) to estimate standard errors. BRR requires a balanced survey design that includes 2 PSU per stratum and works by selecting first one PSU per stratum, performing the analysis, then taking the remaining PSU in the stratum and the analysis is carried out once more with the average of the 2 runs taken. The NDNS RP does not follow this survey design, as the PSUs are not balanced and there are often an odd number of PSUs per stratum. A modification to BRR, known as the Grouped Balanced Half Sample (GBHS) method (Rao and Shao, 1996) has been implemented in the Brazilian national diet survey (BRASIL IBGE, 2011) which also does not have a survey design with 2 PSU per stratum. This method of standard

error estimation randomly assigns the PSU with the stratum into two groups and then carries out the BRR estimation. It is worth noting that were BRR to be carried out using the GBHS method standard error estimation would still take significantly longer than a similar estimate using NDNS RP data. This is due to the large difference in the number of strata between the two surveys with 16 strata in NHANES survey design versus approximately 700 strata in the NDNS RP years 1-4 design. Using BRR for standard error estimation would require the number of repetitions to be fixed at 700 whereas using bootstrap the number of iterations required can be varied and, as demonstrated in Section 3.5, 50 bootstrap repetitions is suitable for obtaining standard errors with high precision.

## **6.5 Areas of future research**

Section 5.7.9 discussed the complexities of food and nutrient interactions providing an example where the bioavailability of iron is impacted by, amongst other food components, the amount of calcium and vitamin C consumed simultaneously. These interactions occur throughout the diet between nutrients and food components and foods, both at biological level, as discussed, and also when an individual makes choices over what to consume at a given meal. This leads to a multitude of possible adjustments required when modelling usual intake, in spite of the complexity methods of modelling intakes from multiple foods have been proposed. One approach to estimate usual intake of the healthy eating index has been presented by Zhang et al. (2011) that is capable of presenting estimates for a combination of episodically and habitually consumed food components therefore presenting an overall indication of diet quality although the complexity of its implementation means that it is unlikely to be used by all (Carriquiry, 2017). Extending the methods presented here to estimate intake that gives an overall measure of usual intake, considering the interactions that occur with other dietary components is a challenging area of future research.

A further area of future work is the distribution of these methods for use by those estimating usual intakes. As discussed in Section 1.5.2, a significant number of authors



(Adams and White, 2015; McGeoghegan et al., 2015; Murakami and Livingstone, 2016; Syrad et al., 2016; Ziauddeen et al., 2017; Hobbs et al., 2018), do not fully consider the challenges of working with dietary data collected under a multistage sampling plan; this, presumably, is in part due to the lack of easy to implement methods to properly carry out this analysis. Therefore, whilst this thesis has developed and presented techniques to overcome these challenges, ensuring that they are accessible to researchers is of importance. Future work to this end involves the development and publication of an R package containing the methods currently implemented in R (Chapters 4 and 5), with the hope of implementing in R the work presented in Chapter 3 that is currently implemented in SAS. Further methodological developments include the consideration of alternative distributions to the generalised gamma distributions for the analysis of intake in the two part model, and extension of the two part model to accommodate a group population that never consume.

## **7 Conclusion**

In summary, this thesis presented novel approaches to the analysis of dietary intake collected using multistage sampling. The methods were carefully designed to provide models of good fit to the data, account for the data variability and the sampling design. The utility of the methods was demonstrated by addressing three common research questions arising from dietary surveillance: analysing the mean dietary status of the target population, identifying groups of low or high dietary consumption, and analysing the distribution of prescription expenditure in a country, informed by nutrient status. The computational implementation of these methods was also provided to make them readily available.

## References

- Abdulla, M., Andersson, I., Asp, N., Berthelsen, K., Birkhed, D., Dencker, I., Johansson, C. G., Jägerstad, M., Kolar, K., Nair, B. M., Nilsson-Ehle, P., Nordén, A., Rassner, S., Akesson, B., and Ockerman, P. A. (1981). Nutrient intake and health status of vegans. Chemical analyses of diets using the duplicate portion sampling technique. *American Journal of Clinical Nutrition*, 34(11):2464–2477.
- Adams, J. and White, M. (2015). Characterisation of UK diets according to degree of food processing and associations with socio-demographics and obesity: cross-sectional analysis of UK National Diet and Nutrition Survey (2008–12). *The International Journal of Behavioral Nutrition and Physical Activity*, 12.
- Aerts, M. and Claeskens, G. (1999). Bootstrapping Pseudolikelihood Models for Clustered Binary Data. *Annals of the Institute of Statistical Mathematics*, 51(3):515–530.
- Allen, L. H. (2002). Iron Supplements: Scientific Issues Concerning Efficacy and Implications for Research and Programs. *Journal of Nutrition*, 132(4):813S–819S.
- Allen, N. B., Holford, T. R., Bracken, M. B., Goldstein, L. B., Howard, G., Wang, Y., and Lichtman, J. H. (2010). Geographic variation in one-year recurrent ischemic stroke rates for elderly Medicare beneficiaries in the USA. *Neuroepidemiology*, 34(2):123–129.
- Andrews, D. W. and Buchinsky, M. (1996). On the Number of Bootstrap Repetitions for Bootstrap Standard Errors, Confidence Intervals, and Tests. *Cowles Foundation Discussion Papers 1141R*, Cowles Foundation for Research in Economics, Yale University.
- Annweiler, C., Rolland, Y., Schott, A. M., Blain, H., Vellas, B., Herrmann, F. R., and Beauchet, O. (2012). Higher Vitamin D Dietary Intake Is Associated With Lower Risk of Alzheimer’s Disease: A 7-Year Follow-up. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 67(11):1205–1211.

- Archer, E., Hand, G. A., and Blair, S. N. (2013). Validity of U.S. Nutritional Surveillance: National Health and Nutrition Examination Survey Caloric Energy Intake Data, 1971–2010. *PLOS ONE*, 8(10):e76632.
- Ashton, C. M., Petersen, N. J., Soucek, J., Menke, T. J., Yu, H. J., Pietz, K., Eigenbrodt, M. L., Barbour, G., Kizer, K. W., and Wray, N. P. (1999). Geographic variations in utilization rates in Veterans Affairs hospitals and clinics. *New England Journal of Medicine*, 340(1):32–39.
- Asparouhov, T. (2006). General Multi-Level Modeling with Sampling Weights. *Communications in Statistics - Theory and Methods*, 35(3):439–460.
- Barroso, F., Allard, S., Kahan, B. C., Connolly, C., Smethurst, H., Choo, L., Khan, K., and Stanworth, S. (2011). Prevalence of maternal anaemia and its predictors: a multi-centre study. *European Journal of Obstetrics, Gynecology, and Reproductive Biology*, 159(1):99–105.
- Bates, B., Lennox, A., Prentice, A., Bates, C., Page, P., Nicholson, S., Milne, A., and Swan, G. (2014a). National Diet and Nutrition Survey Rolling Programme (NDNS RP). Results from Years 1–4 (combined) for Scotland (2008/9-2011/12). Technical report.
- Bates, B., Lennox, A., Prentice, A., Bates, C., Page, P., Nicholson, S., and Swan, G. (2014b). National Diet and Nutrition Survey. Results of the National Diet and Nutrition Survey (NDNS) rolling programme for 2008 and 2009 to 2011 and 2012. Technical report.
- Bates, C. J., Prentice, A., Cole, T. J., van der Pols, J. C., Doyle, W., Finch, S., Smithers, G., and Clarke, P. C. (1999). Micronutrients: highlights and research challenges from the 1994-5 National Diet and Nutrition Survey of people aged 65 years and over. *British Journal of Nutrition*, 82(1):7–15.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014c). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*.

- Beard, J. (2007). Recent Evidence from Human and Animal Studies Regarding Iron Status and Infant Development. *The Journal of Nutrition*, 137(2):524S–530S.
- Beechwood Medical Practice, Bristol CCG (2016). *Changes to prescribing at this Surgery*.
- Bingham, S. A. (1987). The Dietary Assessment of Individuals; Methods, Accuracy, New Techniques and Recommendations. *Nutrition Abstracts Review*, 57(2):705–742.
- Bingham, S. A. (1997). Dietary assessments in the European prospective study of diet and cancer (EPIC). *European journal of cancer prevention: the official journal of the European Cancer Prevention Organisation (ECP)*, 6(2):118–124.
- Bingham, S. A., Gill, C., Welch, A., Cassidy, A., Runswick, S. A., Oakes, S., Lubin, R., Thurnham, D. I., Key, T. J., Roe, L., Khaw, K. T., and Day, N. E. (1997). Validation of dietary assessment methods in the UK arm of EPIC using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin C and carotenoids as biomarkers. *International Journal of Epidemiology*, 26(suppl 1):S137.
- Biro, G., Hulshof, K. F. A. M., Ovesen, L., Amorim Cruz, J. A., and EFCOSUM Group (2002). Selection of methodology to assess food intake. *European Journal of Clinical Nutrition*, 56 Suppl 2:S25–32.
- Black, A. E. and Cole, T. J. (2001). Biased over- or under-reporting is characteristic of individuals whether over time or by different assessment methods. *Journal of the American Dietetic Association*, 101(1):70–80.
- Bondell, H. D., Reich, B. J., and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838.
- Box, G. E. P. and Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.
- BRASIL IBGE (2011). Pesquisa de orçamentos familiares 2008–2009: Análise do consumo alimentar pessoal no Brasil. *Rio de Janeiro: IBGE*.

- Buchowski, M. S. (2014). Doubly Labeled Water Is a Validated and Verified Reference Standard in Nutrition Research. *Journal of Nutrition*, 144(5):573–574.
- Buck, R. J., Hammerstrom, K. A., and Ryan, P. B. (1995). Estimating long-term exposures from short-term measurements. *Journal of Exposure Analysis and Environmental Epidemiology*, 5(3):359–373.
- Cade, J. E., Burley, V. J., Warm, D. L., Thompson, R. L., and Margetts, B. M. (2004). Food-frequency questionnaires: a review of their design, validation and utilisation. *Nutrition Research Reviews*, 17(01):5–22.
- Cancelo-Hidalgo, M. J., Castelo-Branco, C., Palacios, S., Haya-Palazuelos, J., Ciria-Recasens, M., Manasanch, J., and Pérez-Edo, L. (2013). Tolerability of different oral iron supplements: a systematic review. *Current Medical Research and Opinion*, 29(4):291–303.
- Canty, A. J. and Davison, A. C. (1999). Resampling-Based Variance Estimation for Labour Force Surveys. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 48(3):379–391.
- Carrquiry, A. L. (2003). Estimation of Usual Intake Distributions of Nutrients and Foods. *Journal of Nutrition*, 133(2):601S–608S.
- Carrquiry, A. L. (2017). Understanding and Assessing Nutrition. *Annual Review of Statistics and Its Application*, 4(1):123–146.
- Carroll, R. J. (2014). Estimating the Distribution of Dietary Consumption Patterns. *Statistical Science*, 29(1):2–8.
- CCG Planning and Assessment National Team (2016). CCG Assurance Annual Assessment 2015/16. Technical report.
- CDC and National Center for Health Statistics (2013). NHANES - Continuous NHANES Web Tutorial - Sample Design.

- Chen, H.-C., Jia, W., Yue, Y., Li, Z., Sun, Y.-N., Fernstrom, J. D., and Sun, M. (2013). Model-based measurement of food portion size for image-based dietary assessment using 3d/2d registration. *Measurement Science and Technology*, 24(10):105701.
- Cheng, C.-L., Chen, Y.-C., Liu, T.-M., and Kao Yang, Y.-H. (2011). Using spatial analysis to demonstrate the heterogeneity of the cardiovascular drug-prescribing pattern in Taiwan. *BMC Public Health*, 11:380.
- Conway, J. M., Ingwersen, L. A., Vinyard, B. T., and Moshfegh, A. J. (2003). Effectiveness of the US Department of Agriculture 5-step multiple-pass method in assessing food intake in obese and nonobese women. *The American journal of clinical nutrition*, 77(5):1171–1178.
- Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, pages 829–844.
- Danesh, J. and Appleby, P. (1999). Coronary heart disease and iron status: meta-analyses of prospective studies. *Circulation*, 99(7):852–854.
- Darmon, N. and Drewnowski, A. (2008). Does social class predict diet quality? *The American Journal of Clinical Nutrition*, 87(5):1107–1117.
- Dekkers, A. L. M., Verkaik-Kloosterman, J., Rossum, C. T. M. v., and Ocké, M. C. (2014). SPADE, a New Statistical Program to Estimate Habitual Dietary Intake from Multiple Food Sources and Dietary Supplements. *Journal of Nutrition*, 144(12):2083–2091.
- Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a Stochastic Approximation Version of the EM Algorithm. *Annals of Statistics*, 27(1):94–128.
- Department for Communities and Local Government (2015). The English Indices of Deprivation. Technical report.
- Department of Health (2010). *Healthy lives, healthy people: our strategy for public health in England*, volume 7985. The Stationery Office.

- Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., and Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.
- Dodd, K. (1999). Estimation of a distribution function from survey data. *Retrospective Theses and Dissertations*.
- Dodd, K. W. (2011). Accounting for complex survey design in modeling usual intake.
- Dolce, J. J., Crisp, C., Manzella, B., Richards, J. M., Hardin, J. M., and Bailey, W. C. (1991). Medication adherence patterns in chronic obstructive pulmonary disease. *Chest*, 99(4):837–841.
- Domanski, M., Antman, E. M., McKinlay, S., Varshavsky, S., Platonov, P., Assmann, S. F., and Norman, J. (2004). Geographic variability in patient characteristics, treatment and outcome in an International Trial of Magnesium in acute myocardial infarction. *Controlled Clinical Trials*, 25(6):553–562.
- Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983). A comparison of alternative models for the demand for medical care. *Journal of business & economic statistics*, 1(2):115–126.
- Dubnicka, S. R. (2007). A confidence interval for the median of a finite population under unequal probability sampling: A model-assisted approach. *Journal of Statistical Planning and Inference*, 137(7):2429–2438.
- Dwyer, J., Picciano, M. F., and Raiten, D. J. (2003). Collection of Food and Dietary Supplement Intake Data: What We Eat in America - NHANES. *Journal of Nutrition*, 133(2):590S–600S.
- Finch, S. (1998). *National Diet and Nutrition Survey: people aged 65 years and over*. Stationery Office.
- Fitt, E., Cole, D., Ziauddeen, N., Pell, D., Stickley, E., Harvey, A., and Stephen, A. M. (2014). DINO (Diet In Nutrients Out) - an integrated dietary assessment system. *Public Health Nutrition*, pages 1–8.

- Food and Nutrition Board, Coordinating Committee on Evaluation of Food Consumption Surveys, Subcommittee on Criteria for Dietary Evaluation, National Research Council, Division on Earth and Life Studies, and Commission on Life Sciences (1986). *Nutrient Adequacy: Assessment Using Food Consumption Surveys*. National Academies Press.
- Food Standards Agency (1988). *Food Portion Sizes*. TSO, London, 3rd edition.
- Food Standards Agency (2002). *McCance and Widdowson's The Composition of Foods*. Royal Society of Chemistry, Cambridge, 6th edition.
- Food Standards Agency (2008). The Bread and Flour Regulations 1998. Technical report.
- Forouhi, N. G., Harding, A. H., Allison, M., Sandhu, M. S., Welch, A., Luben, R., Bingham, S., Khaw, K. T., and Wareham, N. J. (2007). Elevated serum ferritin levels predict new-onset type 2 diabetes: results from the EPIC-Norfolk prospective study. *Diabetologia*, 50(5):949–956.
- Francisco, C. A. and Fuller, W. A. (1991). Quantile Estimation with a Complex Survey Design. *The Annals of Statistics*, 19(1):454–469.
- Galarza, C. and Lachos, H. V. (2015). *qrLMM: Quantile Regression for Linear Mixed-Effects Model*.
- Galarza, C. E., Bandyopadhyay, D., and Lachosa, V. H. (2015). Quantile Regression for Linear Mixed Models: A Stochastic Approximation EM approach.
- Gasche, C., Lomer, M. C. E., Cavill, I., and Weiss, G. (2004). Iron, anaemia, and inflammatory bowel diseases. *Gut*, 53(8):1190–1197.
- Geraci, M. (2013). Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Statistical Methods in Medical Research*, page 0962280213484401.
- Geraci, M. (2014). Linear Quantile Mixed Models: The lqmm Package for Laplace Quantile Regression. *Journal of Statistical Software*, 57(13).



- Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8(1):140–154.
- Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing*, 24(3):461–479.
- Geraci, M. and Salvati, N. (2007). The geographical distribution of the consumption expenditure in Ecuador: Estimation and mapping of the regression quantiles. *Statistica Applicata - Italian Journal of Applied Statistics*, 19:167–183.
- Golub, G. H. and Welsch, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of computation*, 23(106):221–230.
- Gregory, J., Foster, K., Tyler, H., and Wiseman, M. (1990). *The dietary and nutritional survey of British adults*. HMSO Publications Centre.
- Gregory, J. R., Collins, D. L., Davies, P. S. W., Hughes, J. M., and Clarke, P. C. (1995). *National Diet and Nutrition Survey: children aged 1.5 to 4.5 years*. HMSO Publications Centre.
- Guenther, P. M., Dodd, K. W., Reedy, J., and Krebs-Smith, S. M. (2006). Most Americans Eat Much Less than Recommended Amounts of Fruits and Vegetables. *Journal of the American Dietetic Association*, 106(9):1371–1379.
- Haas, J. D. and Brownlie, T. (2001). Iron Deficiency and Reduced Work Capacity: A Critical Review of the Research to Determine a Causal Relationship. *The Journal of Nutrition*, 131(2):676S–690S.
- Haines, P. S., Hama, M. Y., Guilkey, D. K., and Popkin, B. M. (2003). Weekend Eating in the United States Is Linked with Greater Energy, Fat, and Alcohol Intake. *Obesity Research*, 11(8):945–949.
- Hallberg, L. and Hulthen, L. (2000). Prediction of dietary iron absorption: an algorithm for calculating absorption and bioavailability of dietary iron. *The American Journal of Clinical Nutrition*, 71(5):1147–1160.

- Haubrock, J., Nöthlings, U., Volatier, J.-L., Dekkers, A., Ocké, M., Harttig, U., Illner, A.-K., Knüppel, S., Andersen, L. F., Boeing, H., and European Food Consumption Validation Consortium (2011). Estimating usual food intake distributions by using the multiple source method in the EPIC-Potsdam Calibration Study. *Journal of Nutrition*, 141(5):914–920.
- Haynes, R. B., Ackloo, E., Sahota, N., McDonald, H. P., and Yao, X. (2008). Interventions for enhancing medication adherence. *The Cochrane Database of Systematic Reviews*, (2):CD000011.
- He, F. J., Campbell, N. R. C., Ma, Y., MacGregor, G. A., Cogswell, M. E., and Cook, N. R. (2018). Errors in estimating usual sodium intake by the Kawasaki formula alter its relationship with mortality: implications for public health. *International Journal of Epidemiology*.
- Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 475–492. NBER.
- Heckman, J. J. (1977). *Sample selection bias as a specification error (with an application to the estimation of labor supply functions)*. National Bureau of Economic Research Cambridge, Mass., USA.
- Henderson, L., Gregory, J., Irving, K., and Swan, G. (2004). The National Diet & Nutrition Survey: adults aged 19 to 64 years. *Energy, protein, fat and carbohydrate intake, 2*.
- Herrick, K. A., Rossen, L. M., Parsons, R., and Dodd, K. W. (2018). Estimating Usual Dietary Intake From National Health and Nutrition Examination Survey Data Using the National Cancer Institute Method. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (178):1–63.

- Hoare, J., Henderson, L., Bates, C. J., Prentice, A., Birch, M., Swan, G., and Farron, M. (2004). *The national diet & nutrition survey: adults aged 19 to 64 years. Summary report*. Office For National Statistics London, UK.
- Hobbs, D. A., Givens, D. I., and Lovegrove, J. A. (2018). Yogurt consumption is associated with higher nutrient intake, diet quality and favourable metabolic profile in children: a cross-sectional analysis using data from years 1–4 of the National diet and Nutrition Survey, UK. *European Journal of Nutrition*, pages 1–14.
- Hoffmann, K., Boeing, H., Dufour, A., Volatier, J. L., Telman, J., Virtanen, M., Becker, W., De Henauw, S., and EFCOSUM Group (2002). Estimating the distribution of usual dietary intake by short-term measurements. *European Journal of Clinical Nutrition*, 56 Suppl 2:S53–62.
- House of Commons Library (2016). Population estimates & GP registers: why the difference?
- Huang, J., Sun, J., Li, W. X., Wang, L. J., Huo, J. S., Chen, J. S., Chen, C. M., and Wang, A. X. (2009). Efficacy of Different Iron Fortificants in Wheat Flour in Controlling Iron Deficiency. *Biomedical and Environmental Sciences*, 22(2):118–121.
- Huang, X. (2003). Iron overload and its association with cancer risk in humans: evidence for iron as a carcinogenic metal. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 533(1–2):153–171.
- Iacobucci, G. (2017). Doctors call for national rules on OTC prescribing. *BMJ*, 356:j1442.
- Jackson, K. A., Byrne, N. M., Magarey, A. M., and Hills, A. P. (2008). Minimizing random error in dietary intakes assessed by 24-h recall, in overweight and obese adults. *European Journal of Clinical Nutrition*, 62(4):537–543.
- Joint Formulary Committee (2017). *British National Formulary (online)*.
- Judkins, D. R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6(3):223–239.

- Kazzazi, F., Haggie, R., Forouhi, P., Kazzazi, N., and Malata, C. M. (2018). Utilizing the Total Design Method in medicine: maximizing response rates in long, non-incentivized, personal questionnaire postal surveys. *Patient Related Outcome Measures*, 9:169–172.
- Keogh, R. H. and White, I. R. (2014). A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Statistics in Medicine*, pages 2137–2155.
- Key, T. J. A., Thorogood, M., Appleby, P. N., and Burr, M. L. (1996). Dietary habits and mortality in 11 000 vegetarians and health conscious people: results of a 17 year follow up. *BMJ*, 313(7060):775–779.
- Killip, S., Bennett, J. M., and Chambers, M. D. (2007). Iron deficiency anemia. *American Family Physician*, 75(5):671–8.
- Kish, L. and Frankel, M. R. (1970). Balanced repeated replications for standard errors. *Journal of the American Statistical Association*, 65(331):1071–1094.
- Koenker, R. and Bassett, G. J. (1978). Regression Quantiles. *Econometrica*, 46(1):33–50.
- Koenker, R. and Hallock, K. F. (2001). Quantile Regression. *The Journal of Economic Perspectives*, 15(4):143–156.
- Lacerte, P., Pradipasen, M., Temcharoen, P., Imamee, N., and Vorapongsathorn, T. (2011). Determinants of adherence to iron/folate supplementation during pregnancy in two provinces in Cambodia. *Asia-Pacific Journal of Public Health / Asia-Pacific Academic Consortium for Public Health*, 23(3):315–323.
- Laureano, G. H. C., Torman, V. B. L., Crispim, S. P., Dekkers, A. L. M., and Camey, S. A. (2016). Comparison of the ISU, NCI, MSM, and SPADE Methods for Estimating Usual Intake: A Simulation Study of Nutrients Consumed Daily. *Nutrients*, 8(3):166.
- Leal, J., Gray, A. M., Prieto-Alhambra, D., Arden, N. K., Cooper, C., Javaid, M. K., Judge, A., and REFReSH study group (2016). Impact of hip fracture on hospital care costs: a population-based study. *Osteoporosis International*, 27(2):549–558.

- Lee, S. R. and Eltinge, J. L. (1999). Confidence Bounds for Survey-Weighted Quantile Plots and Offset-Function Plots. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002)*, 61(1):106–132.
- Lennox, A., Bluck, L., Page, P., Pell, D., Cole, D., Ziauddeen, N., Steer, T., Nicholson, S., Goldberg, G., and Prentice, A. (2014). Appendix X: Misreporting in the National Diet and Nutrition Survey Rolling Programme (NDNS RP): summary of results and their interpretation. In *National Diet and Nutrition Survey Results from Years 1, 2, 3 and 4 (combined) of the Rolling Programme (2008/2009 - 2011/2012)*. MRC-HNR.
- Li, Y., Graubard, B. I., and Korn, E. L. (2010). Application of nonparametric quantile regression to body mass index percentile curves from survey data. *Statistics in medicine*, 29(5):558–572.
- Littell, R. C., Stroup, W. W., Milliken, G. A., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for Mixed Models, Second Edition*. SAS Institute.
- Little, P., Gould, C., Williamson, I., Warner, G., Gantley, M., and Kinmonth, A. L. (1997). Reattendance and complications in a randomised trial of prescribing strategies for sore throat: the medicalising effect of prescribing antibiotics. *BMJ (Clinical research ed.)*, 315(7104):350–352.
- Liu, L., Ma, J. Z., and Johnson, B. A. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Statistics in Medicine*, 27(18):3528–3539.
- Liu, L., Strawderman, R. L., Cowen, M. E., and Shih, Y.-C. T. (2010). A flexible two-part random effects model for correlated medical costs. *Journal of Health Economics*, 29(1):110–123.
- Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., and Murray, C. J. L. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, 367(9524):1747–1757.
- Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19.

- Lumley, T. (2014). *survey: analysis of complex survey samples*.
- Ma, Y., Olendzki, B. C., Pagoto, S. L., Hurley, T. G., Magner, R. P., Ockene, I. S., Schneider, K. L., Merriam, P. A., and Hébert, J. R. (2009). Number of 24-hour diet recalls needed to estimate energy intake. *Annals of epidemiology*, 19(8):553–559.
- Macdiarmid, J. and Blundell, J. (1998). Assessing dietary intake: Who, what and why of under-reporting. *Nutrition Research Reviews*, 11(02):231–253.
- Mackerras, D. and Rutishauser, I. (2005). 24-hour national dietary survey data: how do we interpret them most effectively? *Public Health Nutrition*, 8(6):657–665.
- Manning, W. G. (1981). A two-part model of the demand for medical care : preliminary results from the Health Insurance Study. *Health, economics, and health economics : proceedings of the World Congress on Health Economics, Leiden, The Netherlands, September 1980*.
- Manning, W. G., Basu, A., and Mullahy, J. (2005). Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics*, 24(3):465–488.
- Mariappan, A., Bosch, M., Zhu, F., Boushey, C. J., Kerr, D. A., Ebert, D. S., and Delp, E. J. (2009). Personal dietary assessment using mobile devices. In *Proc IS&T/SPIE*, volume 7246, pages 72460Z–72460Z–12, San Jose, CA.
- McGeoghegan, L., Muirhead, C. R., and Almoosawi, S. (2015). Association between an anti-inflammatory and anti-oxidant dietary pattern and diabetes in British adults: results from the national diet and nutrition survey rolling programme years 1-4. *International Journal of Food Sciences and Nutrition*, 67(5):553–561.
- McIntosh, W. A., Kubena, K. S., Walker, J., Smith, D., and Landmann, W. A. (1990). The relationship between beliefs about nutrition and dietary practices of the elderly. *Journal of the American Dietetic Association*, 90(5):671–676.

- McLean, E., Cogswell, M., Egli, I., Wojdyla, D., and de Benoist, B. (2009). World-wide prevalence of anaemia, WHO Vitamin and Mineral Nutrition Information System, 1993–2005. *Public Health Nutrition*, 12(04):444–454.
- Meza, C., Osorio, F., and De la Cruz, R. (2012). Estimation in nonlinear mixed-effects models using heavy-tailed distributions. *Statistics and Computing*, 22(1):121–139.
- Mid Essex CCG (2015). *Your views on proposed changes to healthcare in your area*.
- Milliron, B.-J., Vitolins, M., and Tooze, J. (2014). Usual dietary intake among female breast cancer survivors is not significantly different from women with no cancer history: Results of the National Health And Nutrition Examination Survey, 2003-2006. *Journal of the Academy of Nutrition and Dietetics*, 114(6):932–937.
- Moshfegh, A. J., Rhodes, D. G., Baer, D. J., Murayi, T., Clemens, J. C., Rumpler, W. V., Paul, D. R., Sebastian, R. S., Kuczynski, K. J., Ingwersen, L. A., Staples, R. C., and Cleveland, L. E. (2008). The US Department of Agriculture Automated Multiple-Pass Method reduces bias in the collection of energy intakes. *The American Journal of Clinical Nutrition*, 88(2):324–332.
- Moulton, L. H. and Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, 51(4):1570–1578.
- Mullins, R. J., Clark, S., and Camargo Jr, C. A. (2009). Regional variation in epinephrine autoinjector prescriptions in Australia: more evidence for the vitamin D–anaphylaxis hypothesis. *Annals of Allergy, Asthma & Immunology*, 103(6):488–495.
- Murakami, K. and Livingstone, M. B. E. (2016). Associations between energy density of meals and snacks and overall diet quality and adiposity measures in British children and adolescents: the National Diet and Nutrition Survey. *British Journal of Nutrition*, 116(9):1633–1645.
- Neelon, B., O'Malley, A. J., and Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview:

Modeling zero-modified count and semicontinuous data in health services research  
Part 1: background and overview. *Statistics in Medicine*.

Nelson, M., Black, A. E., Morris, J. A., and Cole, T. J. (1989). Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. *The American journal of clinical nutrition*, 50(1):155–167.

Nelson, M., Erens, B., Bates, B., Church, S., and Boshier, T. (2007). Low income diet and nutrition survey.

NHS Choices (2011). The eatwell plate. Internet <http://www.nhs.uk/Livewell/Goodfood/Pages/eatwell-plate.asp> x (accessed 10th December 2012).

NHS Choices (2016). *5 A DAY: what counts? - Live Well - NHS Choices*.

NHS Digital (2015). Quality and Outcomes Framework.

Nicholson, S., Pot, G., Bates, C., Prentice, A., Cox, L., and Page, P. (2014). Blood Analytes. In *National Diet and Nutrition Survey Results from Years 1, 2, 3 and 4 (combined) of the Rolling Programme (2008/2009 – 2011/2012)*.

Nusser, S. M., Battese, G. E., and Fuller, W. A. (1990). Method of Moments Estimation of Usual Nutrient Intake Distributions. Center for Agricultural and Rural Development (CARD) Publications 90-wp52, Center for Agricultural and Rural Development (CARD) at Iowa State University.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996a). A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions. *Journal of the American Statistical Association*, 91(436):1440–1449.

Nusser, S. M., Fuller, W. A., and Guenther, P. (1996b). Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24-h food intake data. In *Survey Measurement and Process Quality*. Wiley, New York.



- Office for National Statistics, E. L. a. S. A. (2010). *SOC2010 volume 3: the National Statistics Socio-economic classification (NS-SEC rebased on SOC2010)*.
- Olsen, M. K. and Schafer, J. L. (2001). A Two-Part Random-Effects Model for Semi-continuous Longitudinal Data. *Journal of the American Statistical Association*, 96(454):730–745.
- Oppenheimer, S. J. (2001). Iron and Its Relation to Immunity and Infectious Disease. *The Journal of Nutrition*, 131(2):616S–635S.
- Osler, M., Milman, N., and Heitmann, B. L. (1999). Consequences of Removing Iron Fortification of Flour on Iron Status among Danish Adults: Some Longitudinal Observations between 1987 and 1994. *Preventive Medicine*, 29(1):32–36.
- Papanikolaou, Y., Brooks, J., Reider, C., and Fulgoni, V. L. (2014). U.S. adults are not meeting recommended levels for fish and omega-3 fatty acid intake: results of an analysis using observational data from NHANES 2003–2008. *Nutrition Journal*, 13(1).
- Penny, M. E., Creed-Kanashiro, H. M., Robert, R. C., Narro, M. R., Caulfield, L. E., and Black, R. E. (2005). Effectiveness of an educational intervention delivered through the health services to improve nutrition in young children: a cluster-randomised controlled trial. *Lancet (London, England)*, 365(9474):1863–1872.
- Preston, J. (2009). Rescaled bootstrap for stratified multistage sampling. *Survey Methodology*, 35(2):227–234.
- Price, G. M., Paul, A. A., Key, F. B., Harter, A. C., Cole, T. J., Day, K. C., and Wadsworth, M. E. J. (1995). Measurement of diet in a large national survey: comparison of computerized and manual coding of records in household measures. *Journal of Human Nutrition and Dietetics*, 8(6):417–428.
- Primary Care Support England (2018). Registrations - Primary Care Services England.

- Public Health England (2014). *National Diet and Nutrition Survey: results from Years 1 to 4 (combined) of the rolling programme for 2008 and 2009 to 2011 and 2012 - Publications - GOV.UK.*
- R Core Team (2016). R: A Language and Environment for Statistical Computing.
- Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(4):805–827.
- Rao, J. N. K. and Shao, J. (1996). On Balanced Half-Sample Variance Estimation in Stratified Random Sampling. *Journal of the American Statistical Association*, 91(433):343–348.
- Reddy, M. B., Hurrell, R. F., and Cook, J. D. (2006). Meat consumption in a varied diet marginally influences nonheme iron absorption in normal individuals. *Journal of Nutrition*, 136(3):576–581.
- Rennie, K. L., Coward, A., and Jebb, S. A. (2007). Estimating under-reporting of energy intake in dietary surveys using an individualised method. *British Journal of Nutrition*, 97(6):1169–1176.
- Rickard, A. P., Chatfield, M. D., Conway, R. E., Stephen, A. M., and Powell, J. J. (2009). An algorithm to assess intestinal iron availability for use in dietary surveys. *British Journal of Nutrition*, 102(11):1678–1685.
- Riley, H. (2010). NEWS AND VIEWS: NDNS rolling programme - what do the Year 1 results show? *Nutrition Bulletin*, 35(3):235–239.
- Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):637–654.
- Roberts, C., Steer, T., Maplethorpe, N., Cox, L., Meadows, S., Nicholson, S., Page, P., and Swan, G. (2018). National Diet and Nutrition Survey. Results from Years 7 and 8 (combined) of the Rolling Programme (2014/2015 to 2015/2016). Technical report.

- Rockett, H. R. H., Berkey, C. S., and Colditz, G. A. (2003). Evaluation of dietary assessment instruments in adolescents. *Current Opinion in Clinical Nutrition and Metabolic Care*, 6(5):557–562.
- Román-Montoya, Y., Rueda, M., and Arcos, A. (2008). Confidence intervals for quantile estimation using Jackknife techniques. *Computational Statistics*, 23(4):573–585.
- Rose, D., Pevalin, D. J., and O'Reilly, K. (2005). *The National Statistics Socio-economic Classification: origins, development and use*. Palgrave Macmillan Basingstoke.
- Rowlands, G. P., Mehay, A., Hampshire, S., Phillips, R., Williams, P., Mann, A., Steptoe, A., Walters, P., and Tylee, A. T. (2013). Characteristics of people with low health literacy on coronary heart disease GP registers in South London: a cross-sectional study. *BMJ Open*, 3(1):e001503.
- Rowlingson, B., Lawson, E., Taylor, B., and Diggle, P. J. (2013). Mapping English GP prescribing data: a tool for monitoring health-service inequalities. *BMJ Open*, 3(1):e001363.
- SACN (Scientific Advisory Committee on Nutrition) (2015). Carbohydrates and Health. Technical report.
- Sadighi, J., Sheikholeslam, R., Mohammad, K., Pouraram, H., Abdollahi, Z., Samadpour, K., Kolahdooz, F., and Naghavi, M. (2008). Flour fortification with iron: a mid-term evaluation. *Public Health*, 122(3):313–321.
- Sargen, M. R., Hoffstad, O. J., Wiebe, D. J., and Margolis, D. J. (2012). Geographic variation in pharmacotherapy decisions for U.S. Medicare enrollees with diabetes. *Journal of Diabetes and Its Complications*, 26(4):301–307.
- SAS Institute Inc (2011). *Base SAS® 9.3 Procedures Guide [computer program]*. SAS Institute Inc; Cary, NC.
- Schatzkin, A., Kipnis, V., Carroll, R. J., Midthune, D., Subar, A. F., Bingham, S., Schoeller, D. A., Troiano, R. P., and Freedman, L. S. (2003). A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological co-

- hort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study. *International Journal of Epidemiology*, 32(6):1054–1062.
- Scientific Advisory Committee on Nutrition (2011). *Iron and health*. Stationery Office, London. OCLC: 847858399.
- Scientific Advisory Committee on Nutrition (SACN), editor (1991). *Dietary reference values for food energy and nutrients for the United Kingdom: report*. Number 41 in Report on health and social subjects. TSO, London, 18. impression edition.
- Shahar, D. R., Yerushalmi, N., Lubin, F., Froom, P., Shahar, A., and Kristal-Boneh, E. (2001). Seasonal Variations in Dietary Intake Affect the Consistency of Dietary Assessment. *European Journal of Epidemiology*, 17(2):129–133.
- Shang, J., Sundara-Rajan, K., Lindsey, L., Mamishev, A., Johnson, E., Teredesai, A., and Kristal, A. (2011). A pervasive Dietary Data Recording System. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 307–309.
- Shohaimi, S., Welch, A., Bingham, S., Luben, R., Day, N., Wareham, N., and Khaw, K.-T. (2004). Residential area deprivation predicts fruit and vegetable consumption independently of individual educational level and occupational social class: a cross sectional population study in the Norfolk cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk). *Journal of Epidemiology & Community Health*, 58(8):686–691.
- Slob, W. (1993). Modeling long-term exposure of the whole population to chemicals in food. *Risk analysis: an official publication of the Society for Risk Analysis*, 13(5):525–530.
- Slob, W. (2006). Probabilistic dietary exposure assessment taking into account variability in both amount and frequency of consumption. *Food and Chemical Toxicology*, 44(7):933–951.
- Smithers, G. (1993). MAFF's Nutrient Databank. *Nutrition & Food Science*, pages 16–19.

- Smithers, G., Gregory, J. R., Bates, C. J., Prentice, A., Jackson, L. V., and Wenlock, R. (2000). The National Diet and Nutrition Survey: young people aged 4–18 years. *Nutrition Bulletin*, 25(2):105–111.
- Sousa, A. G. and Costa, T. H. M. d. (2015). Usual coffee intake in Brazil: results from the National Dietary Survey 2008–9. *British Journal of Nutrition*, 113(10):1615–1620.
- Souverein, O. W., Dekkers, A. L., Geelen, A., Haubrock, J., de Vries, J. H., Ocke, M. C., Harttig, U., Boeing, H., and van 't Veer, P. (2011). Comparing four methods to estimate usual intake distributions. *European Journal of Clinical Nutrition*, 65(S1):S92–S101.
- Stacy, E. W. (1962). A Generalization of the Gamma Distribution. *The Annals of Mathematical Statistics*, 33(3):1187–1192.
- Stallone, D. D., Brunner, E. J., Bingham, S. A., and Marmot, M. G. (1997). Dietary assessment in Whitehall II: the influence of reporting bias on apparent socioeconomic variation in nutrient intakes. *European Journal of Clinical Nutrition*, 51(12):815–825.
- Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7):1–46.
- Steer, T., Roberts, C., Nicholson, S., Pell, D., Page, P., Lennox, A., Prynne, C., and Swann, G. (2014). Dietary intakes. In *National Diet and Nutrition Survey Results from Years 1, 2, 3 and 4 (combined) of the Rolling Programme (2008/2009 - 2011/2012)*.
- Stephen, A. M., Sommerville, J. P., Henderson, H., Pell, D. A., and Allen, R. E. (2013). Food consumption in the Diet and Nutrition Survey of Infants and Young Children 2011 (DNSIYC). *Proceedings of the Nutrition Society*, 72(OCE3):E121.
- Stevens, L. and Nelson, M. (2011). The contribution of school meals and packed lunch to food consumption and nutrient intakes in UK primary school children from a low income population. *Journal of Human Nutrition and Dietetics*, 24(3):223–232.
- Su, L., Tom, B. D. M., and Farewell, V. T. (2009). Bias in 2-part mixed models for longitudinal semicontinuous data. *Biostatistics (Oxford, England)*, 10(2):374–389.

- Subar, A. F., Kipnis, V., Troiano, R. P., Midthune, D., Schoeller, D. A., Bingham, S. A., Sharbaugh, C. O., Trabulsi, J., Runswick, S., Ballard-Barbash, R., Sunshine, J., and Schatzkin, A. (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. *American journal of epidemiology*, 158(1):1–13.
- Subar, A. F., Mittl, B., Zimmerman, T. P., Kirkpatrick, S. I., Schap, T. E., Miller, A., Wilson, M. M., Kaefer, C., and Potischman, N. (2016). The Automated Self-Administered 24-Hour (ASA24) is Now Mobile and Can Collect Both 24-Hour Recalls and Food Records. *The FASEB Journal*, 30(1 Supplement):1153.6–1153.6.
- Subcommittee on Interpretation and Uses of Dietary Reference Intakes and the Standing Committee on the Scientific Evaluation of Dietary Reference Intakes (2003). *Dietary Reference Intakes: Applications in Dietary Planning*. The National Academies Press.
- Sui, Z., Zheng, M., Zhang, M., and Rangan, A. (2016). Water and beverage consumption: Analysis of the Australian 2011-2012 national nutrition and physical activity survey. *Nutrients*, 8(11).
- Syrad, H., Llewellyn, C. H., van Jaarsveld, C. H. M., Johnson, L., Jebb, S. A., and Wardle, J. (2016). Energy and nutrient intakes of young children in the UK: findings from the Gemini twin cohort. *The British Journal of Nutrition*, 115(10):1843–1850.
- Tang, W., Lu, N., Chen, T., Wang, W., Gunzler, D., Han, Y., and Tu, X. (2015). On Performance of Parametric and Distribution-free Models for Zero-inflated and Overdispersed Count Responses. *Statistics in medicine*, 34(24):3235–3245.
- Taylor, R. W. and Goulding, A. (1998). Validation of a short food frequency questionnaire to assess calcium intake in children aged 3 to 6 years. *European Journal of Clinical Nutrition*, 52(6):464–465.
- Tedstone, A., Anderson, S., and Allen, R. (2014). Sugar reduction responding to the challenge. *London: Public Health England*.

- Thane, C. W., Walmsley, C. M., Bates, C. J., Prentice, A., and Cole, T. J. (2000). Risk factors for poor iron status in British toddlers: further analysis of data from the National Diet and Nutrition Survey of children aged 1.5–4.5 years. *Public Health Nutrition*, 3(04):433–440.
- Tippett, K. S. and Cleveland, L. E. (2001). Results From USDA's 1994-96 Diet and Health Knowledge Survey. Technical Report 96-4, United States Department of Agriculture.
- Tipping, S. (2014). Appendix B: Sampling and weighting the NDNS. In *National Diet and Nutrition Survey Results from Years 1, 2, 3 and 4 (combined) of the Rolling Programme (2008/2009 - 2011/2012)*. MRC-HNR.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, pages 24–36.
- Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data with clumping at zero. *Statistical methods in medical research*, 11(4):341–355.
- Tooze, J. A., Kipnis, V., Buckman, D. W., Carroll, R. J., Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., and Dodd, K. W. (2010). A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Statistics in Medicine*, 29(27):2857–2868.
- Tooze, J. A., Midthune, D., Dodd, K. W., Freedman, L. S., Krebs-Smith, S. M., Subar, A. F., Guenther, P. M., Carroll, R. J., and Kipnis, V. (2006). A New Statistical Method for Estimating the Usual Intake of Episodically Consumed Foods with Application to Their Distribution. *Journal of the American Dietetic Association*, 106(10):1575–1587.
- Troy, K. L., Mancuso, M. E., Butler, T. A., and Johnson, J. E. (2018). Exercise Early and Often: Effects of Physical Activity and Exercise on Women's Bone Health. *International Journal of Environmental Research and Public Health*, 15(5).
- Tzavidis, N., Salvati, N., Schmid, T., Flouri, E., and Midouhas, E. (2016). Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Co-

- hort Study children in England using M-quantile random-effects regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):427–452.
- von Haehling, S., Jankowska, E. A., van Veldhuisen, D. J., Ponikowski, P., and Anker, S. D. (2015). Iron deficiency and cardiovascular disease. *Nature Reviews. Cardiology*, 12(11):659–669.
- Waijers, P. M. C. M., Dekkers, A. L., Boer, J. M., Boshuizen, H. C., and Rossum, C. T. M. v. (2006). The Potential of AGE MODE, an Age-Dependent Model, to Estimate Usual Intakes and Prevalences of Inadequate Intakes in a Population. *The Journal of Nutrition*, 136(11):2916–2920.
- Wallace, L. A., Duan, N., and Ziegenfus, R. (1994). Can long-term exposure distributions be predicted from short-term measurements? *Risk analysis: an official publication of the Society for Risk Analysis*, 14(1):75–85.
- Wawro, N., Kleiser, C., Himmerich, S., Gedrich, K., Boeing, H., Knueppel, S., and Linseisen, J. (2017). Estimating Usual Intake in the 2nd Bavarian Food Consumption Survey: Comparison of the Results Derived by the National Cancer Institute Method and a Basic Individual Means Approach. *Annals of Nutrition and Metabolism*, 71(3-4):164–174.
- Weiss, R., Stumbo, P. J., and Divakaran, A. (2010). Automatic Food Documentation and Volume Computation using Digital Imaging and Electronic Transmission. *Journal of the American Dietetic Association*, 110(1):42.
- WHO (2012). *Global Health Estimates (GHE) deaths by age, sex and cause*.
- WHO and CDC (2008). Worldwide prevalence of anaemia 1993-2005. *WHO global database on anaemia*.
- Willett, W. C., Sampson, L., Browne, M. L., Stampfer, M. J., Rosner, B., Hennekens, C. H., and Speizer, F. E. (1988). The Use of a Self-Administered Questionnaire to Assess Diet Four Years in the Past. *American Journal of Epidemiology*, 127(1):188–199.



- World Health Organisation (2001). *WHO | Iron deficiency anaemia: assessment, prevention and control*.
- Wu, Y. and Liu, Y. (2009). Stepwise multiple quantile regression estimation using non-crossing constraints. *Statistics and Its Interface*, 2:299–310.
- Yu, K. and Zhang, J. (2005). A Three-Parameter Asymmetric Laplace Distribution and Its Extension. *Communications in Statistics: Theory & Methods*, 34(9/10):1867–1879.
- Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L., and Carroll, R. J. (2011). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *The Annals of Applied Statistics*, 5(2B):1456–1487.
- Zhang, Z., Gillespie, C., Welsh, J. A., Hu, F. B., and Yang, Q. (2015). Usual Intake of Added Sugars and Lipid Profiles Among the U.S. Adolescents: National Health and Nutrition Examination Survey, 2005–2010. *Journal of Adolescent Health*, 56(3):352–359.
- Ziauddeen, N., Almiron-Roig, E., Penney, T. L., Nicholson, S., Kirk, S. F. L., and Page, P. (2017). Eating at Food Outlets and “On the Go” Is Associated with Less Healthy Food Choices in Adults: Cross-Sectional Data from the UK National Diet and Nutrition Survey Rolling Programme (2008–2014). *Nutrients*, 9(12):1315.
- Zijp, I. M., Korver, O., and Tijburg, L. B. (2000). Effect of tea and other dietary factors on iron absorption. *Critical Reviews in Food Science and Nutrition*, 40(5):371–398.

# Appendices

## **A Example of the advance letter sent to prospective NDNS participants**

The following is an example of the letter sent to participants selected to take part in the NDNS RP Y1-4 (2008-2012). This is to inform individuals about the survey, the data handling process and that an interviewer will be visiting from NatCen to invite them to take part and that they will be compensated for their time. Also included is contact details should further information be required.

## Figure 37

The letter sent in advance of an interviewer visit providing participant information on the NDNS RP years 1-4 (2008-2012).



Dear Sir/Madam,

### National Diet and Nutrition Survey.

We are writing to tell you about an important and unique study that collects information on the eating habits and health status of people in the United Kingdom. It involves gathering information about the food people eat, as well as their lifestyles and general health. All answers will be treated in strict confidence in accordance with the Data Protection Act, and the information will only be used for research purposes and food policy planning.

In the next few days, an interviewer from the *National Centre for Social Research (NatCen)* will call at your address and will be able to explain more about the study. The interviewer will then select, at random, up to two people from your household whom we would like to take part in the survey. Each interviewer carries an identity card which includes their photograph and the *NatCen* logo shown on the top of this letter.

We hope that your household will be willing to help us with this study. All parts of the study are optional and selected individuals can take part in some parts and not others. We rely on the goodwill of those invited to take part to make the study a success. As a token of our appreciation, everybody who provides information about their eating patterns will be given **£30 in High Street Vouchers**.

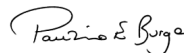
Some questions that you may have about the study are answered on the back of this letter. We also enclose a leaflet which tells you more about the study and why it is being done. If you have any other queries or want further information please contact Pauline Burge at *NatCen* on 0800 652 4572 or visit the National Diet and Nutrition Survey (NDNS) website: [www.natcen.ac.uk/NDNS](http://www.natcen.ac.uk/NDNS).

Many thanks in anticipation of your help.

Yours sincerely,



Gillian Swan  
Nutrition Science and Delivery Branch  
Department of Health



Pauline Burge  
Operations Department  
NatCen

The interviewer who will be calling at your address is: \_\_\_\_\_

Department of Health, Wellington House,  
133 - 155 Waterloo Road, London SE1 8UG  
E-mail: [gillian.swan@dh.gsi.gov.uk](mailto:gillian.swan@dh.gsi.gov.uk)

Operations Department, NatCen, Kings House,  
101-135 Kings Road, Brentwood, Essex CM14 4LX  
E-mail: [ndns@natcen.ac.uk](mailto:ndns@natcen.ac.uk)

## B Variables collected during the NDNS

Table 23 contains a list of variables collected as part of the NDNS RP Y1-4 (2008-2012). In total more than 2,000 variables are available covering socio-demographic characteristics along with food preparation skills, health indicators, supplement and medication intake along dietary intake data from the completed diet diary.

**Table 23**

Variables collected available from the NDNS Rolling Programme Years 1-4, (2008-2012)

Group	Area	Variables (n)
Classification	Household	23
	Individual	16
	Admin	11
	Booklet admin	3
	Education	4
	Employment	3
	Ethnicity	4
	Income	20
	Nurse admin	8
	Relationships	12
	Sample	18
	Weighting	26
Main food provider	Admin	10
	Cooking facilities	6
	Shopping habits	4
	Food preparation	29
	Cooking skills	41
	Ingredients	6
Cooking skills	Adult cooking skills	71
	Child cooking skills	17
School provision	School provision	54
Eating out and other provision	Eating out and other provision	17
Eating habits	Eating habits	24

Continued on next page

Table 23 – continued from previous page

Group	Area	Variables (n)
Food avoidance	Food avoidance	13
General health	General health	120
	Self-assessed health	73
	Longstanding illness	6
Prescribed medicines	General	1
	Drugs affecting blood analytes	17
	Reasons for taking medicines	48
	Sleep time	9
Oral health	Oral health	66
Smoking	Adult general health	16
	Adult current smokers	16
	Adult ex-smokers	7
	Children 8-15	4
Drinking	Adults general	2
	Adults 7 days	6
	Children 8-15	7
	Children 13-15	43
Actigraph	Admin	8
	Measurements	13
Anthropometric measurements	Demi-span admin	24
	Height/weight/infant length admin	174
	Mid upper arm circumference admin	19
	Waist/hip admin	48
	Measurements	9
Recent physical activity questionnaire	Home activities	17
	Activity at work/school/college	39
	Travel to work/school/college	11
	Leisure activities	6
	Adult physical activity profile	7
Sun exposure	Sun exposure at school	138

Continued on next page

Table 23 – continued from previous page

Group	Area	Variables (n)
	Sun exposure at work	18
	General	24
	Use of sun cream	12
	Holidays	105
Supplements	Supplements	51
Doubly labelled water (DLW)	Admin	6
	Measurements	10
Blood pressure	Admin	32
	Measurements	15
Urine sample	Admin	40
	Measurements	43
Blood sample	Admin	190
	Measurements	114
Food level dietary data	Admin	11
	Food groups	10
	Nutrients	60
	Disaggregated foods	30
	Other information	6
Day level dietary data - Foods	Admin	7
	Food groups (not including disaggregated foods)	66
	Other dietary information	13
Day level dietary data - Nutrients	Admin	7
	Nutrients (diet only)	40
	Nutrient (including supplements)	19
	Disaggregated foods	30
	Other dietary information	13
	Supplements	12
Person level dietary data	Admin	6
	Nutrients (diet only)	41
	Nutrients (including supplements)	19

Continued on next page

Table 23 – continued from previous page

Group	Area	Variables (n)
	Dietary reference values/nutrient intakes (percentage of total/food energy)	154
	Food groups (including disaggregated foods)	106
	Supplements	26



## **C Example pages from the NDNS RP food diaries**

Contained here are some example pages from the three estimated diet NDNS RP diaries that vary according to the age of the participant. The first is for those aged 19 and over, the second is participants aged 4-18y and the third is for infants aged under 4 years. As each diary is approximately 70 pages long, for the sake of brevity, a selection of example pages are included, though complete examples of the three diaries are published as Appendix E of the NDNS report (Bates et al., 2014a).

**Figure 38**

The title page of the food diary used for dietary assessment of adults in the NDNS RP years 1-4 (2008-2012).



NATIONAL DIET AND NUTRITION SURVEY  
*Food and Drink Diary*

DIARY START DATE: \_\_\_\_\_

SERIAL NUMBER

CKL

RESPONDENT No

FIRST NAME

Sex: Male / Female

Date of birth:

INTERVIEWER NUMBER:

INTERVIEWER NAME:

## Figure 39

Participant instructions to be read before completing the NDNS RP years 1-4 (2008-2012) diet diary - part 1.

### PLEASE READ THROUGH THESE PAGES BEFORE STARTING YOUR DIARY

We would like you to keep this diary of **everything you eat and drink** over 4 days. Please include all food consumed at home and outside the home e.g. work, college or restaurants. It is very important that you do not change what you normally eat and drink just because you are keeping this record. Please keep to your usual food habits.

#### Day and Date

Please write down the day and date at the top of the page each time you start a new day of recording.

#### Time Slots

Please note the time of each eating occasion into the space provided. For easy use each day is divided into sections, from the first thing in the morning to late evening and through the night.

#### Where and with whom?

For each eating occasion, please tell us what **room or part of the house** you were in when you ate, e.g. kitchen, living room. If you ate at your work canteen, a restaurant, fast food chain or your car, write that location down. We would also like to know **who you share your meals with**, e.g. whether you ate alone or with others. If you ate with others please describe their relationship to you e.g. partner, children, colleagues, or friends. We would also like to know **when you ate at a table** and **when you were watching television whilst eating**. For those occasions where you were **not** at a table or watching TV please write 'Not at table' or 'No TV' rather than leaving it blank.

#### What do you eat?

Please describe the food you eat in as much detail as possible. Be as specific as you can. Pages 16 - 21 will help with the sort of detail we need, like **cooking methods** (fried, grilled, baked etc) and any **additions** (fats, sugar/sweeteners, sauces, pepper etc).

##### □ Homemade dishes

If you have eaten any **homemade dishes** e.g. chicken casserole, please record the name of the recipe, ingredients with amounts (including water or other fluids) for the whole recipe, the number of people the recipe serves, and the cooking method. Write this down in the recipe section at the end of the record day. Record how much of the whole recipe you have eaten in the portion size column (see examples on pages 4 - 15).

##### □ Take-aways and eating out

If you have eaten **take-aways** or **made up dishes not prepared at home** such as at a restaurant or a friend's house, please record as much detail about the ingredients as you can e.g. vegetable curry containing chickpeas, aubergine, onion and tomato.

#### Brand name

Please note the **brand name** (if known). Most packed foods will list a brand name, e.g. Bird's eye, Hovis, or Supermarket own brands.

##### □ Labels/Wrappers

Labels are an important source of information for us. It helps us a great deal if you enclose, in the plastic bag provided, labels from all **ready meals**, labels from **foods of lesser known brands** and also from any **supplements** you take.

## Figure 40

Participant instructions to be read before completing the NDNS RP years 1-4 (2008-2012) diet diary - part 2.

---

### Portion sizes

Examples for how to describe the **quantity** or **portion size** you had of a particular food or drink are shown on pages 16 - 21.

For foods, quantity can be described using:

- **household measures**, e.g. one teaspoon (tsp) of sugar, two thick slices of bread, 4 tablespoons (tbsp) of peas, ½ cup of gravy. Be careful when describing amounts in spoons that you are referring to the correct spoon size. Compare the spoons you use with the life size pictures on page 28 of this diary.
- **weights from labels**, e.g. 4oz steak, 420g tin of baked beans, 125g pot of yoghurt
- **number of items**, e.g. 4 fish fingers, 2 pieces of chicken nuggets, 1 regular size jam filled doughnut
- **picture examples** for specific foods on pages 22-24.

For drinks, quantity can be described using:

- the **size of glass, cup etc** (e.g. large glass) or the **volume** (e.g. 300ml). Examples of typical drinks containers are on pages 26-27.
- **volumes from labels** (e.g. 330ml can of fizzy drink).

We would like to know the **amount that was actually eaten** which means taking **leftovers** into account. You can do this in two ways:

1. Record what was served and make notes of what was not eaten e.g. 3 tbsp of peas, only 2 tbsp eaten; 1 large sausage roll, ate only ½
2. Only record the amount actually eaten i.e. 2 tbsp of peas, ½ a large sausage roll

### Was it a typical day?

After each day of recording you will be prompted to tell us whether this was a typical day or whether there were any reasons why you ate and drank more or less than usual.

### Supplements

At the end of each recording day there is a section for providing information about any supplements you took. Brand name, full name of supplement, strength and the amount taken should be recorded.

### When to fill in the diary

**Please record your eating as you go, not from memory** at the end of the day. Use written notes on a pad if you forget to take your diary with you. Each diary day covers a 24hr period, so please include any food or drinks that you may have had during the night. Remember to include foods and drinks between meals (snacks) including water.

Overleaf you can see 2 example days that have been filled in by different people. These examples show you how we would like you to record your food and drink, for example a ready meal and a homemade dish. Your instruction booklet contains further examples such as how to describe food eaten in a restaurant.

<p>It only takes a few minutes for each eating occasion! For your convenience a separate booklet with instructions and examples is provided.</p>
--

**Thank you for your time – we really appreciate it!**

**Figure 41**

An example of a completed NDNS RP years 1-4 (2008-2012) diet diary - part 1.

Day: Thurs		Date: 31st March		
Time	Where? With Whom? TV on? At table?	Food/Drink description & preparation	Brand Name	Portion size or quantity eaten
<i>How to describe what you had and how much you had can be found on pages 16 - 21</i>				
<b>6am to 9am</b>				
6.30 am	Kitchen Alone No TV Not at table	Filter coffee, decaffeinated milk (fresh, semi-skimmed) Sugar white	Douwe Egberts  Silverspoon	Mug A little 1 level tsp
7.30 am	Kitchen Partner TV on At table	Filter coffee with milk and sugar Cornflakes Milk (fresh, semi-skimmed) Toast, granary medium sliced Light spread Marmalade	As above Tesco's own  Hovis Flora Hartleys	As above 1b drowned 1 slice med spread 1 heaped tsp
<b>9am to 12 noon</b>				
10.15 am	Office desk Alone No TV Not at table	Instant coffee, not decaffeinated Milk (fresh, whole) Sugar brown	Kenco	Mug A little 1 level tsp
11 am	Office desk Alone No TV Not at table	Digestive biscuit – chocolate coated on one side	McVities	2

**Figure 42**

An example of a completed NDNS RP years 1-4 (2008-2012) diet diary - part 2.

Time	Where? With Whom? TV on? At table?	Food/Drink description & preparation	Brand Name	Portion size or quantity eaten
<i>12 noon to 2pm</i>				
12.30 pm	Tea room at work Colleagues No TV At table	Ham salad sandwich from home Bread, wholemeal, thick sliced Light spread  Low fat Mayonnaise Smoked ham thinly sliced Lettuce, iceberg Cucumber with skin  Unsweetened orange juice from canteen  Apple with skin from home, Braeburn	Tesco's own Flora  Hellmans Tesco's own  Tropicana	2 slices thin spread on 1 slice  2 teaspoons 2 slices 1 leaf 4 thin slices  250ml carton  medium size, core left
<i>2pm to 5pm</i>				
3 pm	Meeting room at work With supervisor No TV Not at table	Tea, decaffeinated Milk (fresh, whole) Jaffa cake – mini variety	Twinnings Tesco's own McVities	Mug Some 6

**Figure 43**

An example of a completed NDNS RP years 1-4 (2008-2012) diet diary - part 3.

Time	Where? With Whom? TV on? At table?	Food/Drink description & preparation	Brand Name	Portion size or quantity <u>eaten</u>
<b>5pm to 8pm</b>				
6.30 pm	Pub Partner TV on At table	Gin Tonic water diet Lager 3.8% alcohol Salted peanuts	Gordon's Schweppes Draught, Carlsberg KP	Single measure 1/2 small glass 1 pint 1 handful
8 pm	Dining room Family No TV At table	Spaghetti, wholemeal Bolognese sauce (see recipe) Courgettes (fried in butter) Tinned peaches in juice (juice drained) Single cream UHT  Orange squash No Added Sugar	Tesco's own  Prince's  Sainsbury's own	3b 6 tablespoons 4 tablespoons 3 halves 1 tablespoon  200ml glass, 1 part squash, 3 parts tap water
<b>8pm to 10pm</b>				
9 pm	Sitting room Alone TV on Not at table	Grapes, green, seedless Chocolates, chocolate creams Potato crisps, Prawn Cocktail	Bendicks Walkers	15 2 25g bag (from multipack)
<b>10pm to 6am</b>				
10.30 pm	Bed room Partner No TV Not at table	Camomile tea (no milk or sugar)	Twinnings	1 mug

**Figure 44**

Questions covering whether the day's intake is typical from a NDNS RP years 1-4 (2008-2012) diet diary - part 1.

Was the amount of **food** that you had today about what you usually have, less than usual, or more than usual?

Yes, usual  No, less than usual

*Please tell us why you had less than usual*

No, more than usual

*Please tell us why you had more than usual*

Was the amount you had to **drink** today, including water, tea, coffee and soft drinks [and alcohol], about what you usually have, less than usual, or more than usual?

Yes, usual  No, less than usual

*Please tell us why you had less than usual*

No, more than usual

*Please tell us why you had more than usual*

Went to pub after work



**Figure 45**

Questions covering whether the day's intake is typical from a NDNS RP years 1-4 (2008-2012) diet diary - part 2.

Did you **finish all the food and drink** that you recorded in the diary today?

Yes

No

If no, please go back to the diary and make a note of any leftovers

Did you take any **vitamins, minerals or other food supplements** today?

Yes

No

If yes, please describe the supplements you took below

Brand	Name (in full) including strength	Number of pills, capsules, teaspoons
<i>Healthspan</i>	<i>Omega3 fish oil with vitamin A, C, D &amp; E</i>	<i>2 capsules</i>
<i>Boots</i>	<i>Calcium (1000mg) with vitamin D</i>	<i>1 tablet</i>
<i>Holland &amp; Barrett</i>	<i>Vitamin C 60mg</i>	<i>1 tablet</i>

Please record over the page details of any recipes or (if not already described) ingredients of made up dishes or take-away dishes.

**Figure 46**

An example of a completed recipe from a NDNS RP years 1-4 (2008-2012) diet diary.

<b>Write in recipes or ingredients of made up dishes or take-away dishes</b>			
<b>NAME OF DISH: <i>Bolognese sauce</i></b>		<b>SERVES: 4</b>	
<b>Ingredients</b>	<b>Amount</b>	<b>Ingredients</b>	<b>Amount</b>
<i>Co-op low fat beef mince</i>	<i>500g</i>	<i>Lea &amp; Perrins worcester sauce</i>	<i>dash</i>
<i>garlic</i>	<i>3 cloves</i>		
<i>onion</i>	<i>1 medium</i>		
<i>sweet red pepper</i>	<i>1 medium</i>		
<i>Napoli chopped tomatoes</i>	<i>400g tin</i>		
<i>Tesco tomato puree</i>	<i>1 tablespoon</i>		
<i>Tesco olive oil</i>	<i>1 tablespoon</i>		
<i>mixed herbs</i>	<i>1 dessertspoon</i>		
<b>Brief description of cooking method</b>			
<i>Fry onion &amp; garlic in oil, add mince and fry till brown.</i>			
<i>Add pepper, tomatoes, puree, Worcester sauce &amp; herbs. Simmer for 30 mins</i>			

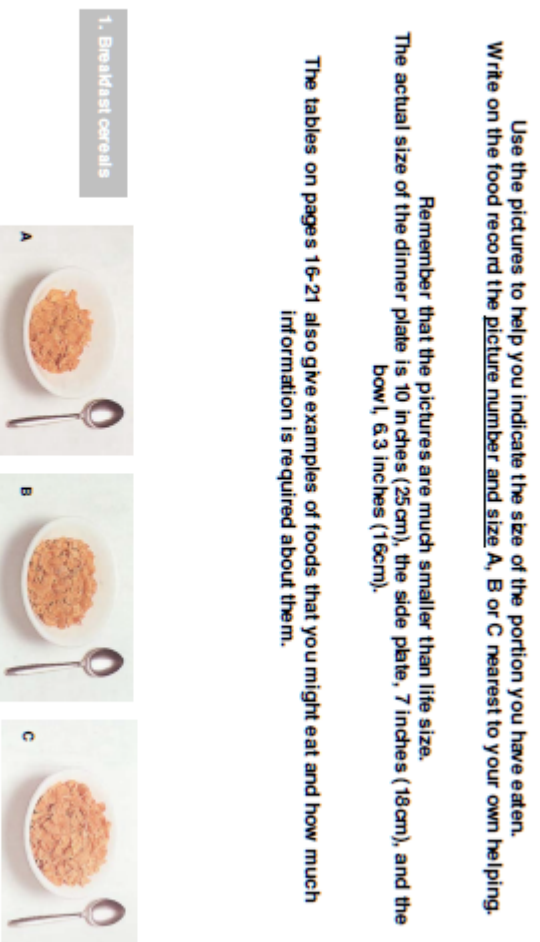
**Figure 47**

Information covering the detail requested for commonly consumed foods from a NDNS RP years 1-4 (2008-2012) diet diary.

Spoon size does matter!!!! When describing amounts check the spoons you use with the life size pictures on page 28 of this diary		
<i>Food/Drink</i>	<i>Description &amp; Preparation</i>	<i>Portion size or quantity</i>
Bacon	Back, middle, streaky; smoked or un-smoked; fat eaten; dry-fried or fried in oil/fat (type used) or grilled rashers	Number of rashers
Baked beans	Standard, reduced salt or reduced sugar	Spoons, weight of tin
Beefburger (hamburger)	Home-made (ingredients), from a packet or take-away; fried (type of oil/fat), microwaved or grilled; economy; with or without bread roll, with or without salad e.g. lettuce, tomato	Large or small, ounces or in grams if info on package
Beer	What sort e.g. stout, bitter, lager; draught, canned, bottled; % alcohol or low-alcohol or home-made	Number of pints or half pints, size of can or bottle
Biscuits	What sort e.g. cheese, wafer, crispbread, sweet, chocolate (fully or half coated), shortbread, home-made	Number, size (standard or mini variety)
Bread (see also sandwiches)	Wholemeal, granary, white or brown; currant, fruit, malt; large or small loaf; sliced or unsliced loaf	Number of slices; thick, medium or thin slices
Bread rolls	Wholemeal, white or brown; alone or with filling; crusty or soft	Size, number of rolls
Breakfast cereal (see also porridge)	What sort e.g. Kellogg's cornflakes; any added fruit and/or nuts; Muesli – with added fruit, no added sugar/salt variety	Spoons or picture 1
Buns and pastries	What sort e.g. iced, currant or plain, jam, custard, fruit, cream; type of pastry; homemade or bought	Size, number
Butter, margarine & fat spreads	Give full product name	Thick/average/thin spread; spoons
Cake	What sort: fruit (rich), sponge, fresh cream, iced, chocolate coated; type of filling e.g. buttercream, jam	Individual or size of slice, packet weight, picture 10

## Figure 48

An example of the food atlas from a NDNS RP years 1-4 (2008-2012) diet diary.



22

**Figure 49**

General dietary intake questions from a NDNS RP years 1-4 (2008-2012) diet diary

**General questions about your food/ drink during the recording period.**

**Special diet**

1. Did you follow a special diet during the recording period e.g. vegetarian, cholesterol lowering, weight reducing?

Yes

*Please specify*

No

**Milk**

2. Which type of milk did you use most often during the recording period?

Whole, fresh,   
pasteurised

Semi-skimmed fresh,   
pasteurised

Skimmed (fat free) fresh,   
pasteurised

1% fat milk,   
pasteurised

Dried

*Type*

Soya

*Type*

Other

*Type*

Did not use

Figure 50

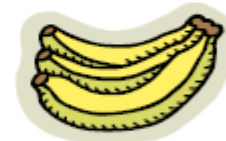
Front page from the NDNS RP years 1-4 (2008-2012) children's diet diary

217



NATIONAL DIET AND NUTRITION SURVEY  
***Food and Drink Diary***

DIARY START DATE: \_\_\_\_\_



SERIAL NUMBER

CKL

RESPONDENT No

FIRST NAME

Sex: Male / Female

Date of birth:

INTERVIEWER NUMBER:

INTERVIEWER NAME:

**Figure 51**

An example of a completed NDNS RP years 1-4 (2008-2012) children's diet diary - part 1.

Day		EXAMPLE	Day: Thursday	Date: March 31 <sup>st</sup>
Time	Where? with whom? TV on? Table?	What	Brand Name	Amount eaten
<b>How to describe what you had and how much you had can be found on pages 12-17</b>				
<b>6am to 9am</b>				
7.30am	Kitchen Family No TV At table	Orange juice, unsweetened, UHT Tea Milk, fresh semi skimmed Sugar white Weetabix Milk as above Sugar as above Toast wholemeal, large loaf Butter unsalted Strawberry Jam	Tesco Tesco Tesco Silverspoon  Hovis Anchor Co-op	Large glass Mug A little 2 level teaspoons 2 Drowned 2 heaped teaspoons 2 thin slices thick spread on both 1 teaspoon on one slice
<b>9am to 12 noon</b>				
11am	School playground With friends	Coca cola diet Potato crisps, Salt and Vinegar	Coca Cola Walkers	330ml can 25g packet from a multipack
12noon	School corridor Alone	Water from water cooler Mars Bar		small plastic cup 1 kingsize
<b>12 noon to 2pm</b>				
12.45pm	School canteen With friends At table	Sandwich, from home White bread, large loaf Spread Ham unsmoked Cheddar cheese Branston Pickle Apple with skin from home Ribena Light, Ready to Drink, Blackcurrant, from canteen Kitkat from home	Kingsmill Flora Light Tesco	2 med slices thin spread on both slices 1 slice 2 medium slices 1 teaspoon 1 (left core) 220ml carton 2 fingers
1.50pm	School corridor Alone	Chewing gum	Orbit Sugar Free	1 piece



NATIONAL DIET AND NUTRITION SURVEY  
*Food and Drink Diary*  
*Children aged 1.5 to 3 years*

DIARY START DATE: \_\_\_\_\_

SERIAL NUMBER (7 digits)    CIL    RESPONDENT No

<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------	----------------------

First name:

Sex: Male / Female      Date of birth: 

<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
----------------------	----------------------	----------------------	----------------------	----------------------	----------------------

INTERVIEWER NUMBER: 

<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
----------------------	----------------------	----------------------	----------------------	----------------------	----------------------

      INTERVIEWER NAME:

NDNS(1) Diary (Parents of children 1.5-3 years), April 09 RECQ Ref: 07/03604/113 [For use from 01/04/17](#)

Figure 52

Front page from the NDNS RP years 1-4 (2008-2012) infant's diet diary



**Figure 53**

An example of a completed NDNS RP years 1-4 (2008-2012) infants's diet diary - part 1.

Day 1: Thurs		Date: 31 March 2007		
Time	Where? With whom? TV on? Table?	Food/Drink description & preparation	Brand Name	Portion size or quantity <u>eaten</u>
<i>How to describe what you had and how much you had can be found on pages 16-21</i>				
<i>6am to 9am</i>				
8am	Living Room Family TV on Not at table	Follow on Milk	SMA Progress	240ml bottle (as usual)
<i>9am to 12 noon</i>				
10am	Kitchen Mother No TV At table	Weetabix Full fat milk  white sugar	Weetabix Sainsbury's  Tate and Lyle	1 biscuit  drowned (about 1 dsp milk leftover) 2 teasp
11.30 am	Living Room Family TV on Not at table	bread  margarine  pure apple juice	Granary from bakers, medium cut  Flora light spread,  Sainsburys	1 slice  medium spread  200ml carton (drank ½ of it)

## **D Two-part models of complex survey data using a generalised gamma distribution: supplementary tables**

Appendix D contains tables relating to work carried out to determine an appropriate number of bootstrap samples to conduct to ensure acceptable precision whilst minimising the duration of time taken for analysis, discussed in Section 3.5. Presented in Section 3.5 are the estimated standard errors for iron from vegetable intake from an average of 50, 100, 200 and 300 bootstrap samples, the following tables provide the standard error estimates for the remaining food groups covered in Section 3 i.e. for iron from breakfast cereals (**Table 24a** and **b**), iron from bread **Table 24c** and **d**), iron from fruit **Table 24e** and **f**) and iron from the fruit and vegetable combined food group **Table 24g** and **h**). Also presented are the percentage differences between standard errors estimated from the average 50 bootstrap samples compared to the average from 100, 200 and 300 bootstrap samples respectively for iron from breakfast cereals (**Table 25a** and **b**), iron from bread (**Table 25c** and **d**), iron from fruit (**Table 25e** and **f**) and iron from the fruit and vegetable combined food group (**Table 25g** and **h**)

**Table 24a**

Estimated parameters of the two-part model for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	0.41	0.10185	0.10168	0.10182	0.10184
Age Group	1.5 -3y (Reference)					
	4-10y	0.91	0.33331	0.33194	0.33270	0.33249
	11-18y	-0.10	0.32629	0.32494	0.32566	0.32545
	19-64y	-0.88	0.28951	0.28810	0.28874	0.28857
	65y and older	0.80	0.31047	0.30898	0.30978	0.30963
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	0.02	0.16171	0.16135	0.16182	0.16172
	Intermediate occupations	0.10	0.21893	0.21818	0.21890	0.21893
	Small employers & own account workers	-0.003	0.20080	0.19974	0.20039	0.20051
	Lower supervisory & technical occupations	0.005	0.20601	0.20490	0.20521	0.20534
	Semi-routine occupations	-0.10	0.19097	0.19082	0.19113	0.19099
	Routine occupations	-0.40	0.20183	0.20148	0.20194	0.20192
	Never worked	-0.001	0.36100	0.35977	0.36001	0.36041
	Other	0.50	0.38884	0.39148	0.39104	0.39173

**Table 24b**

Estimated parameters of the two-part model for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	-0.11	0.03123	0.03122	0.03124	0.03122
Age Group	1.5 -3y (Reference)					
	4-10y	0.48	0.08895	0.08866	0.08900	0.08878
	11-18y	0.84	0.09054	0.09023	0.09060	0.09030
	19-64y	0.37	0.07760	0.07729	0.07758	0.07738
	65y and older	-0.002	0.08302	0.08270	0.08305	0.08284
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	0.03	0.04784	0.04780	0.04796	0.04786
	Intermediate occupations	0.01	0.06463	0.06456	0.06476	0.06466
	Small employers & own account workers	-0.01	0.06101	0.06081	0.06096	0.06095
	Lower supervisory & technical occupations	-0.01	0.06361	0.06316	0.06313	0.06242
	Semi-routine occupations	0.01	0.05763	0.05766	0.05775	0.05760
	Routine occupations	0.01	0.06217	0.06225	0.06245	0.06229
	Never worked	-0.002	0.11086	0.11026	0.11067	0.11066
	Other	-0.002	0.11644	0.11801	0.11766	0.11771
$\hat{k}$ , GG distribution shape parameter		2.93	0.07427	0.07415	0.07345	0.07440
$\hat{\sigma}$ , GG distribution scale parameter		0.01	0.05543	0.05531	0.05489	0.05544
Variance components	$\hat{\sigma}_u$	6.01	0.20948	0.20848	0.20850	0.20899
	$\hat{\sigma}_v$	0.32	0.01840	0.01860	0.01866	0.01802
	$c\hat{o}v(u, v)$	0.24	0.06453	0.06458	0.06475	0.06503

**Table 24c**

Estimated parameters of the two-part model for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	-0.38	0.08346	0.08097	0.08182	0.08142
Age Group	1.5 -3y (Reference)					
	4-10y	0.38	0.26927	0.26347	0.26581	0.26465
	11-18y	-0.04	0.25845	0.25312	0.25528	0.25414
	19-64y	0.01	0.23197	0.22728	0.22918	0.22823
	65y and older	1.01	0.25167	0.24669	0.24861	0.24764
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.56	0.13378	0.13120	0.13199	0.13153
	Intermediate occupations	-0.61	0.17834	0.17531	0.17621	0.17576
	Small employers & own account workers	-0.26	0.16508	0.16267	0.16362	0.16285
	Lower supervisory & technical occupations	-0.47	0.16879	0.16564	0.16660	0.16594
	Semi-routine occupations	-0.48	0.15519	0.15226	0.15332	0.15274
	Routine occupations	-0.54	0.16419	0.16083	0.16197	0.16136
	Never worked	-0.99	0.28528	0.27647	0.28102	0.27915
	Other	-0.45	0.31010	0.30341	0.30437	0.30416

**Table 24d**

Estimated parameters of the two-part model for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	-0.24	0.01745	0.01735	0.01734	0.01733
Age Group	1.5 -3y (Reference)					
	4-10y	0.32	0.05790	0.05772	0.05770	0.05763
	11-18y	0.49	0.05640	0.05626	0.05625	0.05617
	19-64y	0.58	0.05044	0.05028	0.05026	0.05020
	65y and older	0.44	0.05347	0.05329	0.05237	0.05321
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.07	0.02804	0.02794	0.02792	0.02790
	Intermediate occupations	-0.13	0.03795	0.03746	0.03735	0.03743
	Small employers & own account workers	-0.07	0.03408	0.03410	0.03404	0.03397
	Lower supervisory & technical occupations	-0.04	0.03527	0.03502	0.03503	0.03497
	Semi-routine occupations	-0.09	0.03283	0.03271	0.03269	0.03266
	Routine occupations	-0.05	0.03451	0.03438	0.03441	0.03437
	Never worked	-0.04	0.06383	0.06322	0.06366	0.06349
	Other	0.04	0.06635	0.06589	0.06575	0.06592
$\hat{k}$ , GG distribution shape parameter		0.50	0.09989	0.09935	0.09907	0.09921
$\hat{\sigma}$ , GG distribution scale parameter		1.22	0.04712	0.04702	0.04689	0.04693
Variance components	$\hat{\sigma}_u$	1.93	0.15788	0.14410	0.14827	0.14648
	$\hat{\sigma}_v$	0.11	0.00629	0.00618	0.00613	0.00612
	$c\hat{o}v(u, v)$	0.20	0.02176	0.02092	0.02118	0.02105

**Table 24e**

Estimated parameters of the two-part model for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	0.13	0.08105	0.08073	0.08089	0.08080
Age Group	1.5 -3y (Reference)					
	4-10y	0.07	0.27256	0.27143	0.27175	0.27147
	11-18y	-0.64	0.26554	0.26452	0.26492	0.26451
	19-64y	-0.23	0.23837	0.23748	0.23786	0.23752
	65y and older	0.86	0.25500	0.25500	0.25420	0.25418
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.52	0.12988	0.12957	0.12966	0.12955
	Intermediate occupations	-0.93	0.17642	0.17506	0.17533	0.17512
	Small employers & own account workers	-0.36	0.15921	0.15842	0.15844	0.15870
	Lower supervisory & technical occupations	-0.80	0.16335	0.16281	0.16286	0.16291
	Semi-routine occupations	-1.16	0.15160	0.15122	0.15137	0.15122
	Routine occupations	-1.18	0.16003	0.15923	0.15942	0.15926
	Never worked	-1.01	0.29357	0.29173	0.29192	0.29231
	Other	-0.39	0.31379	0.31241	0.31393	0.3135

**Table 24f**

Estimated parameters of the two-part model for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	-0.02	0.03159	0.03125	0.03142	0.03138
Age Group	1.5 -3y (Reference)					
	4-10y	0.15	0.09598	0.09483	0.09256	0.09520
	11-18y	0.20	0.09739	0.09626	0.09673	0.09662
	19-64y	0.34	0.08344	0.08246	0.08286	0.08279
	65y and older	0.56	0.08896	0.08796	0.08837	0.08828
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.15	0.04803	0.04755	0.04772	0.04767
	Intermediate occupations	-0.16	0.06800	0.06691	0.06727	0.06712
	Small employers & own account workers	-0.13	0.05980	0.05903	0.05942	0.05934
	Lower supervisory & technical occupations	-0.22	0.06212	0.06145	0.06177	0.06177
	Semi-routine occupations	-0.23	0.05855	0.05808	0.05827	0.05820
	Routine occupations	-0.29	0.06343	0.06255	0.06297	0.06282
	Never worked	-0.13	0.11668	0.11507	0.11539	0.11530
	Other	-0.13	0.11485	0.11404	0.11555	0.11494
$\hat{k}$ , GG distribution shape parameter		0.60	0.07278	0.07130	0.07175	0.07158
$\hat{\sigma}$ , GG distribution scale parameter		0.22	0.05048	0.04957	0.04932	0.04968
Variance components	$\hat{\sigma}_u$	2.22	0.12774	0.12703	0.12735	0.12714
	$\hat{\sigma}_v$	0.38	0.01894	0.01827	0.01858	0.01849
	$c\hat{o}v(u, v)$	0.55	0.04161	0.04067	0.04105	0.04098



**Table 24g**

Estimated parameters of the two-part model for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	0.58	0.08985	0.08905	0.08877	0.08896
Age Group	1.5 -3y (Reference)					
	4-10y	0.15	0.29815	0.29603	0.29628	0.29559
	11-18y	-0.46	0.28346	0.28211	0.28227	0.28175
	19-64y	0.49	0.25891	0.25738	0.25772	0.25716
	65y and older	1.15	0.28348	0.28133	0.28188	0.28124
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.23	0.15037	0.14846	0.14886	0.14845
	Intermediate occupations	-0.85	0.19320	0.19150	0.19175	0.19139
	Small employers & own account workers	-0.17	0.18407	0.18150	0.18198	0.18140
	Lower supervisory & technical occupations	-0.67	0.18307	0.18122	0.18150	0.18111
	Semi-routine occupations	-1.02	0.16773	0.16619	0.16615	0.16588
	Routine occupations	-1.01	0.17421	0.17277	0.17280	0.17249
	Never worked	-0.99	0.30608	0.30650	0.30632	0.30631
Other	-0.38	0.34662	0.34588	0.34692	0.34662	

**Table 24h**

Estimated parameters of the two-part model for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012) showing the impact upon standard error estimation from the average of 50, 100, 200 and 300 bootstrap samples: Part 2

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
Sex	Males (Reference)					
	Females	-0.02	0.02945	0.02951	0.02954	0.02946
Age Group	1.5 -3y (Reference)					
	4-10y	0.30	0.09464	0.09440	0.09468	0.09439
	11-18y	0.36	0.09188	0.09369	0.09358	0.09354
	19-64y	0.79	0.08204	0.08190	0.08215	0.08190
	65y and older	0.76	0.08731	0.08717	0.08745	0.08716
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.08	0.04590	0.04588	0.04598	0.04585
	Intermediate occupations	-0.10	0.06301	0.06322	0.06339	0.06319
	Small employers & own account workers	-0.03	0.05681	0.05675	0.05690	0.05674
	Lower supervisory & technical occupations	-0.09	0.05892	0.05891	0.05899	0.05882
	Semi-routine occupations	-0.15	0.05519	0.05522	0.05520	0.05506
	Routine occupations	-0.19	0.05820	0.05836	0.05833	0.05822
	Never worked	-0.11	0.10574	0.10734	0.10707	0.10699
	Other	-0.11	0.11096	0.11161	0.11175	0.11159
$\hat{k}$ , GG distribution shape parameter		0.91	0.09581	0.09475	0.09487	0.09440
$\hat{\sigma}$ , GG distribution scale parameter		0.25	0.04761	0.04820	0.04837	0.04830
Variance components	$\hat{\sigma}_u$	1.66	0.14924	0.14733	0.14699	0.14685
	$\hat{\sigma}_v$	0.23	0.02464	0.02362	0.02449	0.02415
	$c\hat{o}v(u, v)$	0.35	0.04227	0.04184	0.04204	0.04184

**Table 25a**

Percentage difference between standard error estimates for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1

		Standard Error (Number of bootstrap samples)				
		Estimates	SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	0.41	0.10185	0.17	0.03	0.01
Age Group	1.5 -3y (Reference)					
	4-10y	0.91	0.33331	0.41	0.18	0.25
	11-18y	-0.10	0.32629	0.41	0.19	0.26
	19-64y	-0.88	0.28951	0.49	0.27	0.32
	65y and older	0.80	0.31047	0.48	0.22	0.27
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	0.02	0.16171	0.22	-0.07	-0.01
	Intermediate occupations	0.10	0.21893	0.34	0.01	0
	Small employers & own account workers	-0.003	0.20080	0.53	0.20	0.14
	Lower supervisory & technical occupations	0.005	0.20601	0.54	0.39	0.33
	Semi-routine occupations	-0.10	0.19097	0.08	-0.08	-0.01
	Routine occupations	-0.40	0.20183	0.17	-0.05	-0.04
	Never worked	-0.001	0.36100	0.34	0.27	0.16
Other	0.50	0.38884	-0.68	-0.57	-0.74	

**Table 25b**

Percentage difference between standard error estimates for iron intake from breakfast cereals in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2

		Standard Error (Number of bootstrap samples)				
		Estimates	SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	-0.11	0.03123	0.03	-0.03	0.03
Age Group	1.5 -3y (Reference)					
	4-10y	0.48	0.08895	0.33	-0.06	0.19
	11-18y	0.84	0.09054	0.34	-0.07	0.27
	19-64y	0.37	0.07760	0.40	0.03	0.28
	65y and older	-0.002	0.08302	0.39	-0.04	0.22
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	0.03	0.04784	0.08	-0.25	-0.04
	Intermediate occupations	0.01	0.06463	0.11	-0.20	-0.05
	Small employers & own account workers	-0.01	0.06101	0.33	0.08	0.10
	Lower supervisory & technical occupations	-0.01	0.06361	0.71	0.75	1.87
	Semi-routine occupations	0.01	0.05763	-0.05	-0.21	0.05
	Routine occupations	0.01	0.06217	-0.13	-0.45	-0.19
	Never worked	-0.002	0.11086	0.54	0.17	0.18
	Other	-0.002	0.11644	-1.35	-1.05	-1.09
$\hat{k}$ , GG distribution shape parameter		2.93	0.07427	0.48	0.47	0.23
$\hat{\sigma}$ , GG distribution scale parameter		0.01	0.05543	-1.09	-1.41	2.07
Variance components	$\hat{\sigma}_u$	6.01	0.20948	0.16	1.1	-0.18
	$\hat{\sigma}_v$	0.32	0.01840	0.22	0.97	-0.02
	$c\hat{\sigma}v(u,v)$	0.24	0.06453	-0.08	-0.34	-0.77

**Table 25c**

Percentage difference between standard error estimates for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1

		Standard Error (Number of bootstrap samples)				
		Estimates	SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	-0.38	0.08151	-0.06	-0.52	0.02
Age Group	1.5 -3y (Reference)					
	4-10y	0.38	0.26540	0.15	-0.21	0.21
	11-18y	-0.04	0.25475	0.07	-0.26	0.16
	19-64y	0.01	0.22879	0.1	-0.27	0.17
	65y and older	1.01	0.24822	0.09	-0.24	0.16
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.56	0.13152	-0.21	-0.46	-0.07
	Intermediate occupations	-0.61	0.17622	0.23	-0.15	0.21
	Small employers & own account workers	-0.26	0.16310	-0.06	-0.26	0.08
	Lower supervisory & technical occupations	-0.47	0.16614	0.11	-0.34	0.05
	Semi-routine occupations	-0.48	0.15300	-0.01	-0.27	0.10
	Routine occupations	-0.54	0.16165	-0.09	-0.20	0.12
	Never worked	-0.99	0.28115	0.91	0.30	0.65
Other	-0.45	0.30509	0.49	-0.16	0.32	

**Table 25d**

Percentage difference between standard error estimates for iron intake from bread in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2

		Standard Error (Number of bootstrap samples)				
		Estimates	SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	-0.24	0.01743	0.80	0.63	0.57
Age Group	1.5 -3y (Reference)					
	4-10y	0.32	0.05795	0.78	0.47	0.55
	11-18y	0.49	0.05646	0.76	0.41	0.51
	19-64y	0.58	0.05047	0.79	0.40	0.53
	65y and older	0.44	0.05351	0.78	0.47	0.54
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.07	0.02805	0.75	0.57	0.57
	Intermediate occupations	-0.13	0.03746	0.51	0.13	0.13
	Small employers & own account workers	-0.07	0.03412	0.53	0.53	0.47
	Lower supervisory & technical occupations	-0.04	0.03510	0.83	0.37	0.43
	Semi-routine occupations	-0.09	0.03293	1.00	0.88	0.85
	Routine occupations	-0.05	0.03458	0.84	0.69	0.61
	Never worked	-0.04	0.06361	0.72	0.06	0.14
	Other	0.04	0.06612	1.04	0.24	0.41
$\hat{k}$ , GG distribution shape parameter		0.50	0.0992	0.42	0.02	0.03
$\hat{\sigma}$ , GG distribution scale parameter		1.22	0.04706	0.57	0.30	0.30
Variance components	$\hat{\sigma}_u$	1.93	0.14740	0.47	-1.39	0.38
	$\hat{\sigma}_v$	0.11	0.0062	1.61	2.42	1.45
	$c\hat{\sigma}v(u,v)$	0.20	0.02134	1.87	0.80	1.31

**Table 25e**

Percentage difference between standard error estimates for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	0.13	0.08105	0.39	0.20	0.31
Age Group	1.5 -3y (Reference)					
	4-10y	0.07	0.27256	0.41	0.30	0.40
	11-18y	-0.64	0.26554	0.24	0.17	0.25
	19-64y	-0.23	0.23837	0.37	0.21	0.36
	65y and older	0.86	0.25500	0.31	0.16	0.32
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.52	0.12988	0.24	0.17	0.25
	Intermediate occupations	-0.93	0.17642	0.77	0.62	0.74
	Small employers & own account workers	-0.36	0.15921	0.50	0.23	0.32
	Lower supervisory & technical occupations	-0.80	0.16335	0.33	0.30	0.27
	Semi-routine occupations	-1.16	0.15160	0.25	0.15	0.25
	Routine occupations	-1.18	0.16003	0.50	0.38	0.48
	Never worked	-1.01	0.29357	0.63	0.56	0.43
	Other	-0.39	0.31379	0.44	-0.04	0.09

**Table 25f**

Percentage difference between standard error estimates for iron intake from fruit in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 2

		Standard Error (Number of bootstrap samples)				
		Estimates	SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	-0.02	0.03159	1.08	0.54	0.66
Age Group	1.5 -3y (Reference)					
	4-10y	0.15	0.09598	1.20	0.75	0.81
	11-18y	0.20	0.09739	1.16	0.68	0.79
	19-64y	0.34	0.08344	1.17	0.70	0.78
	65y and older	0.56	0.08896	1.12	0.66	0.76
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.15	0.04803	1.00	0.65	0.75
	Intermediate occupations	-0.16	0.06800	1.60	1.07	1.29
	Small employers & own account workers	-0.13	0.05980	1.29	0.64	0.77
	Lower supervisory & technical occupations	-0.22	0.06212	1.08	0.56	0.56
	Semi-routine occupations	-0.23	0.05855	0.80	0.48	0.60
	Routine occupations	-0.29	0.06343	1.39	0.73	0.96
	Never worked	-0.13	0.11668	1.38	1.11	1.18
	Other	-0.13	0.11485	0.71	-0.61	-0.08
$\hat{k}$ , GG distribution shape parameter		0.60	0.07278	2.03	1.42	1.65
$\hat{\sigma}$ , GG distribution scale parameter		0.22	0.05048	1.8	2.3	1.58
Variance components	$\hat{\sigma}_u$	2.22	0.12774	0.56	0.31	0.47
	$\hat{\sigma}_v$	0.38	0.01894	3.54	1.9	2.38
	$c\hat{\sigma}v(u,v)$	0.55	0.04161	2.26	1.35	1.51



**Table 25g**

Percentage difference between standard error estimates for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 replicas: Part 1

		Estimates	Standard Error (Number of bootstrap samples)			
			SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	0.58	0.08985	0.89	0.78	0.99
Age Group	1.5 -3y (Reference)					
	4-10y	0.15	0.29815	0.71	0.63	0.86
	11-18y	-0.46	0.28346	0.48	0.42	0.60
	19-64y	0.49	0.25891	0.59	0.46	0.68
	65y and older	1.15	0.28348	0.76	0.56	0.79
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.23	0.15037	1.27	1.00	1.28
	Intermediate occupations	-0.85	0.19320	0.88	0.75	0.94
	Small employers & own account workers	-0.17	0.18407	1.40	1.14	1.45
	Lower supervisory & technical occupations	-0.67	0.18307	1.01	0.86	1.07
	Semi-routine occupations	-1.02	0.16773	0.92	0.94	1.10
	Routine occupations	-1.01	0.17421	0.83	0.81	0.99
	Never worked	-0.99	0.30608	-0.14	-0.08	-0.02
Other	-0.38	0.34662	0.21	-0.09	0	

**Table 25h**

Percentage difference between standard error estimates for iron intake from fruit and vegetables in the UK using data from NDNS RP Years 1-4 (2008-2012), with 100, 200 and 300 bootstrap samples and 50 bootstrap samples: Part 2

		Standard Error (Number of bootstrap samples)				
		Estimates	SE (50)	SE (100)	SE (200)	SE (300)
		Percent difference from 50 bootstrap samples				
Sex	Males (Reference)					
	Females	-0.02	0.02945	-0.2	-0.31	-0.03
Age Group	1.5 -3y (Reference)					
	4-10y	0.30	0.09464	0.25	-0.04	0.26
	11-18y	0.36	0.09351	0	-0.26	-0.01
	19-64y	0.79	0.08204	0.17	-0.13	0.17
	65y and older	0.76	0.08731	0.04	-0.17	0.11
NS-SEC	Higher managerial & professional occupations (Reference)					
	Lower managerial & professional occupations	-0.08	0.04590	0.04	-0.17	0.11
	Intermediate occupations	-0.10	0.06301	-0.33	-0.6	-0.29
	Small employers & own account workers	-0.03	0.05681	0.11	-0.16	0.12
	Lower supervisory & technical occupations	-0.09	0.05892	0.02	-0.12	0.17
	Semi-routine occupations	-0.15	0.05519	-0.05	-0.02	0.24
	Routine occupations	-0.19	0.05820	-0.27	-0.22	-0.
	Never worked	-0.11	0.10574	-1.51	-1.26	-1.18
	Other	-0.11	0.11096	-0.59	-0.71	-0.57
$\hat{k}$ , GG distribution shape parameter		0.91	0.09581	1.11	0.98	1.47
$\hat{\sigma}$ , GG distribution scale parameter		0.25	0.04761	-1.24	-1.6	-1.45
Variance components	$\hat{\sigma}_u$	1.66	0.14924	1.28	1.51	1.6
	$\hat{\sigma}_v$	0.23	0.02464	4.14	0.61	1.99
	$c\hat{\sigma}v(u, v)$	0.35	0.04227	1.02	0.54	1.02

## E Two-part models of complex survey data using a generalised gamma distribution: quadrature point comparison

Specifying the number of quadrature points used in maximum likelihood estimation of the two part model can have an impact upon time taken for the model to converge with little advantage in increasing from 5 to 10 points (Liu et al., 2010). To test this claim here four scenarios were explored estimating iron intake from vegetables using 5, 10, 15 and 20 quadrature points using 50 bootstrap replications to estimate standard errors. **Figure 54** shows the variation in each of the 31 estimated standard errors for each of the 4 scenarios and highlights that there was little difference in the standard errors of the four models.

The time taken to run of the 50 fitted models is presented in **Table 26**, which shows using 5 quadrature points takes an average of 18 minutes and 16 seconds to run each model and 15 hours 13 minutes and 7 seconds in total for 50 bootstrap replications. This increased dramatically when using 10 quadrature points taking an average of 1 hour and 41 minutes and a little over 89 hours in total, to run 50 bootstrapped models with 15 quadrature points took more than 313 hours and using 20 quadrature points took more than 577 hours or 24 days to run 50 models.

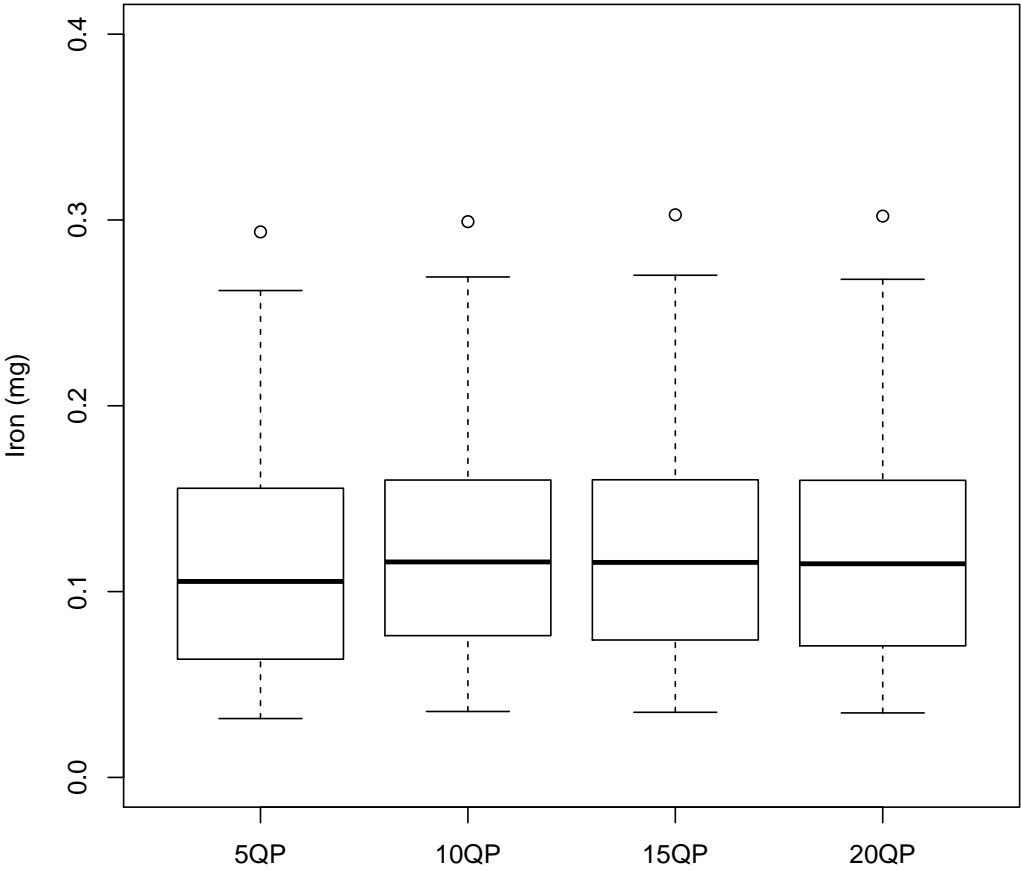
**Table 26**

Difference in time taken for model convergence using 5,10,15 and 20 quadrature points when estimating iron intake from vegetables of 50 bootstrap replicates using data from the NDNS RP Y1-4 (2008-2012) for 4156 participants aged 1.5 years and over.

	Quadrature points			
	5	10	15	20
Average time (h:m:s)	0:18:16	1:41:00	6:16:26	11:33:15
sd (h:m:s)	0:04:15	0:50:53	2:51:36	4:44:21
Total time (h:m:s)	15:13:08	84:09:49	313:41:47	577:42:19

**Figure 54**

A boxplot showing the difference in estimated standard errors of iron intake from vegetables from four models using 5,10,15 and 20 quadrature points using data from NDNS RP Years 1-4 (2008-2012) for 4156 participants aged 1.5 years and over.



## **F Two-part models of complex survey data using a generalised gamma distribution: data table**

Below is a sample of the NDNS RP data used to fit a two-part model for iron intake from bread, where `seriali` is the participant's ID number, typically over 4 rows representing each of the 4 collected days of intake. `BREAD` is the amount of iron consumed in the bread food group this is used to model the amount of iron consumed in part 2 of the two-part model. `IndicatorBread` is a binary variable used to indicate consumption or zero consumption used in the logistic regression model (part 1) of the two-part model. The next variable `agegr1` indicates the age group the participant belongs to, out of a possible 5 groups. the variable `area` refers to the primary sampling unit within the strata named `cluster`. The variable `wti_Y1234` is the sample weighting and `nssec8` is the categorical variable denoting the 9 groups of the NSSEC.

**Table 27**

Sample of NDNS RP Y1-4 (2008-2012) data used to estimate iron intake from the bread food group

seriali	BREAD	IndicatorBread	agegr1	Sex	area	cluster	wti_Y1234	nssec8
10101032	0	0	3	1	10101	1081	0.848841418	6
10101032	4.56768	1	3	1	10101	1081	0.848841418	6
10101032	0	0	3	1	10101	1081	0.848841418	6
10101032	2.45952	1	3	1	10101	1081	0.848841418	6
10101042	0	0	2	1	10101	1081	0.731303388	5
10101042	0	0	2	1	10101	1081	0.731303388	5
10101042	0.864	1	2	1	10101	1081	0.731303388	5
10101042	0	0	2	1	10101	1081	0.731303388	5
10101111	1.728	1	4	1	10101	1081	1.421674093	7
10101111	5.588	1	4	1	10101	1081	1.421674093	7
10101111	4.5502	1	4	1	10101	1081	1.421674093	7
10101111	1.728	1	4	1	10101	1081	1.421674093	7
10101151	4.33536	1	4	0	10101	1081	1.886809257	6
10101151	6.10176	1	4	0	10101	1081	1.886809257	6
10101151	1.94992	1	4	0	10101	1081	1.886809257	6
10101151	1.93536	1	4	0	10101	1081	1.886809257	6
10101161	1.44	1	4	1	10101	1081	0.715086395	7
10101161	1.152	1	4	1	10101	1081	0.715086395	7
10101161	1.152	1	4	1	10101	1081	0.715086395	7
10101161	1.728	1	4	1	10101	1081	0.715086395	7

## G Two-part models of complex survey data using a generalised gamma distribution: code

This script was used to implement the mixed-effects two-part model developed in Chapter 3 and was implemented in SAS software (v9.3) (SAS Institute Inc, 2011). The code is written in two parts: the first to calculate empirical point estimates that do not adjust for the survey design and the second, which produces bootstrapped estimates of variance. The first section of code opens the NDNS RP dataset and creates dummy variables for the five age group categories and nine NSSEC categories.

---

```
1 /*Input NDNS*/
2 proc import datafile="H:\PhD\Dropbox\Two-Part Model\code\ndns.csv" out=NDNS
   dbms=csv replace;
3   getnames=yes;
4   run;
5   DATA NDNS;
6   SET NDNS;
7   agegroup_2 = 0;
8   agegroup_3 = 0;
9   agegroup_4 = 0;
10  agegroup_5 = 0;
11  IF (agegr1=2) THEN agegroup_2 =1;
12  IF (agegr1=3) THEN agegroup_3 =1;
13  IF (agegr1=4) THEN agegroup_4 =1;
14  IF (agegr1=5) THEN agegroup_5 =1;
15  RUN;
16  DATA NDNS;
17  SET NDNS;
18  nssec_2 = 0;
19  nssec_3 = 0;
20  nssec_4 = 0;
21  nssec_5 = 0;
```

```

22 nssec_6 = 0;
23 nssec_7 = 0;
24 nssec_8 = 0;
25 nssec_9 = 0;
26 IF (nssec8=2) THEN nssec_2 =1;
27 IF (nssec8=3) THEN nssec_3 =1;
28 IF (nssec8=4) THEN nssec_4 =1;
29 IF (nssec8=5) THEN nssec_5 =1;
30 IF (nssec8=6) THEN nssec_6 =1;
31 IF (nssec8=7) THEN nssec_7 =1;
32 IF (nssec8=8) THEN nssec_8 =1;
33 IF (nssec8=99) THEN nssec_9 =1;
34 RUN;

```

---

This example is for iron in bread and the continuous variable is denoted as *Bread* with the binary indicator denoted as *IndicatorBread*. The function used to obtain maximum likelihood estimates is *NLMIXED* and arguments of interest include the number of quadrature points specified *QPOINTS* which equals 5 and *TECH = QUANEW* which specifies the optimization algorithm. In this case *QUANEW* refers to the quasi-Newton which approximates second-order derivatives and is thus faster for problems of the size used here. Also specified are the convergence criteria using *GCONV* and *textitmaxfunc*. Model starting parameters were specified using the *PARMS* command, this can be modified to take initial parameters estimated using simple regression. In the following section *Loglikelihood: Part 1 (II1)*, the first part of the two-part model is specified using a logit model with a random intercept then the second part is detailed in *Loglikelihood: Part 1 (II1)*, where the generalised gamma distribution is used and a random intercept is incorporated. The two parts are summed and the random intercepts are allowed to be correlated and the survey weighting is applied using the *REPLICATE* function and the model is then fitted. The correlated random effects are specified at the participant level using the *RANDOM* and *SUBJECT* arguments.

---

```

1 /* define some macro terms */

```



```

2  %let Amount = BREAD;
3  %let Indicator = IndicatorBREAD;
4  /*Save parameter estimates to be used in variance procedure*/
5  /* Reference agegroup is 1 (1.5-3years) */
6  /* Reference nssec8 is 1 (Higher managerial and professional) */
7  /* Reference sex is 0 (Females) */
8  ods output "Parameter Estimates"=parest;
9
          PROC NL MIXED DATA=NDNS MAXITER=10000 QPOINTS=5
          GCONV=1e-3 TECH=QUANEW maxfunc=10000;
10
          PARMS
          a0=1.8 a1=0.4 a2=0.0 a3=-0.5
          a4=0.4 a5=1.0 a6=-0.4 a7=-0.9 a8=-0.2 a9=-0.7
          a10=-1.1 a11=-1.0 a12=-1.0 a13=-0.4
11
          b0=-0.3 b1=0.0 b2=0.3
          b3=0.4 b4=0.8
          b5=0.7 b6=-0.1
          b7=-0.1 b8=-0.0
          b9=-0.1 b10=0.0
          b11=-0.2 b12=-0.1
          b13=-0.1
12
          sda=1.6 sdb=0.2 k=4.4
          d0=1
13
          covab=0.3;
14
15
          *Loglikelihood: Part 1 (l11);
16
          y=a0+au0+a1*sex + a2*agegroup_2 + a3*agegroup_3 +
          a4*agegroup_4 + a5*agegroup_5 +
17
          a6*nssec_2 + a7*nssec_3 + a8*nssec_4 + a9*nssec_5
          + a10*nssec_6 + a11*nssec_7 + a12*nssec_8 +
          a13*nssec_9;
18
          p=exp(y)/(1+exp(y));
19
          l11= log((1-p)**(1-&Indicator)) +
          log(p**(&Indicator));

```

```

20         IF &Indicator=0 THEN L=l11;
21
22         *Loglikelihood: Part 2 (l12);
23         IF &Indicator=1 THEN DO;
24         mu=b0+bu0+ b1*sex + b2*agegroup_2 + b3*agegroup_3
           + b4*agegroup_4 + b5*agegroup_5 +
25         b6*nssec_2 + b7*nssec_3 + b8*nssec_4 + b9*nssec_5
           + b10*nssec_6 + b11*nssec_7 + b12*nssec_8 +
           b13*nssec_9;
26         sigma=exp((d0)/2);
27         eta=abs(k) ** (-2);
28         u=sign(k)*(log(&Amount)-mu)/sigma;
29         value1=eta *log (eta) - log(sigma) -.5 *log(eta) -
           lgamma(eta);
30         l12 = value1 + u *sqrt(eta) - eta * exp(abs(k)*u);
31         L=l11+l12;
32         REPLICATE wti_Y1234;
33         END;
34         MODEL &Amount ~ GENERAL(L);
35         RANDOM au0 bu0 ~ NORMAL([0,0],[sda, covab, sdb])
           SUBJECT=seriali;
36         RUN;
37         quit;
38         /*model end */
39 ods output close;

```

---

The following code details the step used to obtain bootstrapped variance estimates that accounts for the clustering of the NDNS RP. The first step extracts a list of PSUs then a loop is used to sample the PSUs with replacement. The next steps merge the NDNS RP data with the list of PSUs and their selection weights and then repeat the rows based on the selection weight. The model is then ran on each of the created datasets.

---

```

1  /*return total number of clusters*/
2      proc freq data=NDNS noprint;
3      tables area / out=ClusterIDList(drop=count percent);
4      run;
5  /******LOOP THROUGH THIS TO GET N REP WEIGHTS*****/
6  /*produce cluster weights*/
7  %macro ClusterWeights(n);
8      %do i=1 %to &n;
9          proc surveyselect data=ClusterIDList out=Sample&i method=urs
10             n=722 noprint;
11             run;
12 %end;
13 %mend ClusterWeights;
14 %ClusterWeights(50)
15 /******LOOP TO CREATE DATASETS CONTAINING REP WEIGHTS AND NDNS
16     DATA*****/
17 %macro DataGen(n);
18     %do i=1 %to &n;
19         data Sample&i;
20             merge Sample&i(in=sample) NDNS(in=all);
21             by area;
22             if Sample and All;
23             run;
24     %end;
25 %mend DataGen;
26 %DataGen(50)
27 /******LOOP TO REPLICATE OBSERVATIONS BASED ON NEW
28     VARIABLE*****/
29 %macro RepObs(n);
30     %do i=1 %to &n;
31         Data Sample&i (drop=i);
32         set Sample&i;

```

```

30         do i =1 to NumberHits;
31         output;
32         end;
33         run;
34 %end;
35 %mend RepObs;
36 %RepObs(50) /*here RepObs contains the number of iterations to be carried out*/
37 /*-----*/
38 /*ODS TRACE ON;*/
39 /*now n data sets have been created loop through each one and apply the
    following code*/
40 %macro GG2pmBoot(n);
41 %do i=1 %to &n;
42 ODS OUTPUT ParameterEstimates = _parest&i;
43         PROC NLMIXED DATA=Sample&i MAXITER=10000 QPOINTS=5
44         GCONV=1e-3 TECH=QUANEW maxfunc=10000;
45         PARSMS                a0=1.8 a1=0.4 a2=0.0 a3=-0.5
46         a4=0.4 a5=1.0 a6=-0.4 a7=-0.9 a8=-0.2 a9=-0.7
47         a10=-1.1 a11=-1.0 a12=-1.0 a13=-0.4
48         b0=-0.3 b1=0.0 b2=0.3
49         b3=0.4 b4=0.8
50         b5=0.7 b6=-0.1
51         b7=-0.1 b8=-0.0
52         b9=-0.1 b10=0.0
53         b11=-0.2 b12=-0.1
54         b13=-0.1
55         sda=1.6 sdb=0.2 k=4.4
56         d0=1
57         covab=0.3;
58
59         *Loglikelihood: Part 1 (l11);

```

```

50 y=a0+au0+a1*sex + a2*agegroup_2 + a3*agegroup_3 +
    a4*agegroup_4 + a5*agegroup_5 +
51 a6*nssec_2 + a7*nssec_3 + a8*nssec_4 + a9*nssec_5
    + a10*nssec_6 + a11*nssec_7 + a12*nssec_8 +
    a13*nssec_9;
52 p=exp(y)/(1+exp(y));
53 l11= log((1-p)**(1-IndicatorBread)) +
    log(p**(IndicatorBread));
54 IF IndicatorBread=0 THEN L=l11;
55
56 *Loglikelihood: Part 2 (l12);
57 IF IndicatorBread=1 THEN DO;
58 mu=b0+bu0+ b1*sex + b2*agegroup_2 + b3*agegroup_3
    + b4*agegroup_4 + b5*agegroup_5 +
59 b6*nssec_2 + b7*nssec_3 + b8*nssec_4 + b9*nssec_5
    + b10*nssec_6 + b11*nssec_7 + b12*nssec_8 +
    b13*nssec_9;
60 sigma=exp((d0)/2);
61 eta=abs(k) ** (-2);
62 u=sign(k)*(log(BREAD)-mu)/sigma;
63 value1=eta *log (eta) - log(sigma) -.5 *log(eta) -
    lgamma(eta);
64 l12 = value1 + u *sqrt(eta) - eta * exp(abs(k)*u);
65 L=l11+l12;
66 REPLICATE wti_Y1234;
67 END;
68 MODEL BREAD ~ GENERAL(L);
69 RANDOM au0 bu0 ~ NORMAL([0,0],[sda, covab, sdb])
    SUBJECT=seriali;
70 RUN;
71
    %end;
72 %mend GG2pmBoot;

```

73 %GG2pmBoot(50);

74 ODS OUTPUT CLOSE;

---

## H Quantile regression of dietary intake in complex samples: data table

**Table 28** displays a sample of the data used for the analysis described in Chapter 4. SubjectID refers to the participant's ID number, Age gives the participant's age, sex is a binary variable with categories female (2) and male (1), Ironmg indicates the participant's iron intake (mg) for the day, NSSEC8 is the NSSEC group that the participant belongs to. The survey design elements, PSU and Strata, are given in the next two variables along with the survey weighting given by the variable Weighting. Weekday is a binary variable indicating a weekend (0) or weekday (1). Age2 and Age3 are cubic and quadratic terms relating to age.

**Table 28**

Sample of NDNS RP Y1-4 (2008-2012) data used to estimate quantiles of dietary iron intake

Subject ID	Age	Sex	Ironmg	NSSEC8	PSU	Strata	Weighting	Weekday	Age2	Age3
10101032	11	2	7.98378	6	10101	1081	0.848841418	1	1.21	0.1331
10101032	11	2	7.2149	6	10101	1081	0.848841418	1	1.21	0.1331
10101032	11	2	6.52425	6	10101	1081	0.848841418	0	1.21	0.1331
10101032	11	2	7.29087	6	10101	1081	0.848841418	0	1.21	0.1331
10101042	10	2	5.9214	5	10101	1081	0.731303388	0	1	0.1
10101042	10	2	7.9447	5	10101	1081	0.731303388	1	1	0.1
10101042	10	2	5.6914	5	10101	1081	0.731303388	1	1	0.1
10101042	10	2	5.6531	5	10101	1081	0.731303388	0	1	0.1
10101111	32	2	10.6905	7	10101	1081	1.421674093	0	10.24	3.2768
10101111	32	2	7.4665	7	10101	1081	1.421674093	0	10.24	3.2768
10101111	32	2	15.327	7	10101	1081	1.421674093	1	10.24	3.2768
10101111	32	2	22.8288	7	10101	1081	1.421674093	1	10.24	3.2768
10101151	64	1	13.72186	6	10101	1081	1.886809257	0	40.96	26.2144
10101151	64	1	12.45836	6	10101	1081	1.886809257	0	40.96	26.2144
10101151	64	1	18.19694	6	10101	1081	1.886809257	1	40.96	26.2144
10101151	64	1	12.54572	6	10101	1081	1.886809257	1	40.96	26.2144
10101161	61	2	10.188	7	10101	1081	0.715086395	0	37.21	22.6981
10101161	61	2	6.337	7	10101	1081	0.715086395	0	37.21	22.6981
10101161	61	2	7.9696	7	10101	1081	0.715086395	1	37.21	22.6981
10101161	61	2	10.306	7	10101	1081	0.715086395	1	37.21	22.6981



## I Quantile regression of dietary intake in complex samples: code

The following is R code (R Core Team, 2016) (v3.3.2) used to compute the point estimates of the quantile regression parameters using NDNS RP y1-4 data in Chapter 4. Here the 2.5<sup>th</sup>, 25<sup>th</sup>, median, 75<sup>th</sup> and 97.5<sup>th</sup> quantiles of dietary iron intake are modelled as a function of age, age<sup>2</sup>, age<sup>3</sup>, sex, NSSEC and day of the week. A selection weighting is applied to ensure that estimates are adjusted according to survey non-response, for unequal selection probability and to compensate for over- and under-representation of some individuals within the population. Note that the clustering in the NDNS RP survey design will only impact upon variance estimation and is therefore not included in the point estimation code.

The first section of the code calls the `lqmm` and `survey` packages and then opens the NDNS RP data that has undergone some cleaning, following this participants with a missing NSSEC value are removed from the dataset. The weekend indicator variable is created in the next line, with weekdays counted as Monday, Tuesday, Wednesday, Thursday and Friday. Weekend days are Saturday and Sunday.

---

```
1 library(lqmm)
2 library(survey)
3 source("H:/PhD/Analysis/R/Quantile Regression/AmountOnly/NDNSData.r")
4 NDNS <- NDNS[complete.cases(NDNS[,2]),]
5 NDNS$weekday <- ifelse(NDNS$Day.of.Week == "Saturday"|
  NDNS$Day.of.Week=="Sunday",1,0)
6 NDNS$Age2 <- (I(NDNS$Age^2)/100)
7 NDNS$Age3 <- (I(NDNS$Age^3)/10000)
```

---

This section details the model implementation. The first line specifies the model then in the second line gauss-laguerre quadrature is specified using the command `type=robust`, 5 quadrature points are specified. The `tau` argument takes the value for the quantile: here the median is specified. The convergence criteria are given in the `control` arguments.

---

```
1 LQMM05PE <- lqmm(Iron~Age + Age2 + Age3 + Gender + nssec8 + weekday,
2               random=~1, type="robust", group=ISerial, nK=5, tau=0.5,
3               weights=NDNS$wti_Y1234, data=NDNS,
4               control=list(LP_tol_ll=1e-4, LP_max_iter=2000))
```

---

In this part the variance estimates are calculated. The survey design is implemented using the `withReplicates`, `as.svrepdesign` and `svydesign` functions. The two survey design functions include the PSU and strata variables together with a weighting. The `withReplicates` function then takes this information and creates a number of data sets sampled using `type="bootstrap` and the `replicates=50` function. Then the model is ran, in this case 50 times, using the bootstrap weights. The coefficients are then extracted

and averaged to give an output which contains bootstrapped point and variance estimates.

---

```
1 LQMM05 <- withReplicates(as.svrepdesign(  
2   svydesign(id=~point, strata=~strata, weights=~wti_Y1234, data=NDNS), type="bootstrap",  
3     replicates=50),  
4   quote(coef(lqmm(Iron~Age + Age2 + Age3 + Gender + nssec8 + weekday,  
5     random =~1, type="robust", group=ISerial, nK=5, tau=0.5,  
6     weights=.weights, data=NDNS,  
7     control=list(LP_tol_ll=1e-4, LP_max_iter=2000))))))
```

---

## J Iron prescription costs across the UK: code

The following is R code (R Core Team, 2016) (v3.3.2) used to compute the median amount spent on iron medication by health boards in the UK adjusting for the number of registered patients, index of multiple deprivation and a second model as previously that includes bioavailable iron intake. Regression coefficients are presented in Table 30 in Appendix M. The code is in two main parts; the first section of the code involves cleaning and preparing the data prior to fitting models in the second part. The first few lines (2:7) call the relevant libraries used, then the data from each country is compiled into a single data frame (lines 10:13). The next section (lines 16:27) create a data frame with containing index of multiple deprivation values as a percentage, firstly within each country, then overall.

---

```
1 #Libraries
2 library(lqmm) #for quantile regression
3 library(data.table) #for data handling
4 library(plyr) #for faster aggregate
5 library(dplyr) #to create sample data frame
6 library(reshape2) #for long to wide
7 library(doBy) #for summaryBy
8
9 #Combine data into a single file
10 Wales <- as.data.frame(fread("/IronUk/Wales.csv",header=TRUE))
11 England <- as.data.frame(fread("/IronUk/England.csv", header=TRUE))
12 NorthernIreland <- as.data.frame(fread("/IronUK/NorthernIreland.csv",header=TRUE))
13 Scotland <- as.data.frame(fread("/IronUK/Scotland.csv",header=TRUE))
14
15 #CREATE overall IMD percentage
16 IMD <- rbindlist( list (data.frame(PRACTICE=Wales$PRACTICE,IMD=Wales$IMD_RANK/max(Wales$IMD_RANK)),
17 data.frame(PRACTICE=England$PRACTICE,IMD=England$IMD_RANK/max(England$IMD_RANK)),
18 data.frame(PRACTICE=Scotland$PRACTICE,IMD=Scotland$IMD_RANK/max(Scotland$IMD_RANK)),
19 data.frame(PRACTICE=NorthernIreland$PRACTICE,IMD=NorthernIreland$IMD_RANK/max(NorthernIreland$IMD_RANK))))
20
21 IMD <- aggregate(IMD,by=list(PRACTICE=IMD$PRACTICE),FUN=mean)
22 IMD <- IMD[ -c(2) ]
23 IMD$RANK <- rank(IMD$IMD,ties.method= "random")
24
25 #Combine into one object
26 UKData <- as.data.frame(rbindlist( list (Wales,Scotland,England,NorthernIreland), fill=TRUE))
27 UKData <- merge(x = UKData, y = IMD[ , c("PRACTICE", "RANK")], by = "PRACTICE", all.x=TRUE)
```

---

This next section makes adjustments to the names of health boards to ensure they match those used in the map data, removing "NHS", "CCG", "University" and "and" replacing this last case with "&". Following this the health boards are matched to the government office regions used by the NDNS RP (lines 11:245).

---

```

1 UKData$HEALTHBOARD <- gsub('NHS ', ' ', UKData$HEALTHBOARD)
2 UKData$HEALTHBOARD <- gsub(' CCG ', ' ', UKData$HEALTHBOARD)
3 UKData$HEALTHBOARD <- gsub(' University', ' ', UKData$HEALTHBOARD)
4 UKData$HEALTHBOARD <- gsub(' and ', ' & ', UKData$HEALTHBOARD)
5
6 #####
7 #Match HEALTHBOARD to GOR
8 #####
9 UKData$GOR <- NA
10
11 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Cambridgeshire & Peterborough","East of England",UKData$GOR)
12 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Ipswich & East Suffolk","East of England",UKData$GOR)
13 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Herts Valleys","East of England",UKData$GOR)
14 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Mid Essex","East of England",UKData$GOR)
15 UKData$GOR=ifelse(UKData$HEALTHBOARD=="North Norfolk","East of England",UKData$GOR)
16 UKData$GOR=ifelse(UKData$HEALTHBOARD=="South Norfolk","East of England",UKData$GOR)
17 UKData$GOR=ifelse(UKData$HEALTHBOARD=="West Essex","East of England",UKData$GOR)
18 UKData$GOR=ifelse(UKData$HEALTHBOARD=="West Suffolk","East of England",UKData$GOR)
19 UKData$GOR=ifelse(UKData$HEALTHBOARD=="East & North Hertfordshire","East of England",UKData$GOR)
20 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Great Yarmouth & Waveney","East of England",UKData$GOR)
21 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Luton","East of England",UKData$GOR)
22 UKData$GOR=ifelse(UKData$HEALTHBOARD=="North East Essex","East of England",UKData$GOR)
23 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Norwich","East of England",UKData$GOR)
24 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Thurrock","East of England",UKData$GOR)
25 UKData$GOR=ifelse(UKData$HEALTHBOARD=="West Norfolk","East of England",UKData$GOR)
26 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Basildon & Brentwood","East of England",UKData$GOR)
27 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Southend","East of England",UKData$GOR)
28 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Corby","East Midlands",UKData$GOR)
29 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Erewash","East Midlands",UKData$GOR)
30 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Leicester City","East Midlands",UKData$GOR)
31 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Mansfield & Ashfield","East Midlands",UKData$GOR)
32 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Nene","East Midlands",UKData$GOR)
33 UKData$GOR=ifelse(UKData$HEALTHBOARD=="North Derbyshire","East Midlands",UKData$GOR)
34 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Nottingham North & East","East Midlands",UKData$GOR)
35 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Rushcliffe","East Midlands",UKData$GOR)
36 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Southern Derbyshire","East Midlands",UKData$GOR)
37 UKData$GOR=ifelse(UKData$HEALTHBOARD=="East Leicestershire & Rutland","East Midlands",UKData$GOR)
38 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Hardwick","East Midlands",UKData$GOR)
39 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Lincolnshire West","East Midlands",UKData$GOR)
40 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Milton Keynes","East Midlands",UKData$GOR)

```

41 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Newark & Sherwood","East Midlands",UKData\$GOR)

42 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Nottingham City","East Midlands",UKData\$GOR)

43 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Nottingham West","East Midlands",UKData\$GOR)

44 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South West Lincolnshire","East Midlands",UKData\$GOR)

45 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="West Leicestershire","East Midlands",UKData\$GOR)

46 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bedfordshire","East Midlands",UKData\$GOR)

47 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Barnet","London",UKData\$GOR)

48 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Brent","London",UKData\$GOR)

49 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Camden","London",UKData\$GOR)

50 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Croydon","London",UKData\$GOR)

51 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Enfield","London",UKData\$GOR)

52 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Greenwich","London",UKData\$GOR)

53 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Haringey","London",UKData\$GOR)

54 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Havering","London",UKData\$GOR)

55 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Islington","London",UKData\$GOR)

56 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Lambeth","London",UKData\$GOR)

57 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Newham","London",UKData\$GOR)

58 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Richmond","London",UKData\$GOR)

59 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Merton","London",UKData\$GOR)

60 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Tower Hamlets","London",UKData\$GOR)

61 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wandsworth","London",UKData\$GOR)

62 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Central London (Westminster)","London",UKData\$GOR)

63 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Barking & Dagenham","London",UKData\$GOR)

64 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bexley","London",UKData\$GOR)

65 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bromley","London",UKData\$GOR)

66 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="City & Hackney","London",UKData\$GOR)

67 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Ealing","London",UKData\$GOR)

68 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Hounslow","London",UKData\$GOR)

69 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Hammersmith & Fulham","London",UKData\$GOR)

70 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Harrow","London",UKData\$GOR)

71 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Hillingdon","London",UKData\$GOR)

72 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Kingston","London",UKData\$GOR)

73 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Lewisham","London",UKData\$GOR)

74 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Redbridge","London",UKData\$GOR)

75 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Southwark","London",UKData\$GOR)

76 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Sutton","London",UKData\$GOR)

77 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Waltham Forest","London",UKData\$GOR)

78 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="West London","London",UKData\$GOR)

79 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Durham Dales, Easington & Sedgefield","North East",UKData\$GOR)

80 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Hartlepool & Stockton-on-Tees","North East",UKData\$GOR)

81 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Tees","North East",UKData\$GOR)

82 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Sunderland","North East",UKData\$GOR)

83 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Darlington","North East",UKData\$GOR)

84 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North Durham","North East",UKData\$GOR)

85 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Northumberland","North East",UKData\$GOR)

86 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Tyneside","North East",UKData\$GOR)

87 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North Tyneside","North East",UKData\$GOR)

88 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Newcastle Gateshead","North East",UKData\$GOR)

89 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Blackpool","North West",UKData\$GOR)

90 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bury","North West",UKData\$GOR)

91 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Chorley & South Ribble","North West",UKData\$GOR)

92 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="East Lancashire","North West",UKData\$GOR)

93 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Heywood, Middleton & Rochdale","North West",UKData\$GOR)

94 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Halton","North West",UKData\$GOR)

95 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Cumbria","North West",UKData\$GOR)

96 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Lancashire North","North West",UKData\$GOR)

97 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Manchester","North West",UKData\$GOR)

98 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Sefton","North West",UKData\$GOR)

99 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Stockport","North West",UKData\$GOR)

100 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Tameside & Glossop","North West",UKData\$GOR)

101 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Vale Royal","North West",UKData\$GOR)

102 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="West Cheshire","North West",UKData\$GOR)

103 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wigan Borough","North West",UKData\$GOR)

104 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Airedale, Wharfedale & Craven","North West",UKData\$GOR)

105 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bassetlaw","North West",UKData\$GOR)

106 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Calderdale","North West",UKData\$GOR)

107 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wirral","North West",UKData\$GOR)

108 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Liverpool","North West",UKData\$GOR)

109 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Blackburn with Darwen","North West",UKData\$GOR)

110 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bolton","North West",UKData\$GOR)

111 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Central Manchester","North West",UKData\$GOR)

112 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Oldham","North West",UKData\$GOR)

113 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Eastern Cheshire","North West",UKData\$GOR)

114 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Greater Preston","North West",UKData\$GOR)

115 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Salford","North West",UKData\$GOR)

116 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Knowsley","North West",UKData\$GOR)

117 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North Manchester","North West",UKData\$GOR)

118 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Cheshire","North West",UKData\$GOR)

119 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Southport & Formby","North West",UKData\$GOR)

120 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="St Helens","North West",UKData\$GOR)

121 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Trafford","North West",UKData\$GOR)

122 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Warrington","North West",UKData\$GOR)

123 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="West Lancashire","North West",UKData\$GOR)

124 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Fylde & Wyre","North West",UKData\$GOR)

125 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Northern","Northern Ireland",UKData\$GOR)

126 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Belfast","Northern Ireland",UKData\$GOR)

127 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Western","Northern Ireland",UKData\$GOR)

128 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Eastern","Northern Ireland",UKData\$GOR)

129 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Southern","Northern Ireland",UKData\$GOR)

130 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Borders","Scotland",UKData\$GOR)

131 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Fife","Scotland",UKData\$GOR)

132 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Grampian","Scotland",UKData\$GOR)

133 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Highland","Scotland",UKData\$GOR)

134 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Lothian","Scotland",UKData\$GOR)

135 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Shetland","Scotland",UKData\$GOR)

136 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Western Isles","Scotland",UKData\$GOR)

137 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Ayrshire & Arran","Scotland",UKData\$GOR)

138 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Dumfries & Galloway","Scotland",UKData\$GOR)

139 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Forth Valley","Scotland",UKData\$GOR)

140 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Greater Glasgow & Clyde","Scotland",UKData\$GOR)

141 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Lanarkshire","Scotland",UKData\$GOR)

142 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Orkney","Scotland",UKData\$GOR)

143 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Tayside","Scotland",UKData\$GOR)

144 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Brighton & Hove","South East",UKData\$GOR)

145 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Eastbourne, Hailsham & Seaford","South East",UKData\$GOR)

146 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Crawley","South East",UKData\$GOR)

147 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="East Surrey","South East",UKData\$GOR)

148 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Hastings & Rother","South East",UKData\$GOR)

149 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Horsham & Mid Sussex","South East",UKData\$GOR)

150 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Kent Coast","South East",UKData\$GOR)

151 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Swale","South East",UKData\$GOR)

152 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bracknell & Ascot","South East",UKData\$GOR)

153 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North Hampshire","South East",UKData\$GOR)

154 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Isle of Wight","South East",UKData\$GOR)

155 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North & West Reading","South East",UKData\$GOR)

156 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Portsmouth","South East",UKData\$GOR)

157 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Eastern Hampshire","South East",UKData\$GOR)

158 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Southampton","South East",UKData\$GOR)

159 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="West Hampshire","South East",UKData\$GOR)

160 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wokingham","South East",UKData\$GOR)

161 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Castle Point & Rochford","South East",UKData\$GOR)

162 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Surrey Downs","South East",UKData\$GOR)

163 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="High Weald Lewes Havens","South East",UKData\$GOR)

164 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Ashford","South East",UKData\$GOR)

165 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Canterbury & Coastal","South East",UKData\$GOR)

166 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Coastal West Sussex","South East",UKData\$GOR)

167 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Dartford, Gravesham & Swanley","South East",UKData\$GOR)

168 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Guildford & Waverley","South East",UKData\$GOR)

169 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Medway","South East",UKData\$GOR)

170 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North West Surrey","South East",UKData\$GOR)

171 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Surrey Heath","South East",UKData\$GOR)

172 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Thanet","South East",UKData\$GOR)

173 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Chiltern","South East",UKData\$GOR)

174 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Fareham & Gosport","South East",UKData\$GOR)

175 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Newbury & District","South East",UKData\$GOR)

176 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Oxfordshire","South East",UKData\$GOR)

177 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Slough","South East",UKData\$GOR)

178 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Reading","South East",UKData\$GOR)



179 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Aylesbury Vale","South East",UKData\$GOR)

180 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Windsor, Ascot & Maidenhead","South East",UKData\$GOR)

181 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="West Kent","South East",UKData\$GOR)

182 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North East Hampshire & Farnham","South East",UKData\$GOR)

183 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bristol","South West",UKData\$GOR)

184 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Gloucestershire","South West",UKData\$GOR)

185 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North Somerset","South West",UKData\$GOR)

186 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Gloucestershire","South West",UKData\$GOR)

187 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wiltshire","South West",UKData\$GOR)

188 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Devon & Torbay","South West",UKData\$GOR)

189 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Bath & North East Somerset","South West",UKData\$GOR)

190 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Dorset","South West",UKData\$GOR)

191 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Kernow","South West",UKData\$GOR)

192 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Somerset","South West",UKData\$GOR)

193 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Swindon","South West",UKData\$GOR)

194 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Northern, Eastern & Western Devon","South West",UKData\$GOR)

195 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Abertawe Bro Morgannwg","Wales",UKData\$GOR)

196 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Betsi Cadwaladr","Wales",UKData\$GOR)

197 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Cwm Taf","Wales",UKData\$GOR)

198 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Powys","Wales",UKData\$GOR)

199 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Aneurin Bevan","Wales",UKData\$GOR)

200 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Cardiff & Vale","Wales",UKData\$GOR)

201 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Hywel Dda","Wales",UKData\$GOR)

202 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Birmingham South & Central","West Midlands",UKData\$GOR)

203 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Coventry & Rugby","West Midlands",UKData\$GOR)

204 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="East Staffordshire","West Midlands",UKData\$GOR)

205 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="North Staffordshire","West Midlands",UKData\$GOR)

206 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Redditch & Bromsgrove","West Midlands",UKData\$GOR)

207 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Shropshire","West Midlands",UKData\$GOR)

208 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South East Staffordshire & Seisdon Peninsula","West Midlands",UKData\$GOR)

209 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Worcestershire","West Midlands",UKData\$GOR)

210 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Stoke on Trent","West Midlands",UKData\$GOR)

211 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Walsall","West Midlands",UKData\$GOR)

212 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wyre Forest","West Midlands",UKData\$GOR)

213 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Cannock Chase","West Midlands",UKData\$GOR)

214 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Dudley","West Midlands",UKData\$GOR)

215 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Herefordshire","West Midlands",UKData\$GOR)

216 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Warwickshire North","West Midlands",UKData\$GOR)

217 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Sandwell & West Birmingham","West Midlands",UKData\$GOR)

218 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Solihull","West Midlands",UKData\$GOR)

219 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="South Warwickshire","West Midlands",UKData\$GOR)

220 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Stafford & Surrounds","West Midlands",UKData\$GOR)

221 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Telford & Wrekin","West Midlands",UKData\$GOR)

222 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Wolverhampton","West Midlands",UKData\$GOR)

223 UKData\$GOR=ifelse(UKData\$HEALTHBOARD=="Birmingham Crosscity","West Midlands",UKData\$GOR)

```

224 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Bradford City","Yorkshire and the Humber",UKData$GOR)
225 UKData$GOR=ifelse(UKData$HEALTHBOARD=="East Riding of Yorkshire","Yorkshire and the Humber",UKData$GOR)
226 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Leeds West","Yorkshire and the Humber",UKData$GOR)
227 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Harrogate & Rural District","Yorkshire and the Humber",UKData$GOR)
228 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Leeds South & East","Yorkshire and the Humber",UKData$GOR)
229 UKData$GOR=ifelse(UKData$HEALTHBOARD=="North Kirklees","Yorkshire and the Humber",UKData$GOR)
230 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Rotherham","Yorkshire and the Humber",UKData$GOR)
231 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Sheffield","Yorkshire and the Humber",UKData$GOR)
232 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Wakefield","Yorkshire and the Humber",UKData$GOR)
233 UKData$GOR=ifelse(UKData$HEALTHBOARD=="South Lincolnshire","Yorkshire and the Humber",UKData$GOR)
234 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Barnsley","Yorkshire and the Humber",UKData$GOR)
235 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Bradford Districts","Yorkshire and the Humber",UKData$GOR)
236 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Leeds North","Yorkshire and the Humber",UKData$GOR)
237 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Doncaster","Yorkshire and the Humber",UKData$GOR)
238 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Greater Huddersfield","Yorkshire and the Humber",UKData$GOR)
239 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Hambleton, Richmondshire & Whitby","Yorkshire and the
      Humber",UKData$GOR)
240 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Hull","Yorkshire and the Humber",UKData$GOR)
241 UKData$GOR=ifelse(UKData$HEALTHBOARD=="North East Lincolnshire","Yorkshire and the Humber",UKData$GOR)
242 UKData$GOR=ifelse(UKData$HEALTHBOARD=="North Lincolnshire","Yorkshire and the Humber",UKData$GOR)
243 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Scarborough & Ryedale","Yorkshire and the Humber",UKData$GOR)
244 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Vale of York","Yorkshire and the Humber",UKData$GOR)
245 UKData$GOR=ifelse(UKData$HEALTHBOARD=="Lincolnshire East","Yorkshire and the Humber",UKData$GOR)

```

---

The following continues with the preparation of the data prior to model fitting and includes collapsing age groups provided to where extra information on age is provided, reducing the size of the data frame, dropping practices with 0 patients and merging bioavailable iron data.

```

1 #####
2 #merge old age groups
3 #####
4 UKData$M65plus <- UKData$M6574 + UKData$M7584 + UKData$M85plus
5 UKData$F65plus <- UKData$F6574 + UKData$F7584 + UKData$F85plus
6 UKData <- UKData[, c("PRACTICE", "BNFCODE", "ITEMS", "COST", "QUANTITY", "PERIOD", "COUNTRY.ID", "HEALTHBOARD",
7 "M04", "M514", "M1544", "M4564", "M65plus", "F04", "F514", "F1544", "F4564", "F65plus",
8 "IMD_RANK", "RANK", "GOR"), drop = FALSE]
9
10 #make wales country id uppercase
11 UKData$COUNTRY.ID <- toupper(UKData$COUNTRY.ID)
12
13 #drop practices without patients
14 UKData <- UKData[UKData$PRACTICE!=30561 & UKData$PRACTICE!=46589, ]
15 UKData <- UKData[UKData$PRACTICE!=65931 & UKData$PRACTICE!=70963, ]

```

```

16
17 #Open adjusted NDNS Iron
18 source("IronAlgorithm.R")
19
20 #create a mean iron intake value for each age and sex group
21 #but first create age groups
22 NDNS$AgeGroups <- cut(NDNS$age, c(0,4,14,44,64,max(NDNS$age)))
23
24 #recode NDNS gor into names
25 NDNS$GOR <- as.factor(NA)
26 NDNS$GOR <- ifelse(NDNS$gor==1,"North East",NDNS$GOR)
27 NDNS$GOR <- ifelse(NDNS$gor==2,"North West",NDNS$GOR)
28 NDNS$GOR <- ifelse(NDNS$gor==3,"Yorkshire and the Humber",NDNS$GOR)
29 NDNS$GOR <- ifelse(NDNS$gor==4,"East Midlands",NDNS$GOR)
30 NDNS$GOR <- ifelse(NDNS$gor==5,"West Midlands",NDNS$GOR)
31 NDNS$GOR <- ifelse(NDNS$gor==6,"East of England",NDNS$GOR)
32 NDNS$GOR <- ifelse(NDNS$gor==7,"London",NDNS$GOR)
33 NDNS$GOR <- ifelse(NDNS$gor==8,"South East",NDNS$GOR)
34 NDNS$GOR <- ifelse(NDNS$gor==9,"South West",NDNS$GOR)
35 NDNS$GOR <- ifelse(NDNS$gor==10,"Wales",NDNS$GOR)
36 NDNS$GOR <- ifelse(NDNS$gor==11,"Scotland",NDNS$GOR)
37 NDNS$GOR <- ifelse(NDNS$gor==12,"Northern Ireland",NDNS$GOR)
38
39 #create model
40 model <- lqmm(AvailableIron~GOR+AgeGroups+as.factor(Sex), random = ~ 1, group=seriali, data=NDNS)
41
42 #create data frame with variable groups
43 basic_summ = data.frame(summarise(group_by(NDNS, Sex, AgeGroups, GOR)))
44
45 #combine predicted values by multiplying design matrix by model coefficients
46 results <- cbind(basic_summ,
47 Iron = as.matrix(model.matrix(~basic_summ$GOR + basic_summ$AgeGroups + as.factor(basic_summ$Sex))) %*%
48 as.vector(model$theta[1:17]))
49
50 #aggregate to the practice level to speed things up
51 aggdata <- ddply(UKData, "PRACTICE", head, 1)
52
53 #create a dataframe with columns to work out percentages on
54 pc <- data.frame(aggdata[,c("PRACTICE", "M04", "M514", "M1544", "M4564", "M65plus",
55 "F04", "F514", "F1544", "F4564", "F65plus")])
56 pc[,c(2:11)] = apply(pc[,c(2:11)], 2, function(x) as.numeric(as.character(x)))
57
58 #work out percentage contribution of each age group to total
59 pc <- cbind(pc[1], prop.table(as.matrix(pc[-1]), margin = 1))
60
61 #merge in gor

```

```

62  aggdata <- merge(x=pc,y=aggdata[,c("PRACTICE","GOR")], by="PRACTICE",all=T)
63
64  #Go Long to wide
65  IronWide <- setnames(dcast(data=results, GOR ~ Sex+AgeGroups, value.var =
        "Iron"),c("GOR","M04","M514","M1544","M4564","M65plus"
66  ,"F04","F514","F1544","F4564","F65plus"))
67  #merge on GOR
68  Iron <- merge(x=aggdata,y=IronWide,by="GOR", all.x=T)
69
70  #multiply columns together
71  Iron$M04 <- Iron$M04.x * Iron$M04.y
72  Iron$M514 <- Iron$M514.x * Iron$M514.y
73  Iron$M1544 <- Iron$M1544.x * Iron$M1544.y
74  Iron$M4564 <- Iron$M4564.x * Iron$M4564.y
75  Iron$M65plus <- Iron$M65plus.x * Iron$M65plus.y
76  Iron$F04 <- Iron$F04.x * Iron$F04.y
77  Iron$F514 <- Iron$F514.x * Iron$F514.y
78  Iron$F1544 <- Iron$F1544.x * Iron$F1544.y
79  Iron$F4564 <- Iron$F4564.x * Iron$F4564.y
80  Iron$F65plus <- Iron$F65plus.x * Iron$F65plus.y
81
82  #drop variables
83  drops <- c("M04.x","M04.y","M514.x","M514.y","M1544.x","M1544.y","M4564.x","M4564.y","M65plus.x","M65plus.y",
84  "F04.x","F04.y","F514.x","F514.y","F1544.x","F1544.y","F4564.x","F4564.y","F65plus.x","F65plus.y")
85  Iron <- Iron[ , !(names(Iron) %in% drops)]
86
87  #add iron intakes together
88  Iron$Iron <- rowSums(Iron[,3:12])
89
90  #tidy up
91  rm(list = setdiff (ls () , c("Iron", "UKData")))
92
93  #merge with ukdata
94  Iron <- merge(x=Iron[,c("PRACTICE","Iron")],
        y=UKData[,c("PRACTICE","COST","HEALTHBOARD","IMD_RANK")],by="PRACTICE",all.x=T)
95
96  #make healthboard a factor
97  Iron$HEALTHBOARD <- as.factor(Iron$HEALTHBOARD)
98
99  #####
100 #add in total number of patients per practice
101 #scaled row sum
102 # UKData$TotalPatients <- scale(rowSums(UKData[,c("M04", "M514", "M1544","M4564",
        "M65plus","F04","F514","F1544","F4564","F65plus"))))
103 UKData$TotalPatients <- rowSums(UKData[,c("M04", "M514", "M1544","M4564",
        "M65plus","F04","F514","F1544","F4564","F65plus"))))

```

```

104 #aggregate to the practice level to speed things up...although its not very fast
105 UKData <- ddply(UKData[,c("PRACTICE","TotalPatients")], "PRACTICE", head, 1)
106 #merge
107 Iron <- merge(x=Iron,y=UKData, by=c("PRACTICE"), all.x=TRUE)
108
109 #sum prescription cost at practice level
110 Iron <- merge(x=aggregate(COST ~ PRACTICE, data=Iron, sum),
111 y= unique(Iron[,c("PRACTICE","Iron","HEALTHBOARD","IMD_RANK","TotalPatients")]),
112 by=c("PRACTICE"), all = TRUE)

```

---

In this final section estimates for both models (including and excluding bioavailable iron) are fitted, this includes an explicit command to use Lincolnshire West as the reference health board and tau = 0.5 specifies the median value.

---

```

1 HealthboardIronIMDTotPatients <- summary(lqmm(COST ~ IMD_RANK + Iron + TotalPatients + relevel(HEALTHBOARD, ref
= "Lincolnshire West"), random=~1, group=PRACTICE, tau=0.5,data=Iron,
2 control = lqmmControl(method="gs",UP_max_iter = 200, LP_tol_ll = 1e-2, LP_max_iter = 10000)))
3
4 HealthboardIMDTotPatients <- summary(lqmm(COST ~ IMD_RANK + TotalPatients + relevel(HEALTHBOARD, ref =
"Lincolnshire West"), random=~1, group=PRACTICE, tau=0.5,data=Iron,
5 control = lqmmControl(method="gs",UP_max_iter = 200, LP_tol_ll = 1e-2, LP_max_iter = 10000)))

```

---

## Welsh map

---

```

1
2 library ( latticeExtra )
3 library ( maptools )
4 library ( gpclib )
5 library ( sp ) #for joining spatialpolygons
6 library ( raster ) #to join spatialpolygons – union function
7 library ( rgdal )
8 library ( rgeos ) #unionspatialpolygons
9 library ( readxl )#read.xls
10
11 gadm <- readRDS("IronUK/GBR_adm2.rds")
12 gadm@data$HealthRegion <- NA
13
14 #-----#
15 #WALES
16 #-----#
17
18 #Abertawe Bro Morgannwg
19 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Bridgend","Abertawe Bro
Morgannwg",gadm@data$HealthRegion)

```

```

20 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Neath Port Talbot","Abertawe Bro
    Morgannwg",gadm@data$HealthRegion)
21 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Swansea","Abertawe Bro
    Morgannwg",gadm@data$HealthRegion)
22 #Aneurin Bevan
23 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Blaenau Gwent","Aneurin Bevan",gadm@data$HealthRegion)
24 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Caerphilly","Aneurin Bevan",gadm@data$HealthRegion)
25 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Monmouthshire","Aneurin Bevan",gadm@data$HealthRegion)
26 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Newport","Aneurin Bevan",gadm@data$HealthRegion)
27 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Torfaen","Aneurin Bevan",gadm@data$HealthRegion)
28 #Betsi Cadwaladr
29 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Anglesey","Betsi Cadwaladr",gadm@data$HealthRegion)
30 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Conwy","Betsi Cadwaladr",gadm@data$HealthRegion)
31 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Denbighshire","Betsi Cadwaladr",gadm@data$HealthRegion)
32 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Flintshire","Betsi Cadwaladr",gadm@data$HealthRegion)
33 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Gwynedd","Betsi Cadwaladr",gadm@data$HealthRegion)
34 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Wrexham","Betsi Cadwaladr",gadm@data$HealthRegion)
35 #Cardiff & Vale University
36 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Cardiff","Cardiff & Vale",gadm@data$HealthRegion)
37 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Vale of Glamorgan","Cardiff & Vale",gadm@data$HealthRegion)
38 #Cwm Taf
39 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Merthyr Tydfil","Cwm Taf",gadm@data$HealthRegion)
40 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Rhondda, Cynon, Taff","Cwm Taf",gadm@data$HealthRegion)
41 #Hywel Dda
42 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Carmarthenshire","Hywel Dda",gadm@data$HealthRegion)
43 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Ceredigion","Hywel Dda",gadm@data$HealthRegion)
44 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Pembrokeshire","Hywel Dda",gadm@data$HealthRegion)
45 #Powys Teaching
46 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Powys","Powys",gadm@data$HealthRegion)
47
48 #this function firstly takes the text argument removes formatting (spaces and &), then extracts the matching data to a
    seperate data frame
49 #the id number is set to start from zero
50 #then borders within areas matching the area are removed
51 #assign allows strings to be modified and then saved as an object
52 area.sub <- function(region){
53   Carrier <- assign(paste0(gsub("[:punct :]]\\ s", "", region)), gadm[ grep(region, gadm@data$HealthRegion) , ])
54   Carrier@data$id <- seq(length(Carrier@data$OBJECTID)) - 1 ## add 'id' column
55   Carrier.sub <- unionSpatialPolygons(Carrier, IDs = rep(1, length(Carrier))) ## unify polygons
56   return(Carrier.sub)
57 }
58
59 #Create a new environment that takes all of the newly created objects in the loop
60 Country.env <- new.env()
61
62 for (i in 171:192){#pretty sure this just over writes previously created health board spatialpolygons within the loop

```

```

63   assign(paste0(gsub("[[:punct :]]\\ s", "", .gadm@data$HealthRegion[i]),area.sub(region=gadm@data$HealthRegion[i]),
64   envir=Country.env)
65 }
66
67
68 #this takes all of this objects in the environment (ls()) as objects (eval(parse(text=paste(i))))
69 #and combines them into a list (lapply) then into an object of spatialpolygon class (bind)
70 Wales <- do.call(bind,lapply(ls(Country.env), function(i) {(eval(parse(text=paste0("Country.env$",i))))}))
71 #Combine into dataframe
72 Wales = SpatialPolygonsDataFrame(Sr=Wales, data=data.frame(HealthRegions=ls(Country.env)),FALSE)
73
74 #Tidy up
75 rm(Country.env,i,area.sub,gadm)

```

---

## English Map

---

```

1 #current as of Apr 2015
2 #downloaded from here: https://www.england.nhs.uk/resources/ccg-maps/
3 #ccg
4 England <- readOGR(dsn="IronUK/England/EnglandMap/ccg-boundaries-0415-tab/CCG_BSC_Apr2015.TAB",
   layer="CCG_BSC_Apr2015")
5 England<-spTransform(England,CRS("+proj=longlat")) #change scale to longlat from UTM
6 #Rename ccgs to Health regions
7 colnames(England@data)[colnames(England@data) == "CCGname_short"] <- "HealthRegions"

```

---

## Northern Ireland map

---

```

1
2 #Created on 11/7/2016
3 #https://www.opendatani.gov.uk/dataset/department-of-health-trust-boundaries
4 NorthernIreland = readOGR("IronUK/NorthernIreland/NorthernIrelandMap/health-trust-boundaries.geojson", "OGRGeoJSON")
5 #Currently, because of lakes and whatnot, there are 22 areas for the 5 health boards combine these into 5
6
7 NorthernIreland@data$Id <- NA
8 NorthernIreland@data$Id <- ifelse(NorthernIreland@data$LGDNAMER=="Northern Trust",1,NorthernIreland@data$Id)
9 NorthernIreland@data$Id <- ifelse(NorthernIreland@data$LGDNAMER=="South Eastern Trust",2,NorthernIreland@data$Id)
10 NorthernIreland@data$Id <- ifelse(NorthernIreland@data$LGDNAMER=="Belfast Trust",3,NorthernIreland@data$Id)
11 NorthernIreland@data$Id <- ifelse(NorthernIreland@data$LGDNAMER=="Southern Trust",4,NorthernIreland@data$Id)
12 NorthernIreland@data$Id <- ifelse(NorthernIreland@data$LGDNAMER=="Western Trust",5,NorthernIreland@data$Id)
13 NorthernIreland_union <- unionSpatialPolygons(NorthernIreland, IDs = NorthernIreland@data$Id) ## unify polygons
14 NorthernIreland <- SpatialPolygonsDataFrame(NorthernIreland_union,data=data.frame(HealthRegions = c("Northern",
15 "South Eastern",
16 "Belfast",
17 "Southern",
18 "Western")))

```

```
19 rm(NorthernIreland_union)
```

---

## Scottish Map

---

```
1
2 library ( latticeExtra )
3 library ( maptools )
4 library ( gpclib )
5 library ( sp ) #for joining spatialpolygons
6 library ( raster ) #to join spatialpolygons – union function
7 library ( rgdal )
8
9 gadm <- readRDS("IronUK/GBR_adm2.rds")
10 gadm@data$HealthRegion <- NA
11
12 # -----#
13 #CREATE GROUPS
14 # -----#
15
16 #Ayrshire & Arran
17 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="East Ayrshire","Ayrshire & Arran",gadm@data$HealthRegion)
18 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="North Ayrshire","Ayrshire & Arran",gadm@data$HealthRegion)
19 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="South Ayrshire","Ayrshire & Arran",gadm@data$HealthRegion)
20 #Borders
21 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Scottish Borders","Borders",gadm@data$HealthRegion)
22 #Dumfries & Galloway
23 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Dumfries and Galloway","Dumfries &
    Galloway",gadm@data$HealthRegion)
24 #Fife
25 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Fife","Fife",gadm@data$HealthRegion)
26 #Forth Valley
27 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Clackmannanshire","Forth Valley",gadm@data$HealthRegion)
28 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Falkirk","Forth Valley",gadm@data$HealthRegion)
29 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Stirling","Forth Valley",gadm@data$HealthRegion)
30 #Grampian
31 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Aberdeenshire","Grampian",gadm@data$HealthRegion)
32 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Moray","Grampian",gadm@data$HealthRegion)
33 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Aberdeen","Grampian",gadm@data$HealthRegion)
34 #Greater Glasgow & Clyde
35 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="East Dunbartonshire","Greater Glasgow &
    Clyde",gadm@data$HealthRegion)
36 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="East Renfrewshire","Greater Glasgow &
    Clyde",gadm@data$HealthRegion)
37 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Inverclyde","Greater Glasgow &
    Clyde",gadm@data$HealthRegion)
```



```

38 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Renfrewshire","Greater Glasgow &
    Clyde",gadm@data$HealthRegion)
39 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="West Dunbartonshire","Greater Glasgow &
    Clyde",gadm@data$HealthRegion)
40 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Glasgow","Greater Glasgow &
    Clyde",gadm@data$HealthRegion)
41 #Highland
42 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Argyll and Bute","Highland",gadm@data$HealthRegion)
43 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Highland","Highland",gadm@data$HealthRegion)
44 #Lanarkshire
45 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="North Lanarkshire","Lanarkshire",gadm@data$HealthRegion)
46 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="South Lanarkshire","Lanarkshire",gadm@data$HealthRegion)
47 #Lothian
48 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="East Lothian","Lothian",gadm@data$HealthRegion)
49 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Midlothian","Lothian",gadm@data$HealthRegion)
50 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="West Lothian","Lothian",gadm@data$HealthRegion)
51 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Edinburgh","Lothian",gadm@data$HealthRegion)
52 #Orkney
53 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Orkney Islands","Orkney",gadm@data$HealthRegion)
54 #Shetland
55 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Shetland Islands","Shetland",gadm@data$HealthRegion)
56 #Tayside
57 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Angus","Tayside",gadm@data$HealthRegion)
58 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Perthshire and Kinross","Tayside",gadm@data$HealthRegion)
59 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Dundee","Tayside",gadm@data$HealthRegion)
60 #Western Isles
61 gadm@data$HealthRegion <- ifelse(gadm@data$NAME_2=="Eilean Siar","Western Isles",gadm@data$HealthRegion)
62
63 #this function firstly takes the text argument removes formatting (spaces and &), then extracts the matching data to a
    seperate data frame
64 #the id number is set to start from zero
65 #then borders within areas matching the area are removed
66 #assign allows strings to be modified and then saved as an object
67 area.sub <- function(region){
68   Carrier <- assign(paste0(gsub("[[:punct :]]\\ s", "", region)),gadm[ grep(region, gadm@data$HealthRegion) , ])
69   Carrier@data$id <- seq(length(Carrier@data$OBJECTID)) - 1 ## add 'id' column
70   Carrier.sub <- unionSpatialPolygons(Carrier, IDs = rep(1, length(Carrier))) ## unify polygons
71   return(Carrier.sub)
72 }
73
74 #Create a new environment that takes all of the newly created objects in the loop
75 Country.env <- new.env()
76 ls(Country.env)
77 for (i in 141:170){#pretty sure this just over writes previously create health board spatialpolygons
78   assign(paste0(gsub("[[:punct :]]\\ s", "", gadm@data$HealthRegion[i]),area.sub(region=gadm@data$HealthRegion[i]),
79     envir=Country.env)

```

```

80 }
81
82 #this takes all of this objects in the environment (ls()) as objects (eval(parse(text=paste(i))))
83 #and combines them into a list (lapply) then into an object of spatialpolygon class (bind)
84 Scotland <- do.call(bind,lapply(ls(Country.env), function(i) {(eval(parse(text=paste0("Country.env$",i))))}))
85 #then combine them in THE SAME ORDER AS IN THE ENVIRONMENT into a labelled spatialpolygonsdataframe
86 #Note there is no check performed to ensure that the correct area matches the coordinates it is done on order and then once
    the graphs are constructed by eye to see if the coordinates match the label
87 Scotland = SpatialPolygonsDataFrame(Sr=Scotland, data=data.frame(HealthRegions=ls(Country.env)),FALSE)
88
89 #Tidy up
90 rm(Country.env,i,area.sub,gadm)

```

---

## Combined mapping

---

```

1
2 #CONVERT LQMM OUTPUT TO DATA FRAME
3 HITP <- data.frame(HealthboardIMDTotalPatients$tTable)
4
5 #CHANGE TEXT TO MATCH HEALTHBOARDS
6 HITP$HealthRegions <- gsub("relevel\\(HEALTHBOARD, ref = \"Lincolnshire West\\)\",\"", rownames(HITP))
7
8 #DROP ROWS
9 HITP <- HITP[!HITP$HealthRegions == "(Intercept)",]
10 HITP <- HITP[!HITP$HealthRegions == "IMD_RANK",]
11 HITP <- HITP[!HITP$HealthRegions == "TotalPatients",]
12
13 #CREATE QUINTILES OF INTAKE
14 HITP$quintiles <- cut(HITP$Value, quantile(HITP$Value, seq(0, 1, .2)), include.lowest = TRUE, dig.lab=6)
15
16
17 #CREATE COLOURS BASED ON QUINTILES
18 rbPal <- colorRampPalette(c("red","blue"))
19 HITP$Colour <- rbPal(5)[as.numeric(HITP$quintiles)]
20
21 #MERGE
22 UK@data <- data.frame(UK@data, HITP[match(UK@data["HealthRegions"], HITP["HealthRegions"]),])
23
24
25 #CHANGE REFERENCE CATEGORY COLOUR
26 UK@data$Colour <- ifelse(is.na(UK@data$Colour),"#A3A3A3",UK@data$Colour)
27
28 #PLOT
29 pdf("UKxIron.pdf", width = 6, height = 8 )
30 plot(UK, col=UK@data$Colour)
31 legend(-14,61, c("Lowest Quintile [-2676,-1128]",

```

```

32 "Second Quintile (-1128,-642]",
33 "Third Quintile (-642,-132]",
34 "Fourth Quintile (-132,505]",
35 "Highest Quintile (505,4619]",
36 col=rbPal(5), bty="n", pch=15, cex=0.6)
37 dev.off ()
38
39 #####
40 #START AGAIN FOR MAP WITH IRON
41 rm(list=ls())
42
43 #####
44 #OPEN MAP
45 #####
46
47 load("~/IronPrescriptions/IronPrescriptionModels.RData")
48
49 #####
50 #OPEN MAP
51 #####
52
53 load("~/IronPrescriptions/UKMap.RData")
54
55 #####
56 #CONVERT LQMM OUTPUT TO DATA FRAME
57 #####
58
59 HIITP <- data.frame(HealthboardIronIMDTotalPatients$Table)
60
61 #####
62 #CHANGE TEXT TO MATCH HEALTHBOARDS
63 #####
64
65 HIITP$HealthRegions <- gsub("relevel\\(HEALTHBOARD, ref = \\\"Lincolnshire West\\\"\\)", "", rownames(HIITP))
66
67 #####
68 #DROP ROWS
69 #####
70
71 HIITP <- HIITP[!HIITP$HealthRegions == "(Intercept)",]
72 HIITP <- HIITP[!HIITP$HealthRegions == "IMD_RANK",]
73 HIITP <- HIITP[!HIITP$HealthRegions == "TotalPatients",]
74
75 #####
76 #CREATE QUINTILES OF INTAKE
77 #####

```

```

78
79 HIITP$quintiles <- cut(HIITP$Value, quantile(HIITP$Value, seq(0, 1, .2)), include.lowest = TRUE, dig.lab=6)
80
81 #####
82 #CREATE COLOURS BASED ON QUINTILES
83 #####
84
85 rbPal <- colorRampPalette(c("red","blue"))
86 HIITP$Colour <- rbPal(5)[as.numeric(HIITP$quintiles)]
87
88 #####
89 #MERGE
90 #####
91
92 UK@data <- data.frame(UK@data, HIITP[match(UK@data["HealthRegions"], HIITP["HealthRegions"]),])
93
94 #####
95 #CHANGE REFERENCE CATEGORY COLOUR
96 #####
97
98 UK@data$Colour <- ifelse(is.na(UK@data$Colour),"#A3A3A3",UK@data$Colour)
99
100 #####
101 #PLOT
102 #####
103
104 pdf("UKIron.pdf", width = 6, height = 8 )
105 plot(UK, col=UK@data$Colour)
106 legend(-14,61, c("Lowest Quintile [-3138,-2015]",
107 "Second Quintile (-2015,-1394]",
108 "Third Quintile (-1394.23,-876]",
109 "Fourth Quintile (-876,-191]",
110 "Highest Quintile (-191,2781]"),
111 col=rbPal(5), bty="n", pch=15, cex=0.6)
112 dev.off ()

```

---

## K Iron prescription costs across the UK: data sources

Prescription, registered patients and IMD data were downloaded from open source repositories managed by each country and are available from the following locations.

Prescription data:

From the English data.gov.uk website

<https://data.gov.uk/dataset/prescribing-by-gp-practice-presentation-level>

Northern Ireland from the open data NI

<https://www.opendatani.gov.uk/dataset/gp-prescribing-data>

From the Scottish information services division

<http://www.isdscotland.org/Health-Topics/Prescribing-and-Medicines/Publications/2017-01-17/opendata.asp>

Wales from the primary care services website.

<http://www.primarycareservices.wales.nhs.uk/general-practice-prescribing-data-extrac>

The number of patients registered at a GP practice in England were downloaded from

<http://content.digital.nhs.uk/catalogue/PUB19775/gp-reg-patients-prac-quin-age.csv>

which contains data from January 2016. IMD values were obtained through postcodes

using <http://imd-by-postcode.opendatacommunities.org/> which contains a lookup table

that returns the corresponding IMD value. GP addresses were taken from: [https://](https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data)

[digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data](https://digital.nhs.uk/services/organisation-data-service/data-downloads/gp-and-gp-practice-related-data)

Data containing the number of patients registered at Scottish GP practices were extracted from

[http://www.isdscotland.org/Health-Topics/General-Practice/Publications/2015-12-15/Table6\\_](http://www.isdscotland.org/Health-Topics/General-Practice/Publications/2015-12-15/Table6_Practice_ListSizes_by_gender_age_2005_2015.xlsx)

[Practice\\_ListSizes\\_by\\_gender\\_age\\_2005\\_2015.xlsx](http://www.isdscotland.org/Health-Topics/General-Practice/Publications/2015-12-15/Table6_Practice_ListSizes_by_gender_age_2005_2015.xlsx) and values were taken as of 1st of Oc-

tober 2015. Scottish IMD information was obtained from [http://www.gov.scot/Topics/](http://www.gov.scot/Topics/Statistics/SIMD)

[Statistics/SIMD](http://www.gov.scot/Topics/Statistics/SIMD) and matched to GP practice postcodes found at

<http://www.isdscotland.org/Health-Topics/General-Practice/Workforce-and-Practice-Populations/Practices-and-Their-Populations/>

Welsh GP practice addresses and GP registered patient numbers were found at

<http://gov.wales/docs/statistics/2016/160330-gp-practice-populations-gender-age-group-2015-en.xls>,

Welsh IMD information was obtained from

<http://gov.wales/docs/statistics/2015/150812-wimd-2014-overall-domain-ranks-each-lsoa-revised-en.xlsx>

Patient information regarding the age and sex distribution of patients registered at GP practices in Northern Ireland were requested from the health and social care Northern Ireland team as only the total number of registered patients were available on the website. Northern Irish IMD was not available at the postcode level but was provided in small geographical units (LSOA). Then the LSOA of the GP practice was found by converting the postcode from <http://mapit.mysociety.org/postcode/#> (where # is the GP practice postcode). The IMD scores were then found from: [http://www.nisra.gov.uk/deprivation/nimdm\\_2010.htm](http://www.nisra.gov.uk/deprivation/nimdm_2010.htm). The latest available IMD data is from 2010.

## **L Iron prescription costs across the UK: data table**

Table 29 contains the first 20 rows of the data used to estimate the amount spent of iron prescriptions per health board. The first column practice relates to the GP practices that are clustered within health boards. The second column is the total amount spent by that GP practice in the 12 month period from September 2015 to August 2016. The column titled Iron relates to an estimated median iron intake for patients registered at that GP practice based on the geographic location and the sex and age composition of the patients. The next column, IMD Rank is the ranking of the index of multiple deprivation for the area the GP practice is situated in and finally Total Patients is sum of all patients registered at that GP practice.

**Table 29**

Sample of data used to estimate health board spending on iron prescriptions

Practice	Cost	Iron	Healthboard	IMD Rank	Total Patients
1	165.59	3.742	Belfast	2202	1170
3	1647.11	3.687	Belfast	2202	6583
5	420.42	3.798	Belfast	535	1351
6	1882.59	3.665	Belfast	4483	10068
10	622.56	3.733	Belfast	359	2375
13	553.53	3.676	Belfast	359	3227
14	2720.96	3.631	Belfast	2887	7271
15	3977.4	3.711	Belfast	2202	5034
16	1039.02	3.692	Belfast	4792	5011
17	1294.5	3.679	Belfast	2202	4990
18	1881.71	3.700	Belfast	484	6183
19	832.72	3.628	Belfast	1799	3350
20	850.91	3.706	Belfast	2202	2761
23	1396.56	3.668	Belfast	2202	5055
24	962.95	3.739	Belfast	473	2655
28	765.2	3.696	Belfast	2202	3136
29	865.99	3.668	Belfast	4732	4633
30	1029.99	3.679	Belfast	131	2748
31	1870.11	3.658	Belfast	2545	6693
32	1289.24	3.702	Belfast	535	3465



## **M Iron prescription costs across the UK: Estimated regression parameters for the amount spent by health boards in the UK**

Table 30 presents regression coefficients for two models estimating spending on iron prescriptions by health board that adjusts for the number of patients registered at each GP practice within the health board, the index of multiple deprivation ranking for the GP practices within each health board and excluding and including estimated bioavailable iron intakes for those living within the same region as the health board. Coefficients are grouped alphabetically by country with England first then Northern Ireland, Scotland and finally Wales.

**Table 30**

Estimated regression parameters for Median with standard errors for amount spent by health boards in the UK

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Registered Patients	0.40	0.01	<0.001	0.39	0.01	<0.001
Index of Multiple Deprivation	-0.04	0.003	<0.001	-0.04	0.002	<0.001
Iron				-2395.40	296.10	<0.001
<b>England</b>						
Airedale, Wharfedale & Craven	-118.20	609.46	0.847	-1613.48	515.62	0.003
Ashford	-368.76	380.66	0.337	-905.72	356.56	0.014
Aylesbury Vale	-1881.00	492.61	<0.001	-2384.32	397.48	<0.001
Barking & Dagenham	-792.13	354.29	0.030	-1222.63	265.21	<0.001
Barnet	42.60	360.05	0.906	-255.01	326.91	0.439
Barnsley	247.14	417.33	0.556	-1557.19	330.84	<0.001
Basildon & Brentwood	-915.87	301.30	0.004	-830.32	280.86	0.005
Bassetlaw	493.30	567.07	0.389	-907.04	717.14	0.212
Bath & North East Somerset	-401.10	370.72	0.285	-1475.25	286.76	<0.001
Bedfordshire	155.50	373.02	0.679	100.17	314.18	0.751
Bexley	-308.97	313.19	0.329	-592.95	372.51	0.118
Birmingham Crosscity	1521.28	411.37	0.001	505.78	464.00	0.281
Birmingham South & Central	252.05	433.64	0.564	-791.56	432.52	0.073
Blackburn with Darwen	2580.81	722.94	0.001	990.84	832.68	0.240
Blackpool	1643.71	506.12	0.002	295.25	535.26	0.584
Bolton	-825.66	380.73	0.035	-2412.44	366.71	<0.001
Bracknell & Ascot	-2084.52	465.08	<0.001	-2610.27	339.43	<0.001
Bradford City	4056.04	1064.45	<0.001	1903.20	840.74	0.028
Bradford Districts	1779.71	444.43	<0.001	-169.89	550.86	0.759
Brent	-771.87	346.31	0.030	-959.77	247.19	<0.001

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Brighton & Hove	-2089.58	354.13	<0.001	-2452.67	286.42	<0.001
Bristol	-1045.00	320.89	0.002	-2140.79	350.77	<0.001
Bromley	-878.92	313.46	0.007	-1133.36	236.84	<0.001
Bury	548.91	356.19	0.130	-996.55	359.04	0.008
Calderdale	-209.63	388.70	0.592	-1735.47	435.53	<0.001
Cambridgeshire & Peterborough	-567.53	338.70	0.100	-436.94	300.55	0.152
Camden	-186.85	387.93	0.632	-303.66	415.22	0.468
Cannock Chase	-205.50	313.46	0.515	-1071.53	262.41	<0.001
Canterbury & Coastal	-930.77	412.96	0.029	-1361.49	320.06	<0.001
Castle Point & Rochford	-837.84	338.44	0.017	-1247.08	260.27	<0.001
Central London (Westminster)	-1487.14	392.73	<0.001	-1404.68	328.87	<0.001
Central Manchester	-268.22	400.83	0.507	-1826.87	445.59	<0.001
Chiltern	-1434.39	371.82	<0.001	-1977.53	296.45	<0.001
Chorley & South Ribble	-211.18	310.29	0.499	-1695.67	330.23	<0.001
City & Hackney	-1456.91	336.40	<0.001	-1674.22	320.90	<0.001
Coastal West Sussex	-131.56	316.79	0.680	-561.34	284.22	0.054
Corby	-2675.69	1530.55	0.087	-2699.05	1448.30	0.068
Coventry & Rugby	-1114.12	347.31	0.002	-2044.52	280.64	<0.001
Crawley	2169.92	558.87	<0.001	1625.62	412.66	<0.001
Croydon	-1908.62	273.02	<0.001	-2142.50	281.56	<0.001
Cumbria	-254.30	318.89	0.429	-1614.83	288.72	<0.001
Darlington	-518.40	498.97	0.304	-1785.46	460.60	<0.001
Dartford, Gravesham & Swanley	495.55	390.58	0.211	-62.43	431.79	0.886
Doncaster	963.76	350.77	0.008	-860.66	416.88	0.044
Dorset	395.24	336.85	0.246	-624.99	276.33	0.028
Dudley	1930.00	418.29	<0.001	1003.80	434.29	0.025

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Durham Dales, Easington & Sedgefield	156.02	311.65	0.619	-1042.18	382.45	0.009
Ealing	-108.30	317.84	0.735	-339.32	272.92	0.220
East & North Hertfordshire	-375.92	353.79	0.293	-263.80	311.66	0.401
East Lancashire	1352.79	441.90	0.004	-167.12	441.43	0.707
East Leicestershire & Rutland	104.03	435.24	0.812	134.47	362.46	0.712
East Riding of Yorkshire	1578.02	553.88	0.006	-178.54	404.15	0.661
East Staffordshire	-335.25	342.70	0.333	-1246.85	304.77	<0.001
East Surrey	-507.12	378.60	0.187	-1051.08	305.82	0.001
Eastbourne, Hailsham & Seaford	-171.28	356.83	0.633	-607.65	302.50	0.050
Eastern Cheshire	-761.48	393.00	0.058	-2145.34	347.19	<0.001
Enfield	-547.73	353.02	0.127	-884.01	305.22	0.006
Erewash	-1095.57	428.07	0.014	-1046.90	371.83	0.007
Fareham & Gosport	-810.75	305.29	0.011	-1276.84	325.38	<0.001
Fylde & Wyre	395.74	582.12	0.500	-998.44	490.87	0.047
Gloucestershire	-797.14	284.98	0.007	-1846.02	267.56	<0.001
Great Yarmouth & Waveney	-1415.93	413.54	0.001	-1187.47	355.16	0.002
Greater Huddersfield	681.59	386.07	0.084	-1173.09	447.33	0.012
Greater Preston	229.82	524.74	0.663	-1274.98	473.27	0.010
Greenwich	-1334.15	339.83	<0.001	-1625.91	332.27	<0.001
Guildford & Waverley	-2043.30	412.60	<0.001	-2537.82	413.36	<0.001
Halton	557.37	525.14	0.294	-943.11	499.59	0.065
Hambleton, Richmondshire & Whitby	-488.58	311.67	0.123	-2232.59	394.87	<0.001
Hammersmith & Fulham	-1893.79	331.28	<0.001	-2046.84	337.56	<0.001
Hardwick	71.07	407.11	0.862	146.21	425.97	0.733
Haringey	-1317.89	358.57	0.001	-1497.63	259.74	<0.001
Harrogate & Rural District	-1208.49	458.28	0.011	-2981.89	423.60	<0.001

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Harrow	977.17	466.73	0.041	706.51	361.99	0.057
Hartlepool & Stockton-on-Tees	940.39	636.00	0.146	-315.27	568.86	0.582
Hastings & Rother	-485.35	341.95	0.162	-863.84	323.53	0.010
Havering	-841.02	269.73	0.003	-1098.77	253.21	<0.001
Herefordshire	-2086.82	327.87	<0.001	-2892.12	313.37	<0.001
Herts Valleys	-1066.37	291.24	0.001	-987.07	304.05	0.002
Heywood, Middleton & Rochdale	1413.04	435.12	0.002	-164.51	486.40	0.737
High Weald Lewes Havens	-991.33	274.09	0.001	-1442.37	288.24	<0.001
Hillingdon	-36.93	377.04	0.922	-336.77	364.70	0.360
Horsham & Mid Sussex	-188.90	394.74	0.634	-681.66	357.01	0.062
Hounslow	-384.45	286.44	0.186	-637.45	270.97	0.023
Hull	1532.89	488.46	0.003	-289.00	426.45	0.501
Ipswich & East Suffolk	1996.29	520.23	<0.001	2198.19	420.15	<0.001
Isle of Wight	80.24	449.65	0.859	-267.02	396.23	0.504
Islington	-2197.83	351.26	<0.001	-2317.37	337.93	<0.001
Kernow	934.59	318.39	0.005	-60.99	321.46	0.850
Kingston	-1951.59	373.57	<0.001	-2200.92	418.67	<0.001
Knowsley	655.91	405.99	0.113	-736.47	491.40	0.140
Lambeth	-1850.95	305.28	<0.001	-1984.16	241.71	<0.001
Lancashire North	-908.82	677.26	0.186	-2329.03	644.68	0.001
Leeds North	-39.34	428.78	0.927	-1907.83	402.44	<0.001
Leeds South & East	-853.82	279.16	0.004	-2729.18	350.00	<0.001
Leeds West	-1347.82	544.36	0.017	-3138.41	477.12	<0.001
Leicester City	569.10	677.67	0.405	489.47	486.42	0.319
Lewisham	370.69	381.65	0.336	108.04	301.60	0.722
Lincolnshire East	1359.65	752.08	0.077	-370.68	715.21	0.607

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Liverpool	2014.94	392.59	<0.001	567.86	386.87	0.149
Luton	173.66	443.40	0.697	140.38	401.85	0.728
Mansfield & Ashfield	-504.18	450.50	0.269	-471.84	357.91	0.194
Medway	126.89	341.58	0.712	-406.11	286.77	0.163
Merton	-1346.65	453.83	0.005	-1614.12	300.52	<0.001
Mid Essex	241.17	311.19	0.442	390.57	243.61	0.115
Milton Keynes	251.12	374.06	0.505	137.50	325.93	0.675
Nene	-1221.51	312.96	<0.001	-1234.62	233.60	<0.001
Newark & Sherwood	472.79	426.06	0.273	530.47	404.32	0.196
Newbury & District	-1676.84	381.23	<0.001	-2163.50	332.79	<0.001
Newcastle Gateshead	-223.83	379.95	0.558	-1473.02	300.28	<0.001
Newham	68.31	358.97	0.850	-215.84	329.37	0.515
North & West Reading	-370.58	880.65	0.676	-908.84	528.89	0.092
North Derbyshire	109.36	389.01	0.780	204.97	293.49	0.488
North Durham	-129.94	482.07	0.789	-1374.50	514.79	0.010
North East Essex	862.84	504.79	0.094	1038.65	404.87	0.013
North East Hampshire & Farnham	-1922.80	365.43	<0.001	-2472.06	252.06	<0.001
North East Lincolnshire	315.83	466.36	0.501	-1482.13	454.07	0.002
North Hampshire	-920.49	355.76	0.013	-1437.48	354.53	<0.001
North Kirklees	2161.05	629.47	0.001	277.99	590.97	0.640
North Lincolnshire	2458.49	792.05	0.003	663.13	744.75	0.378
North Manchester	1391.87	565.29	0.017	-212.20	537.19	0.695
North Norfolk	248.48	454.06	0.587	521.99	398.20	0.196
North Somerset	-219.21	301.67	0.471	-1319.99	376.30	0.001
North Staffordshire	627.17	383.64	0.109	-186.95	371.76	0.617
North Tyneside	374.66	344.70	0.282	-902.70	372.08	0.019

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
North West Surrey	-824.75	301.51	0.009	-1363.75	255.46	<0.001
Northern, Eastern & Western Devon	-445.91	280.96	0.119	-1495.93	287.92	<0.001
Northumberland	518.86	374.65	0.172	-660.99	311.78	0.039
Norwich	-724.78	448.44	0.112	-557.08	386.45	0.156
Nottingham City	-254.69	386.41	0.513	-316.23	451.74	0.487
Nottingham North & East	145.32	379.00	0.703	170.20	338.25	0.617
Nottingham West	-648.72	304.62	0.038	-622.50	254.53	0.018
Oldham	3081.90	535.38	<0.001	1426.43	665.06	0.037
Oxfordshire	-563.72	327.54	0.092	-1034.29	271.99	<0.001
Portsmouth	-686.90	631.83	0.282	-1068.54	422.22	0.015
Redbridge	71.46	320.30	0.824	-281.08	346.55	0.421
Redditch & Bromsgrove	-951.22	331.18	0.006	-1832.36	337.44	<0.001
Richmond	-1840.59	371.15	<0.001	-2123.11	248.86	<0.001
Rotherham	4618.98	745.63	<0.001	2781.21	932.33	0.004
Rushcliffe	187.23	310.33	0.549	154.00	364.57	0.675
Salford	-645.42	327.90	0.055	-2212.77	353.25	<0.001
Sandwell & West Birmingham	1916.13	446.98	<0.001	941.73	350.11	0.010
Scarborough & Ryedale	-443.37	327.71	0.182	-2180.24	378.60	<0.001
Sheffield	1565.21	449.81	0.001	-273.27	437.37	0.535
Shropshire	351.88	296.90	0.242	-464.29	302.99	0.132
Slough	2016.68	796.98	0.015	1354.93	850.96	0.118
Solihull	949.59	420.11	0.028	17.00	346.06	0.961
Somerset	-1041.74	317.53	0.002	-2093.77	278.17	<0.001
South Cheshire	-183.19	537.97	0.735	-1618.50	578.97	0.007
South Devon & Torbay	793.79	405.85	0.056	-208.59	388.58	0.594
South East Staffordshire & Seisdon Peninsula	-1020.24	360.44	0.007	-1879.93	299.56	<0.001

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
South Eastern Hampshire	-672.82	313.01	0.037	-1118.30	386.89	0.006
South Gloucestershire	-901.77	388.01	0.024	-2016.10	297.38	<0.001
South Kent Coast	-371.68	396.40	0.353	-816.32	309.31	0.011
South Lincolnshire	3137.65	769.67	<0.001	1343.06	800.75	0.100
South Manchester	-629.62	360.65	0.087	-2162.84	394.86	<0.001
South Norfolk	122.16	392.47	0.757	305.75	386.52	0.433
South Reading	-900.23	340.39	0.011	-1442.61	372.24	<0.001
South Sefton	620.01	357.92	0.090	-824.22	380.19	0.035
South Tees	1876.87	599.43	0.003	605.79	521.45	0.251
South Tyneside	-538.33	338.27	0.118	-1717.76	350.36	<0.001
South Warwickshire	-1138.13	361.74	0.003	-1999.97	292.32	<0.001
South West Lincolnshire	814.57	521.59	0.125	872.31	464.98	0.067
South Worcestershire	-89.30	350.32	0.800	-909.83	283.03	0.002
Southampton	-1924.46	375.98	<0.001	-2379.66	365.84	<0.001
Southend	-990.57	291.34	0.001	-757.42	212.17	0.001
Southern Derbyshire	-862.42	430.19	0.051	-858.25	462.72	0.070
Southport & Formby	-561.10	344.81	0.110	-2014.65	311.12	<0.001
Southwark	422.21	358.19	0.244	230.24	328.43	0.487
St Helens	234.35	399.52	0.560	-1206.58	409.13	0.005
Stafford & Surrounds	-587.90	386.49	0.135	-1387.26	312.19	<0.001
Stockport	-546.62	296.51	0.071	-2031.67	345.75	<0.001
Stoke on Trent	1192.77	379.15	0.003	264.64	374.73	0.483
Sunderland	-709.18	302.57	0.023	-1876.32	296.08	<0.001
Surrey Downs	-1145.53	320.98	0.001	-1666.46	295.77	<0.001
Surrey Heath	-2094.60	698.55	0.004	-2601.61	577.38	<0.001
Sutton	-1305.10	279.98	<0.001	-1573.61	269.70	<0.001



Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Swale	-403.69	346.41	0.250	-912.08	292.55	0.003
Swindon	1207.42	379.36	0.003	68.53	397.06	0.864
Tameside & Glossop	475.60	341.98	0.171	-1047.35	387.72	0.009
Telford & Wrekin	1736.89	600.25	0.006	791.95	644.86	0.225
Thanet	-211.23	422.27	0.619	-681.70	355.11	0.061
Thurrock	-1133.97	323.26	0.001	-1099.73	254.38	<0.001
Tower Hamlets	-779.88	368.64	0.039	-1014.77	337.21	0.004
Trafford	389.78	358.12	0.282	-1172.93	498.87	0.023
Vale of York	-509.55	466.29	0.280	-2268.48	443.05	<0.001
Vale Royal	229.27	568.57	0.689	-1230.35	563.43	0.034
Wakefield	-596.94	326.91	0.074	-2374.46	393.77	<0.001
Walsall	64.96	311.56	0.836	-951.34	374.81	0.014
Waltham Forest	-1418.73	300.17	<0.001	-1678.96	276.32	<0.001
Wandsworth	-1727.56	419.71	<0.001	-1944.61	326.10	<0.001
Warrington	667.89	381.59	0.086	-819.12	417.88	0.056
Warwickshire North	-813.97	322.27	0.015	-1686.44	347.83	<0.001
West Cheshire	1239.76	387.34	0.002	-191.71	386.64	0.622
West Essex	-724.52	379.04	0.062	-607.74	303.08	0.050
West Hampshire	-1123.73	331.33	0.001	-1572.10	218.30	<0.001
West Kent	-604.70	337.35	0.079	-1097.89	246.95	<0.001
West Lancashire	-231.15	325.84	0.481	-1663.00	313.93	<0.001
West Leicestershire	147.44	395.72	0.711	171.95	331.00	0.606
West London	-1640.97	319.68	<0.001	-1702.14	277.69	<0.001
West Norfolk	-527.76	453.53	0.250	-288.09	367.20	0.436
West Suffolk	-1005.81	364.68	0.008	-798.51	309.40	0.013
Wigan Borough	-63.51	321.41	0.844	-1561.17	318.08	<0.001

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Wiltshire	-882.53	278.85	0.003	-1951.26	284.28	<0.001
Windsor, Ascot & Maidenhead	-1464.32	421.03	0.001	-1989.04	373.47	<0.001
Wirral	-1107.46	277.22	<0.001	-2589.40	326.04	<0.001
Wokingham	-1669.46	731.50	0.027	-2201.90	688.23	0.002
Wolverhampton	559.33	316.40	0.083	-394.76	353.53	0.270
Wyre Forest	-271.23	415.50	0.517	-1099.28	358.61	0.004
<b>Northern Ireland</b>						
Belfast	-1925.84	287.64	<0.001	-2561.60	233.05	<0.001
Northern	-1878.17	318.97	<0.001	-2556.34	273.81	<0.001
Western	-2162.45	334.30	<0.001	-2822.83	269.70	<0.001
Southern	-2305.69	275.27	<0.001	-3026.29	258.89	<0.001
South Eastern	-2150.34	291.14	<0.001	-2839.59	269.88	<0.001
<b>Scotland</b>						
Ayrshire & Arran	39.82	298.90	0.895	-909.00	268.84	0.001
Borders	-2036.61	297.34	<0.001	-2987.89	273.28	<0.001
Dumfries & Galloway	-2024.88	298.73	<0.001	-2942.92	276.38	<0.001
Fife	898.25	395.99	0.028	-106.22	380.37	0.781
Forth Valley	-948.59	328.24	0.006	-1946.61	292.18	<0.001
Grampian	-1082.81	308.22	0.001	-2068.00	283.79	<0.001
Greater Glasgow & Clyde	-133.98	300.17	0.657	-1101.63	270.99	<0.001
Highland	-762.50	301.78	0.015	-1633.89	273.28	<0.001
Lanarkshire	-953.35	327.06	0.005	-1953.24	283.47	<0.001
Lothian	-1108.07	306.63	0.001	-2117.30	293.55	<0.001

Table 30 continued from previous page

	Excluding Bioavailable Iron			Including Bioavailable Iron		
	Value	SE	P value	Value	SE	P value
Orkney	-1331.92	314.74	<0.001	-2148.09	334.06	<0.001
Shetland	-652.50	312.81	0.042	-1678.37	323.15	<0.001
Tayside	-2063.25	311.63	<0.001	-3054.80	252.30	<0.001
Western Isles	-1189.52	316.93	<0.001	-2094.73	335.98	<0.001
<b>Wales</b>						
Abertawe Bro Morgannwg	104.21	350.16	0.767	-320.58	291.41	0.277
Aneurin Bevan	-556.51	365.47	0.134	-1002.43	300.05	0.002
Betsi Cadwaladr	-342.50	320.36	0.290	-769.62	214.57	0.001
Cardiff & Vale	20.39	291.87	0.945	-462.89	287.96	0.114
Cwm Taf	-1523.39	321.77	<0.001	-1987.06	274.90	<0.001
Hywel Dda	-652.46	301.37	0.035	-1040.15	313.70	0.002
Powys	-1468.75	359.09	<0.001	-1812.61	275.16	<0.001
$\hat{\sigma}$ AL distribution scale parameter		570.06			570.84	